

Article

Analysis of Fall-from-Height Accidents in Construction Based on Text Mining Technology and Improved Apriori Algorithm

Rongjian Sun ^{1,2} and Junwu Wang ^{1,*}

¹ School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430070, China; sun18689605818@163.com

² Haikou Rural Revitalization Investment and Development Co., Ltd., Haikou 570000, China

* Correspondence: junwuwang@whut.edu.cn

Abstract

In recent years, fall-from-height accidents have frequently occurred in construction activities, posing severe risks to workers' safety and impeding the sustainable development of construction enterprises as well as social stability. Due to the complexity and multifactorial nature of such accidents, traditional safety risk assessment methods face significant limitations in uncovering their underlying causes. To address this issue, this study develops a novel analytical framework that integrates text mining with an improved Apriori algorithm. A standardized text preprocessing pipeline is established, including data collection, construction of a domain-specific lexicon, and synonym-based term unification. Key features are extracted using the TF-IDF method, while thematic patterns are identified through LDA topic modeling. To overcome the contextual insensitivity of conventional association rule mining, the Apriori algorithm is enhanced by introducing time-based constraints, enabling the discovery of accident causation patterns that differ between daytime and nighttime. Using 1064 accident reports from 22 provinces in China, the framework extracted 40 high-frequency accident-causing features and generated a richer set of meaningful association rules compared to the standard algorithm. The results indicate that insufficient safety protection, inadequate worker training, and management deficiencies are the predominant causes of fall-from-height accidents. Building on these insights, the study proposes targeted preventive measures. The findings make significant theoretical contributions by enhancing methodological frameworks for accident analysis while also providing practical insights to improve safety management practices in the construction industry.

Keywords: fall-from-height accidents; text mining; TF-IDF; Apriori algorithm; construction safety

Academic Editor: Ahmed Senouci

Received: 21 May 2025

Revised: 20 December 2025

Accepted: 28 January 2026

Published: 1 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the acceleration of urbanization, the construction industry has expanded rapidly, and safety issues on construction sites have increasingly become a focal point of social concern [1,2]. Although construction safety standards and management systems have been continuously improved, working at heights remains one of the most hazardous activities in construction [3]. Falls from height occur frequently, posing severe threats to workers' lives and causing substantial impacts on enterprise operations as well as social stability [4]. In particular, when protective measures are lacking or monitoring

mechanisms are insufficient, the risks of working at height become even more pronounced, highlighting the urgent need for more systematic and data-driven approaches to uncover accident causation and develop effective countermeasures.

Traditional construction safety management primarily relies on experiential judgment and regulatory frameworks [5]. While these measures reduce accident risks to some extent, their explanatory and predictive power remains limited in the face of the complex environments and multifactorial causes of accidents. In recent years, researchers have attempted to introduce system analysis methods, risk assessment models, and data-driven techniques to overcome the shortcomings of conventional approaches [6]. For instance, some studies have constructed multi-factor causal networks based on systems thinking, revealing complex coupling relationships among accidents; others have integrated location data, workers' physiological characteristics, and managerial factors to establish multidimensional risk assessment frameworks; still others have employed accident case analysis and statistical modeling to identify high-risk scenarios [7–9]. Collectively, these efforts indicate that single-dimensional experiential methods are insufficient for modern construction safety management, while interdisciplinary modeling and data mining are emerging as dominant trends.

Against this backdrop, text mining techniques—owing to their advantages in processing unstructured data—have gradually been introduced into construction safety research. By extracting latent patterns from accident reports, inspection records, and other textual materials, researchers can obtain valuable insights from large volumes of data. For example, previous studies have constructed knowledge graphs to visualize risk relationships, employed Term Frequency-Inverse Document Frequency (TF-IDF) and topic models to identify key causal factors, and applied deep learning techniques for automated accident classification [10–12]. From a methodological perspective, research has evolved from simple keyword extraction to topic modeling and semantic analysis, and further toward causal association mining and dynamic pattern recognition. Nevertheless, two major gaps remain. Firstly, a predominant focus on information extraction and classification without sufficient attention to the systematic mining of causal associations [13]. Secondly, a neglect of temporal dynamics, which limits the understanding of the evolving mechanisms of accidents [14].

Association rule mining provides a promising approach to address these limitations. As a typical data-driven technique, the Apriori algorithm has been widely applied in accident analysis across domains such as transportation, energy, and construction, and is effective in identifying high-frequency combinations of causal factors [15–18]. Existing studies have demonstrated that Apriori-based methods can reveal the complex topological structures underlying accident causation [19,20]. However, conventional algorithms often fail to incorporate temporal variables, resulting in insufficient representation of the dynamic evolution of accidents [21]. This highlights the necessity of improving existing algorithms to enhance their adaptability and explanatory capacity in dynamic scenarios.

To further synthesize existing research progress and reveal methodological trends, this study systematically reviews relevant literature on the application of text mining and association rule methods in construction safety over the past decade (Table 1). As shown in Table 1, research methods have progressed from keyword extraction to topic modeling, and subsequently to causal association modeling and integrated approaches, with analytical depth continuously increasing. However, most studies remain focused on static information extraction and classification, with limited attention to the dynamic evolution of accident causation. In particular, few attempts have been made to incorporate temporal variables into association rule analysis.

Table 1. Research trends in the application of text mining and association rule mining (2015–2024).

Year	Methodological Approaches	Main Research Focus
2015	Text mining (e.g., TF-IDF, LDA) [22]	Keyword extraction, preliminary accident pattern identification
2016	Apriori algorithm [23]	Association rule mining of fall-from-height factors
2017	Deep learning + text mining [24]	Accident classification and high-risk pattern prediction
2018	LDA + system analysis [25]	Multidimensional analysis and causal relationship modeling
2019	Text mining + system dynamics models [26]	Temporal and dynamic analysis of accidents
2020	Improved Apriori algorithm [27]	Incorporating temporal variables for dynamic association analysis
2021	Deep learning + Apriori + text analysis [28]	Deep mining of causal relationships in accidents
2022	Text mining + improved Apriori algorithm [29]	Multi-factor analysis and dynamic prediction of falls from height
2023	Integrated approaches (e.g., TF-IDF + LDA + Apriori) [30]	Multi-method fusion for safety management optimization
2024	Other methods (e.g., network analysis, knowledge graph construction) [31]	Systematic accident causation analysis and risk assessment

In response, this study proposes an analytical framework that integrates text mining techniques with an improved Apriori algorithm to investigate falls-from-height accidents. The proposed method first establishes a standardized text preprocessing procedure to enhance data quality and analytical efficiency. Accident features are then identified using TF-IDF weighting and LDA modeling. Finally, by introducing construction time variables, the Apriori algorithm is improved to systematically mine both the complex associations among causal factors and their evolutionary characteristics. Using 1064 accident reports collected from 22 provinces across China, the study validates the effectiveness of the proposed approach. The results reveal key risk factors and association patterns in falls-from-height accidents and further propose targeted prevention strategies, thereby providing both theoretical insights and practical guidance for improving safety management in construction.

The framework of this paper is as follows: Section 1 introduces research background. Section 2 elaborates the detection method proposed in this article. Section 3 reports the experimental verification. Section 4 analyzes the calculation results, and Section 5 gives the conclusion.

2. Methodology

2.1. TF-IDF Algorithm for Weighted Feature Extraction

In text mining, the TF-IDF algorithm is commonly used to assign weights to feature terms, thereby more accurately reflecting their importance within a document and improving the effectiveness of text analysis [32]. The TF-IDF metric integrates two components: term frequency (TF), which measures how often a term appears in a specific document, and inverse document frequency (IDF), which quantifies how unique or rare the term is across the entire corpus.

Term frequency $TF_{i,j}$ represents the normalized frequency of term t_i in document d_j , indicating that the more frequently a term appears in a document, the more significant it is considered for that document. In contrast, inverse document frequency IDF_i evaluates the rarity of the term across all documents in the corpus. A term that appears frequently across many documents is deemed less effective in distinguishing between them and thus assigned a lower weight. Conversely, terms that occur less frequently across the corpus but appear prominently in a particular document are considered more informative.

By multiplying TF and IDF, the TF-IDF weight $w_{i,j}$ of a term in a specific document can be calculated. This weighting mechanism ensures that common but uninformative terms are down-weighted, while rare yet contextually significant terms are emphasized,

thus enhancing the accuracy and relevance of feature selection in subsequent text analysis tasks [33].

The TF-IDF computation is defined by the following formulas [19]:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \quad (2)$$

$$W_{ij} = tf_{i,j} \times idf_i \quad (3)$$

where $tf_{i,j}$ measures how frequently a term occurs in a document. $n_{i,j}$ is the number of occurrences of term t_i in document d_j . $\sum_k n_{k,j}$ is the total number of term occurrences in document d_j . idf_i measures how important a term is across the entire corpus. $|D|$ is the total number of documents in the corpus. $|\{j: t_i \in d_j\}|$ denotes the number of documents containing term t_i . W_{ij} is used to determine the relevance of a term in the context of a document.

This weighting scheme allows the model to focus on the most relevant terms in a document, mitigating the limitations of frequency-based analysis alone and improving the precision of downstream processing tasks such as feature selection, clustering, and classification.

2.2. Improved Apriori Algorithm for Construction Accident Association Rules

The Apriori algorithm is a classic and widely used method in the field of association rule mining, known for its simplicity, interpretability, and effectiveness in discovering frequent itemsets and potential correlations within datasets. These features enable it to support data-driven decision-making in various domains. However, the standard Apriori algorithm has several limitations that constrain its practical applicability. These include high computational complexity, low efficiency when processing large-scale datasets, poor adaptability to sparse data, and variable quality of the generated rules. Some of these challenges were observed during the TF-IDF-based text mining process discussed in the previous section.

To address the inherent limitations of the traditional Apriori algorithm—such as its high computational complexity, limited scalability to sparse datasets, and insufficient capacity to capture contextual dependencies—this study introduces an improved variant specifically adapted to the temporal characteristics of construction accident data. The methodological innovation lies in the incorporation of a temporal constraint, whereby accident records are partitioned into daytime (06:00–18:00) and nighttime (18:00–06:00) subsets prior to rule extraction. By embedding temporal information into the mining process, the proposed approach enhances the relevance and interpretability of association rules, reduces analytical noise, and improves the efficiency of frequent itemset generation. In contrast to existing enhancements of Apriori, which primarily emphasize computational optimization, the proposed framework advances a domain-oriented extension that explicitly models the temporal variability of accident causation. This enables the discovery of time-sensitive patterns that provide novel insights into the dynamics of construction safety risks.

The improved Apriori algorithm is implemented in six steps as follows:

Step 1: Define the three key metrics used in association rule mining—support, confidence, and lift—and set the minimum thresholds for support and confidence. The formulas are as follows:

Support: Measures how frequently a rule occurs in the dataset.

$$\text{Support}(X \rightarrow Y) = P(X, Y) = \frac{\text{Sum}(X, Y)}{\text{Sum}(\text{samples})} \quad (4)$$

where $\text{Sum}(X, Y)$ is the number of samples where both X and Y occur together. $\text{Sum}(\text{samples})$ is the total number of samples in the dataset.

Confidence: Measures the probability that Y occurs given that X occurs.

$$\text{Confidence}(X \rightarrow Y) = P(Y|X) = \frac{\text{Sum}(X, Y)}{\text{Sum}(X)} \quad (5)$$

where $\text{Sum}(X)$ is the number of samples where X occurs.

Lift: Measures how much more likely Y is to occur given that X occurs, compared to if they were independent.

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(X \rightarrow Y)} = \frac{\text{Sum}(X, Y)}{\text{Sum}(X) \times \text{Sum}(Y)} \quad (6)$$

where $\text{Sum}(Y)$ is the number of samples where Y occurs.

Step 2: Traverse the dataset of construction accident reports and generate the candidate 1-itemsets. Based on the accident occurrence time, partition the dataset into two subsets: daytime and nighttime accident reports.

Step 3: According to the predefined time constraints (06:00 and 18:00), split the candidate 1-itemsets into two groups: candidate 1-itemsets for daytime (A_1^b) and candidate 1-itemsets for nighttime (A_1^h).

Apply the minimum support threshold to derive frequent 1-itemsets for each group: frequent 1-itemsets for daytime (P_1^b) and frequent 1-itemsets for nighttime (P_1^h).

Step 4: Generate candidate 2-itemsets for both groups: A_2^b and A_2^h , followed by frequent 2-itemsets P_2^b and P_2^h .

Step 5: Repeat Step 4 iteratively to generate higher-order itemsets. The candidate itemsets are $A_1^b, A_2^b, \dots, A_n^b$ and $A_1^h, A_2^h, \dots, A_n^h$. The frequent itemsets are $P_1^b, P_2^b, \dots, P_2^b$ and $P_1^h, P_2^h, \dots, P_2^h$.

Step 6: Based on the metrics defined in Step 1, extract strong association rules between accident-related factors and construction accident occurrences under different time periods.

The flowchart of the improved Apriori algorithm is presented in Figure 1, illustrating the integration of time-based filtering into the association rule mining process. By introducing the temporal dimension, this enhanced algorithm enables more granular and context-sensitive analysis of accident causation patterns in the construction industry.

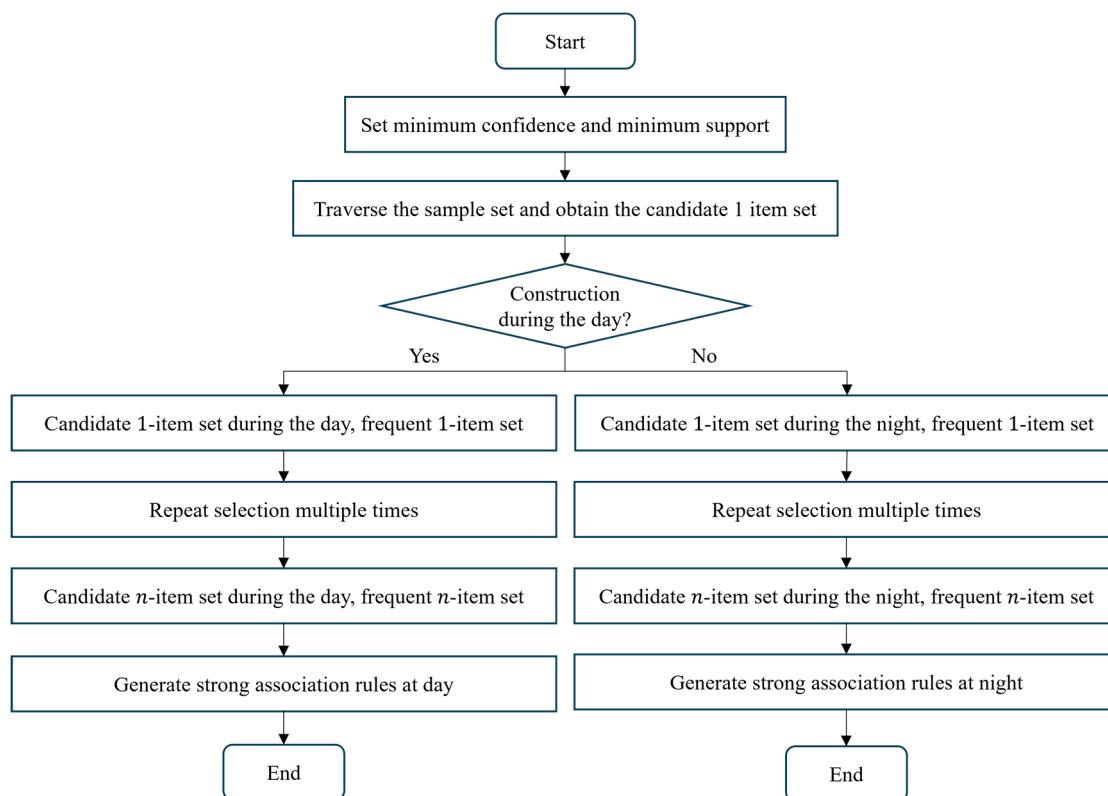


Figure 1. Workflow of the improved Apriori algorithm.

3. Experimental Verification

3.1. Data Processing

As one of the most common types of construction accidents, falls from height are a leading cause of both fatal and non-fatal injuries among construction workers in China. These incidents are often attributed to a combination of individual and systemic factors. Inadequate safety training by managers, worker negligence, and insufficient protective measures are frequent contributors to fall-related accidents, which often result in irreversible outcomes such as severe injury or death.

3.1.1. Text Corpus Data Collection of Fall-from-Height Accidents in Construction

Accident investigation reports are among the most informative and detailed sources for analyzing construction safety incidents. These reports typically document critical contextual information such as the time and location of the incident, the physical conditions of the site, a summary of the sequence of events, casualty figures, and the economic losses incurred. They also provide direct and indirect causes of the accident, assign responsibility, and propose remedial measures. Compared to other types of documentation, accident investigation reports offer a more precise and comprehensive depiction of the incident.

To conduct a robust analysis of construction safety issues, this study utilizes accident reports specifically related to high-fall incidents. Prior to analysis, a substantial effort was made to collect and clean the accident data. Through a combination of on-site investigation and web-based searches, a total of 1064 fall-from-height accident reports were collected from 22 provincial-level administrative regions in China, covering the period from 2013 to 2023, as shown in Table 2. To examine the characteristics of this dataset, a Chi-square (χ^2) goodness-of-fit test was conducted. The results ($\chi^2 = 342.5$, $p < 0.001$) indicate a statistically significant non-uniform distribution across regions. This variation reflects the regional differences in construction industry scale, where provinces with higher reporting frequencies correspond to areas with more intensive construction activities. These reports

were obtained from publicly accessible and authoritative official platforms, including the Ministry of Housing and Urban-Rural Development (MOHURD) and relevant provincial/municipal government departments, ensuring high data transparency and credibility. Although the sample distribution varies across provinces due to regional differences in construction scale, the dataset provides a representative foundation for identifying universal accident causation patterns. The availability of extensive and diverse data sources enables a more comprehensive understanding of the root causes of such incidents.

Table 2. Statistics of provincial administrative regions where accidents occurred.

Provincial administrative Region	Number of Samples	Provincial Administrative Region	Number of Samples
Anhui Province	45	Hebei Province	72
Beijing City	18	Henan Province	97
Chongqing City	23	Zhejiang Province	88
Fujian Province	12	Hubei Province	76
Gansu Province	18	Hunan Province	59
Guangdong Province	67	Jiangsu Province	73
Hainan Province	12	Jiangxi Province	47
Guizhou Province	32	Jilin Province	41
Shandong Province	56	Liaoning Province	35
Shanxi Province	31	Sichuan Province	58
Shaanxi Province	45	Shanghai	59

After collecting the reports, a systematic classification and organization process was undertaken. Each report was categorized based on various dimensions such as time of occurrence, geographic location, type of accident, and severity of casualties. This classification facilitated a clearer understanding of the temporal, regional, and contextual characteristics of fall-from-height incidents. To ensure data quality, duplicate reports were identified through cross-checking of metadata (e.g., report number, accident date, and location) and subsequently removed. Incomplete reports lacking essential information such as accident time, location, casualty details, or causative factors were excluded from the dataset. For cases where minor inconsistencies were found across different fields, information was verified and corrected using authoritative government sources. These procedures reduced analytical noise and improved the reliability of the corpus. Furthermore, an integrated analysis was conducted on contributing factors, including human behavior, environmental conditions, and deficiencies in protective infrastructure prior to the accident, as well as post-accident emergency response, rescue efforts, and responsibility attribution.

Statistical methods were employed to analyze data quantities and trends, investigating accident frequency, severity, and scope of impact. Comparative analyses were also carried out across different regions, construction types, and ownership groups to identify both commonalities and contextual distinctions. These efforts provided a comprehensive foundation for understanding the patterns, causes, and mitigation strategies of high-fall accidents, thereby supporting the development and validation of the subsequent analytical model.

3.1.2. Stop Word List and Jieba Word Segmentation

The construction of a stop word list is an essential component of text preprocessing, designed to filter out commonly used words that carry little or no semantic value in analytical tasks. This process significantly enhances the accuracy and efficiency of text analysis. In this study, special attention was given to the development of a domain-specific stop word list, tailored to the context of construction and safety-related documents.

Typical stop words include everyday functional words such as “is,” “of,” and “has,” which are frequently used but provide no useful information for accident analysis. In addition, personal pronouns like “I,” “you,” and “he,” as well as common verbs such as “do,” “go,” and “come,” were excluded from the analysis, as they primarily serve grammatical purposes without contributing meaningfully to the identification of causative patterns in fall-from-height incidents.

To better understand term distribution across the dataset, preliminary segmentation was applied to the collected accident reports. Word frequency statistics were then computed across the corpus. Frequently occurring words such as “accident,” “safety,” and “production” were found to rank among the highest in term frequency. These terms, while prominent, primarily reflect the general nature of the accident reports rather than offering actionable insights into specific causative factors, and were therefore removed. Similarly, terms like “supervision,” “construction personnel,” and “project start,” although frequently mentioned, were excluded due to their limited analytical contribution in the context of this study’s objectives.

The primary goal of this step was to refine and filter the feature set, enabling a more targeted and effective text analysis. Jieba was employed to tokenize the accident reports into granular semantic units. This segmentation laid the foundation for downstream processing such as feature weighting and association rule mining. Figure 2 presents the frequency distribution of selected key terms extracted from the accident reports following the segmentation process.

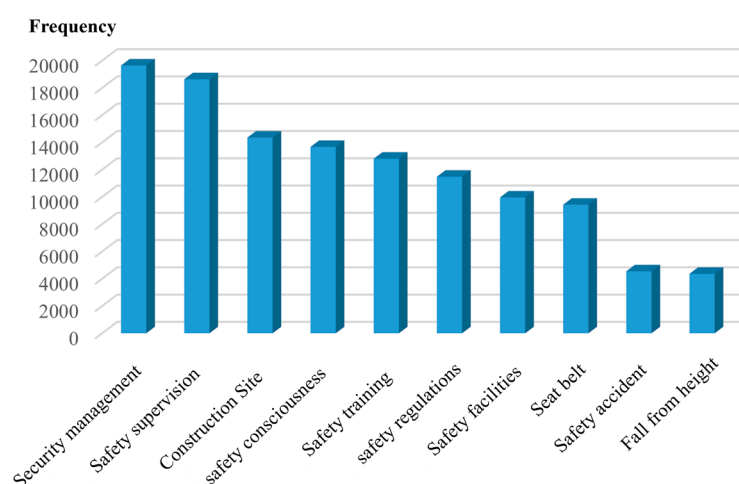


Figure 2. Frequency distribution of top construction safety-related terms extracted from accident reports.

3.1.3. Synonym Database Construction and Expression Merging

In addition to word segmentation using the Jieba, the construction of a synonym dictionary and the merging of synonymous expressions are critical components in the text preprocessing workflow. The objective is to consolidate words or phrases that share the same or similar meanings into unified representations, ensuring greater semantic consistency in the subsequent analytical steps.

For example, the term “fall-from-height” may appear in various forms such as “high fall,” “falling accident,” or “high-altitude operation accident.” These variations were unified under a single standardized term to avoid redundancy and semantic fragmentation. This process significantly reduces ambiguity during text vectorization and improves the accuracy of feature extraction. The synonyms and their corresponding standardized terms used in this study are presented in Table 3.

Table 3. Merged word list.

Replaced Word	Synonyms
Fall-from-height	high fall, falling accident, high-altitude operation accident, high-altitude fall
Construction	construction site, construction site, engineering construction, construction engineering
Accident analysis	accident cause analysis, accident occurrence analysis, accident investigation, accident report
Association model	association rule model, association analysis model, association mining model, association relationship model
Text mining	text analysis, information mining, text analysis technology, text data mining
Safety management	safety measures, safety specifications, safety standards, safety supervision
Risk prevention	risk control, risk management, hazard prevention, risk reduction
Working environment	construction environment, working conditions, workplace, work site
Injury assessment	injury degree assessment, injury classification, injury statistics, injury prediction
Safety awareness	safety education, safety training, safety awareness improvement, safety culture construction
Preventive measures	safety measures, hazard control, accident prevention, risk prevention
Report analysis	accident report, analysis report, research report, investigation report
Safety standards	safety specifications, safety requirements, safety indicators, safety restrictions
Construction projects	engineering construction, construction project, engineering construction, construction engineering
Work management	construction management, site management, project management, site management
...	...

Following preprocessing and initial feature extraction from the accident investigation reports, a set of causative features was identified. Due to the non-standardized format of these reports, the same root cause could be expressed using different phrasings. For example, phrases such as “inadequate safety management” and “deficient safety oversight” may convey identical meanings. Without synonym unification, such variations may result in fragmented term frequencies, leading to an underrepresentation of critical causative factors in the analysis.

By implementing the synonym merging process and eliminating redundant high-frequency terms, the dimensional consistency of the text vector space was enhanced. This allowed for more robust feature representation, reducing analytical noise and emphasizing meaningful patterns. As a result, 40 high-frequency accident-causing feature items were extracted, which served as key indicators in the subsequent association rule mining. These are summarized in Table 4.

A critical comparison with previous studies shows that many of these features—such as inadequate safety training, improper use of protective equipment, poor site management, and worker negligence—are consistent with prior findings, supporting the reliability of our extraction process. Additionally, several context-specific factors, including environmental and temporal conditions, were identified, which have been less emphasized in earlier research. These findings not only confirm established risk factors but also provide new insights for targeted preventive strategies.

Table 4. Main accident-causing characteristics of height-falling accidents.

Index	Causal Item	Index	Causal Item
T_1	Incomplete safety facilities	T_{21}	Lack of emergency response plan
T_2	Improper operation by operators	T_{22}	Incomplete safety monitoring system
T_3	Irregular operations at heights	T_{23}	Lack of warning signs at the workplace
T_4	Lack of effective safety training	T_{24}	Messy work site
T_5	Inadequate safety management	T_{25}	Unqualified work tools
T_6	Complex working environment	T_{26}	Lack of emergency rescue equipment

T_7	Equipment failure	T_{27}	Unclear work process
T_8	Poor construction management	T_{28}	Lack of safety production responsibility system
T_9	Illegal operations	T_{29}	Poor personnel quality
T_{10}	Human negligence	T_{30}	Lack of safety isolation measures at the construction site
T_{11}	Lack of work safety management system	T_{31}	Lack of safety production education
T_{12}	Poor supervision	T_{32}	Excessive working time
T_{13}	Lack of safety awareness	T_{33}	Excessive fatigue
T_{14}	Unskilled skills	T_{34}	Insufficient safety supervisors
T_{15}	Excessive work intensity	T_{35}	Untimely equipment maintenance
T_{16}	Inadequate safety inspections	T_{36}	Lack of emergency response capabilities of workers
T_{17}	High work pressure	T_{37}	Incomplete safety culture construction
T_{18}	Not wearing safety protection equipment	T_{38}	Chaotic on-site management
T_{19}	Chaotic management	T_{39}	Lack of regular safety drills
T_{20}	Narrow work area	T_{40}	Lack of effective communication mechanism during construction

3.2. Extracting Association Rules

3.2.1. LDA Topic Mining Model

The extraction of association rules and the construction of the Latent Dirichlet Allocation (LDA) topic model are two essential steps in the text mining process. Together, they enable a deeper understanding of the relational structure and latent themes embedded within textual data.

Association rule mining is a data mining technique used to identify frequent patterns and relationships among items within a dataset. In the context of text mining, this approach can be applied to detect co-occurring terms and explore the hidden relationships between words. For example, by analyzing term co-occurrence patterns, it is possible to uncover associations such as that between “inadequate safety management” and “poor construction quality,” thereby offering valuable insights into the root causes of construction safety incidents and informing targeted interventions.

LDA is a widely used probabilistic model for topic discovery in large-scale text corpora. It assumes that each document is a mixture of multiple topics, and each topic is characterized by a distribution over words. By analyzing the distribution of terms across documents, LDA can identify the latent thematic structures within the corpus. In the context of construction safety, LDA facilitates the identification of key themes such as safety management, construction quality, and worker training from accident reports, thereby helping stakeholders understand the mechanisms and patterns underlying safety failures.

LDA also supports visualization of keyword–topic relationships, making it easier to interpret the semantic structure of large text datasets. To illustrate the mechanics of the LDA model, a small example is provided.

Table 5 presents excerpts from four sample accident reports. These samples were pre-processed through tokenization, stop word removal, and synonym merging before being transformed into a document–term matrix, which served as input for LDA modeling. The number of topics was set to three, determined through a combination of domain knowledge and coherence score evaluation to ensure interpretability.

Table 5. Excerpts from the accident report.

Reports	Content
1	Fall-from-height is one of the common accidents in construction...
2	Safety belts at construction sites are important equipment to prevent accidents caused by fall-from-height...
3	Safety training and regulations are one of the key measures to prevent accidents caused by fall-from-height...

4 Strengthen safety supervision to ensure that safety facilities at construction sites are complete and effective...

The LDA model was trained using Gibbs sampling for 1000 iterations to achieve convergence, which was confirmed by monitoring the stabilization of the model log-likelihood. The hyperparameters were set as $\alpha = 0.1$ and $\beta = 0.01$, based on iterative tuning experiments to balance model fit and topic interpretability. Specifically, multiple configurations were tested, and these values were found to provide the most stable log-likelihood while ensuring that the resulting topics were semantically distinct and aligned with domain knowledge. A smaller α encourages sparse document–topic distributions, reflecting that each accident report typically covers a limited number of topics, while β was selected to minimize noise in word–topic distributions. Model validity was further confirmed through manual inspection of the top words in each topic. The resulting topic–word distributions are presented in Table 6.

Table 6. List of topics and their corresponding word weights.

Topic	Weight	Content
1	0.568	Safety training and regulations are one of the key measures to prevent accidents caused by fall-from-height
	0.144	Fall-from-height is one of the common accidents in construction, and safety awareness and management need to be strengthened
	0.144	Strengthen safety supervision to ensure that the safety facilities at the construction site are complete and effective
	0.144	Safety belts at the construction site are important equipment to prevent accidents caused by fall-from-height
	0.399	Strengthen safety supervision to ensure that the safety facilities at the construction site are complete and effective
2	0.399	Fall-from-height is one of the common accidents in construction, and safety awareness and management need to be strengthened
	0.101	Safety training and regulations are one of the key measures to prevent accidents caused by fall-from-height
	0.101	Safety belts at the construction site are important equipment to prevent accidents caused by fall-from-height
	0.568	Safety belts at the construction site are important equipment to prevent accidents caused by fall-from-height
3	0.144	Strengthen safety supervision to ensure that the safety facilities at the construction site are complete and effective
	0.144	Fall-from-height is one of the common accidents in construction, and safety awareness and management need to be strengthened
	0.144	Safety training and regulations are one of the key measures to prevent accidents caused by fall-from-height

The document–topic distributions, representing the proportion of each document associated with each topic, are shown in Table 7.

Table 7. Topic distribution of documents.

Reports	Topic Distribution
1	[(0, 0.16874394), (1, 0.6625126), (2, 0.16874345)]
2	[(0, 0.16798162), (1, 0.16751395), (2, 0.6645044)]
3	[(0, 0.6645043), (1, 0.16751406), (2, 0.16798168)]
4	[(0, 0.1687426), (1, 0.66251326), (2, 0.16874418)]

As shown in Tables 6 and 7, the LDA model successfully identified three major topics related to fall-from-height accidents in construction settings. A detailed interpretation of the topics is as follows:

Topic 1 predominantly revolves around safety training and regulations, highlighting the critical role of procedural and educational measures in accident prevention.

Topic 2 emphasizes safety supervision and facility completeness, reflecting the importance of active monitoring and the presence of physical safety protections at construction sites.

Topic 3 focuses on the use of personal protective equipment, particularly safety belts, underscoring the significance of individual-level preventive measures.

An examination of the topic distributions reveals that most accident reports simultaneously relate to multiple topics. For instance, Report 1 and Report 4 exhibit a higher contribution from Topic 2, indicating that deficiencies in site supervision and facility management are recurring issues. Reports 2 and 3, in contrast, show stronger associations with Topic 3 and Topic 1, respectively, suggesting a greater emphasis on individual behavior and safety training in these cases.

This thematic decomposition suggests that fall-from-height accidents are not driven by a single isolated factor but rather arise from the interaction of multiple risk domains, including institutional management failures, equipment inadequacies, and insufficient worker education. Importantly, the topic distributions obtained through LDA were further utilized to inform the association rule mining stage. Terms with high loadings within the same topic were grouped into semantically coherent clusters, which served as the basis for constructing itemsets. This approach reduced dimensional redundancy and ensured that the rules extracted captured domain-relevant relationships rather than random co-occurrences. For example, safety training-related terms identified in Topic 1 were treated as a unified group when generating candidate itemsets, thereby enabling the discovery of rules that more accurately reflected causal linkages among managerial, procedural, and behavioral factors.

3.2.2. Extraction of Association Rules

Building upon the previously introduced TF-IDF weighting method, this section applies the technique to construction fall-from-height accident reports to quantify the significance of extracted textual features. These weighted feature items are then used in conjunction with LDA topic modeling for the subsequent extraction of association rules. In the matrix presented in Table 8, D represents individual accident report documents, while T denotes feature terms identified through the TF-IDF process.

Table 8. Text feature quantification based on TF-IDF.

D	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}	T_{13}	T_{14}
1	0	0.8	0.1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0.2	0	0	0.1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0
4	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0.1	0.3	0	0	0.1	0	0	0
6	0	0	0	0	0	0.5	0	0	0	0.3	0	0	0	0
7	0	0	0	0	0	0.2	0	0	0.8	0	0	0	0	0
8	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0.1	0	0	0	0.3	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0.3	0	0	0.6
11	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0.6	0	0	0	0	0.4	0	0	0

13	0	0	0	0.2	0	0.6	0	0	0	0	0	0	0	0.1
14	0	0	0.1	0	0	0	0	0	0	0	0.2	0	0.5	0
15	0.2	0	0	0	0	0	0	0	0	0.5	0	0	0	0.3
16	0.7	0	0	0	0.1	0	0.1	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	1	0	0	0	0	0	0
19	0	0	0	0	0	0	0	1	0	0	0	0	0	0
20	0.4	0	0	0	0	0.1	0	0	0	0	0	0.1	0	0

Falls from height are among the most prevalent and severe types of accidents in the construction industry, often leading to significant injuries or fatalities. To mitigate such incidents, association rule mining can be employed to identify patterns and correlations among accident-related factors. This technique uncovers frequent co-occurrences and dependencies between variables, enabling researchers and safety managers to pinpoint high-risk scenarios and potential causes.

For instance, Table 9 presents examples of association rules derived from a set of fall-related accident reports, highlighting key contributing factors such as adverse weather, lack of safety equipment usage, and limited worker experience.

Table 9. Association rule examples.

Index	Factor	Accident	Support	Confidence	Lift
1	Adverse weather conditions	Fall-from-height accidents	0.10	0.80	1.30
2	Lack of seat belt use	Fall-from-height accidents	0.15	0.90	1.10
3	Insufficient work experience	Fall-from-height accidents	0.20	0.85	1.20

To demonstrate the implementation of the improved Apriori algorithm, a subset of accident cases was divided into daytime and nighttime datasets. Each case includes a variety of accident-related factors such as construction risks, worker casualties, material damage, equipment failure, and structural collapse, as shown in Table 10.

Table 10. Data examples during the day and at night.

Case	Accident Factors	
	Day	Night
1	Construction risks, worker casualties, building material damage	Construction risks, mechanical failure
2	Construction risks, mechanical failure	Damage to building materials, collapse of building structures
3	Building material damage, building structure collapse	Construction risks, worker casualties
4	Worker casualties, building structure collapse	Construction risks, mechanical failure, worker casualties
5	Construction risks, mechanical failure, worker casualties	Construction risks, worker casualties, damage to building materials
6	Construction risks, worker casualties	Construction risks, worker casualties
...	...	...

By setting a minimum support threshold of 0.01 and a minimum confidence threshold of 0.03, a series of strong association rules were extracted for daytime conditions, as summarized in Table 11. These thresholds were selected to balance the extraction of meaningful associations while avoiding weak or irrelevant rules. A support threshold of 0.01 allows for the inclusion of less frequent yet potentially significant patterns, while the 0.03 confidence threshold ensures the reliability of the rules. The impact of these thresholds is

evident: lower thresholds may result in a larger set of rules, whereas higher thresholds focus the analysis on stronger, but fewer, associations.

Table 11. Association rules during the day.

Rules			Support	Confidence	Lift
Worker casualties	=>	Collapse of building structures	0.17	0.25	0.75
Worker casualties	=>	Damage to building materials	0.17	0.25	0.75
Worker casualties	=>	Mechanical failure	0.17	0.25	0.75
Worker casualties	=>	Construction risks	0.50	0.75	1.125
Building material damage	=>	Collapse of building structures	0.17	0.50	1.50
Building material damage	=>	Construction risks	0.17	0.25	0.75
Construction risks	=>	Mechanical failure	0.33	0.50	1.50
Worker casualties, building material damage	=>	Construction risks	0.17	1.00	1.50
Worker casualties, construction risks	=>	Mechanical failure	0.17	0.33	1.00
...		...	-	-	-

Based on these rules, a daytime risk association network was constructed using the improved Apriori algorithm. As shown in Figure 3, nodes represent various risk sources (e.g., worker casualties, material damage), and directed edges denote association rules. The size of nodes reflects their frequency or importance, while edge thickness or labels can represent support, confidence, or lift metrics. This visualization aids in identifying key contributing factors and their relationships under daytime conditions.

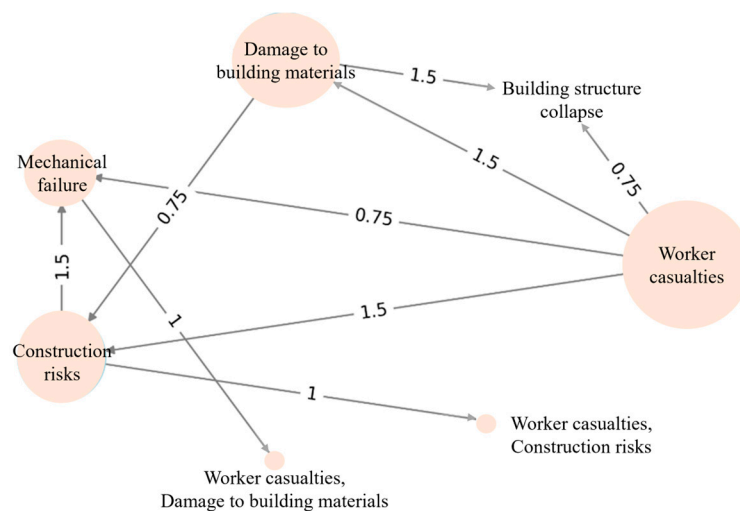


Figure 3. Daytime association rule network model.

Similarly, nighttime accident reports were analyzed using the same method and thresholds. The resulting association rules are presented in Table 12.

Table 12. Association rules during the night.

Rules			Support	Confidence	Lift
Worker casualties	=>	Damage to building materials	0.17	0.25	0.75
Worker casualties	=>	Mechanical failure	0.17	0.25	0.75
Worker casualties	=>	Construction risks	0.67	1.00	1.20
Building material damage	=>	Building structure collapse	0.17	0.50	3.00
Building material damage	=>	Construction risks	0.17	0.50	0.60

Construction risks	=>	Mechanical failure	0.33	0.40	1.20
Worker casualties, building material damage	=>	Construction risks	0.17	1.00	1.50
Worker casualties, construction risks	=>	Mechanical failure	0.17	0.25	0.75
...		...	-	-	-

The nighttime association rule network, shown in Figure 4, reveals a more complex and interconnected structure compared to the daytime model. Quantitatively, the nighttime model identifies a larger set of strong association rules, and its directed edges exhibit significantly more intersections. This increased connectivity suggests that construction activities at night involve a stronger coupling effect among multiple risk factors, such as limited visibility, fatigue, and operational uncertainty. Unlike the relatively dispersed patterns in the daytime (Figure 3), the nighttime risk factors are more likely to act in combination, forming intricate accident causation chains where individual failures rapidly escalate into severe incidents.

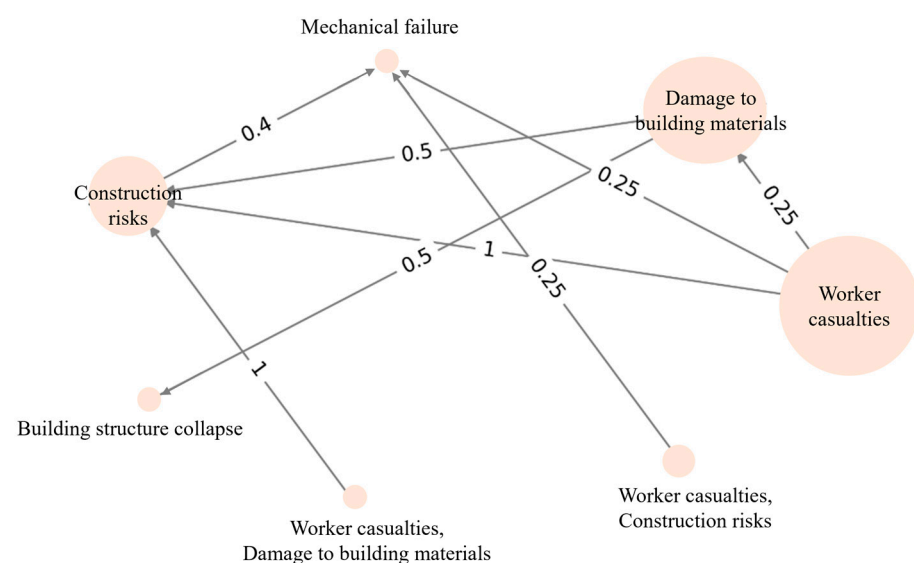


Figure 4. Night association rule network model.

A more in-depth analysis of the network structure shows that during the day, key risk nodes, such as “insufficient safety measures” and “lack of proper training,” are more dispersed, implying that multiple factors contribute to accidents in a less centralized way. In Figure 3, the daytime accident chains tend to be more linear and localized, where a single deficiency (e.g., unqualified tools) directly leads to a specific consequence. In contrast, the nighttime network (Figure 4) exhibits a highly centralized structure with “Construction risks” acting as a critical intersection. This suggests that at night, problems like “fatigue” and “low visibility” do not act in isolation but have a compounded effect, creating a “web-like” coupling of risks. For instance, the path in Figure 4 involving “Mechanical failure → Construction risks → Worker casualties” indicates that equipment issues are more likely to escalate into systemic site risks under poor nighttime lighting conditions. Consequently, for daytime safety, management should focus on “point-to-point” inspections and individual skill training. For nighttime safety, a more holistic approach is required, emphasizing the optimization of the overall working environment (e.g., high-intensity lighting) and the reduction in systemic complexity to break the interdependent accident chains.

4. Results and Analysis

4.1. Ablation Experiment

The improved Apriori algorithm was designed to uncover deeper and more meaningful association rules among the causative factors of construction fall-from-height accidents from textual data. Its objective is to extract a richer set of associations that can better represent the internal connections between accident causes and facilitate effective visualization. To evaluate the performance of the proposed algorithm, a series of comparative experiments were conducted by varying support and confidence thresholds under different time settings (daytime and nighttime). Both the standard Apriori algorithm and the improved version were tested in 10 repeated experiments. The parameter settings for each experiment are listed in Table 13.

Table 13. Comparative experimental parameter settings.

Experiments	Daytime		Night	
	Support	Confidence	Support	Confidence
1	0.01	0.03	0.04	0.1
2	0.01	0.04	0.04	0.4
3	0.01	0.06	0.04	0.6
4	0.01	0.08	0.04	0.7
5	0.01	0.1	0.05	0.8
6	0.02	0.1	0.05	0.1
7	0.03	0.1	0.06	0.4
8	0.04	0.1	0.06	0.6
9	0.05	0.1	0.06	0.8
10	0.06	0.1	0.06	0.9

The changes in the number of association rules extracted under different experimental conditions for both daytime and nighttime datasets are illustrated in Figure 5, respectively. It is evident from the results that, across all parameter settings, the improved Apriori algorithm consistently produces a greater number of association rules compared to the standard algorithm. This indicates that the enhanced version is more effective in capturing nuanced relationships between causative factors in construction fall accidents, thereby offering more comprehensive support for accident prevention efforts.

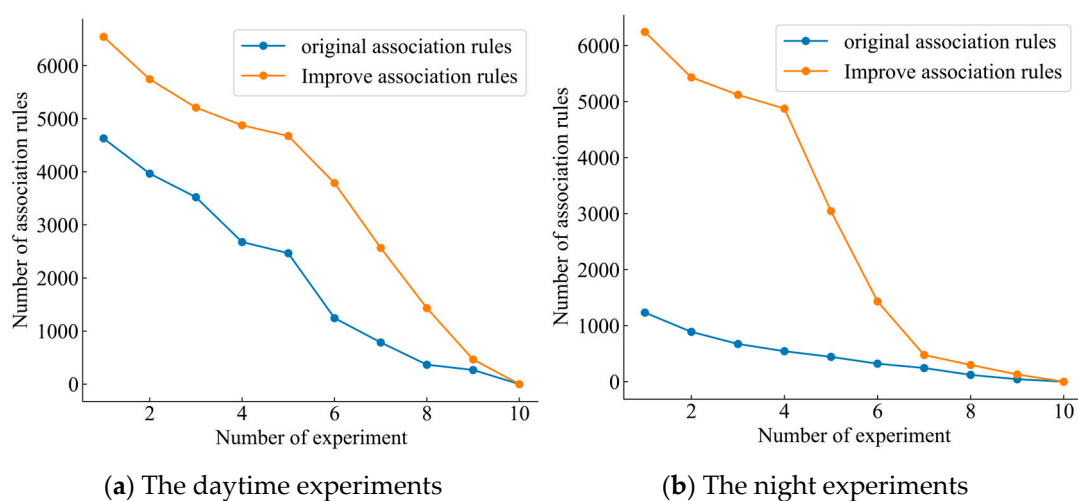


Figure 5. Association rule acquisition experiment: (a) The daytime experiments; (b) The night experiments.

In addition to the overall number of rules extracted, further analysis was conducted to determine whether the improved algorithm identifies a higher proportion of rules containing actual accident-related information. The results suggest that both parameter settings and construction time (day/night) influence the richness of extracted rules. The proportion of rules directly related to fall-from-height accident information varies across different experiment settings.

As shown in Figure 6, the improved algorithm consistently achieved a higher percentage of meaningful associations under most settings. For instance, in Experiment 5 (daytime) and Experiment 4 (nighttime), the proportion of rules containing accident-related information increased by 23.0% and 26.2%, respectively. These findings demonstrate that the improved Apriori algorithm effectively enhances the ability to uncover accident-relevant patterns, which holds significant practical value for accident analysis and risk mitigation in the construction industry.

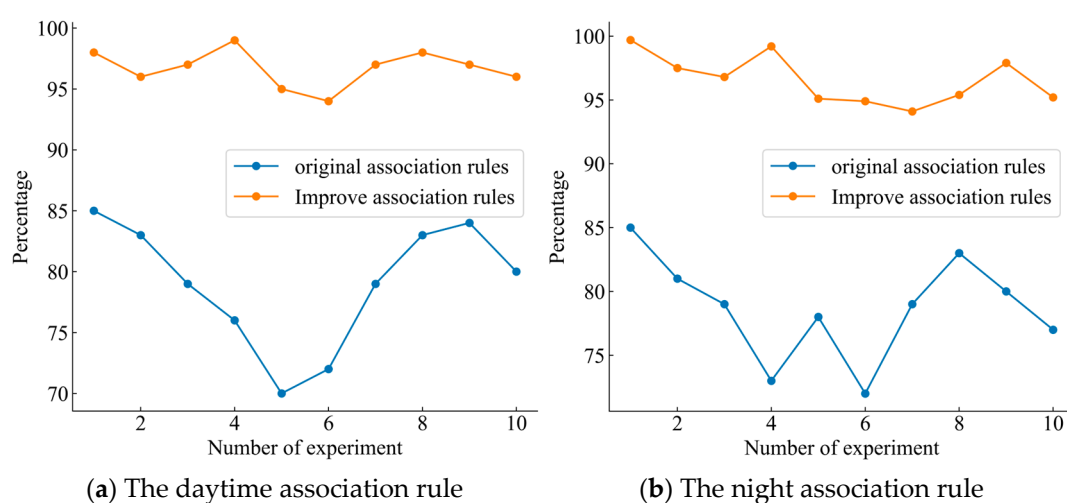


Figure 6. Association rules ratio: (a) The daytime association rule; (b) The night association rule.

4.2. Countermeasures and Suggestions for Preventing Falls from Height in Construction

To better understand the critical causal factors identified through text mining, the study converted the spatial vector model into a word cloud visualization, as shown in Figure 7. The word cloud effectively highlights the most significant keywords, with the size and prominence of each word correlating to its importance in contributing to fall-from-height accidents. Larger and more centrally located words indicate the primary risk factors that must be addressed.

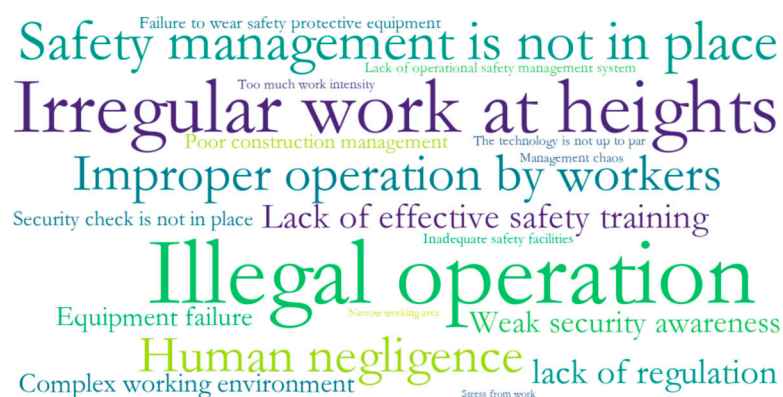


Figure 7. The fall-from-height accident causes word cloud.

Based on the insights derived from LDA topic modeling and association rule mining, a comprehensive analysis of the causes and patterns of fall-from-height incidents in construction was conducted. To mitigate these risks, the following targeted countermeasures are proposed:

- **Enforcing a Robust Safety Culture with Zero Tolerance for Unsafe Practices**

While a zero-tolerance policy toward unsafe behaviors is commonly recommended, a deeper examination reveals that mere policy enforcement is insufficient. It is essential to foster a culture where safety is integrated into the everyday workflow, rather than solely relying on external supervision. Research has shown that enforcement can only be effective if workers internalize safety as a core responsibility. This can be achieved through a combination of behavior-driven safety programs, peer reporting systems, and clear, transparent disciplinary protocols. Additionally, the role of management in promoting safety culture must not be limited to supervisory functions but should extend to active participation in safety decision-making. Effective safety leadership directly influences compliance and can reduce unsafe behaviors by promoting a shared sense of accountability.

- **Targeted and Continuous Safety Training with a Focus on Behavioral Change**

Safety training programs, although standard practice, often fail to produce lasting behavior change. This gap can be bridged by transitioning from generic, one-time training sessions to ongoing, behavior-based safety education. Training should focus not only on the correct use of protective equipment but also on the cognitive aspects of risk recognition and hazard avoidance. Virtual reality (VR) or augmented reality (AR) simulations could be employed to simulate fall hazards in a controlled environment, helping workers to better understand and internalize the risks. Furthermore, personalized training plans, considering the diverse backgrounds and learning speeds of workers, would increase the effectiveness of these programs.

- **Redesigning Site Management Systems to Incorporate Real-Time Risk Monitoring**

Current safety management systems often rely heavily on periodic inspections and static safety protocols. However, accidents frequently occur between inspections or during lapses in safety enforcement. To address this, the adoption of real-time monitoring systems using IoT (Internet of Things) sensors and AI-driven analytics can drastically improve situational awareness. These systems can provide immediate feedback to workers and supervisors, enabling them to take corrective actions before accidents occur. For example, AI-powered drones or wearable devices could alert site managers in real-time if a worker is at risk of falling or not using proper safety equipment, preventing accidents before they happen.

- **Innovating Safety Equipment Design and Maintenance Protocols**

Safety equipment (such as harnesses, safety nets, and guardrails) is essential but often inadequately maintained or improperly used. A more advanced approach would be to introduce IoT-enabled safety devices that can communicate with central monitoring systems to ensure that equipment is always in optimal condition. Wearable safety equipment, such as smart harnesses that detect improper fitting or malfunction, can reduce the risk of accidents caused by equipment failure. Moreover, predictive maintenance models could be implemented, where data from wearables or equipment is analyzed to predict failure before it happens, significantly improving overall safety.

- **Establishing Comprehensive, Data-Driven Emergency Response Plans**

While most construction sites have emergency response plans, these plans are often generic and do not account for the dynamic nature of fall accidents. A more effective approach would be to develop customized emergency protocols that account for specific site layouts, weather conditions, and worker skill levels. The use of drones or automated

rescue systems, such as robotic arms that can retrieve fallen workers quickly, should be explored. Moreover, integrating data-driven simulation systems that predict accident scenarios could help refine response plans, ensuring quicker and more coordinated actions during an emergency.

- **Strengthening Regulatory Oversight and Data Transparency**

Although inspections and compliance monitoring are routine in construction safety, a more robust and transparent regulatory framework is needed. Regulatory agencies should leverage big data analytics to not only monitor real-time compliance but also analyze historical accident data to predict high-risk patterns. Furthermore, public reporting systems should be enhanced to include real-time incident data, helping to increase awareness and accountability at the industry level. Penalties for non-compliance should also be designed to be more stringent and directly tied to accident rates, creating a direct incentive for companies to prioritize safety.

By implementing these advanced countermeasures, construction companies can significantly reduce the occurrence of fall-from-height accidents. Not only will these measures protect the health and safety of workers, but they will also enhance the operational efficiency of construction projects by minimizing delays caused by accidents. Furthermore, creating a culture of safety, driven by real-time data and innovative technologies, will help lay the foundation for a more sustainable and accident-free construction industry. Successful implementation of these strategies requires a collaborative effort from all stakeholders, including government bodies, construction enterprises, and site managers, to create a robust, data-driven safety management framework.

5. Conclusions and Future Work

This study provides a comprehensive analysis of safety accidents caused by falls from height in the construction industry, integrating text mining techniques and association rule modeling to uncover deep-seated causal relationships and inform safety management practices. The main conclusions are as follows:

- A theoretical framework for construction accident analysis based on text mining was established. An improved Apriori algorithm was proposed to enhance the expressiveness of association rules by incorporating temporal variables. This framework offers a structured and scalable approach for identifying key causal patterns in fall-related accidents, providing a foundation for future in-depth analyses of construction safety data.
- By collecting and analyzing 1064 fall-from-height accident reports from 22 provincial regions across China, the study categorized accident types, locations, and personnel roles. This analysis highlighted specific construction activities and environmental contexts that require heightened attention from both site managers and frontline workers.
- Foundational tools for text preprocessing were developed, including a domain-specific dictionary, stop word list, and synonym mapping database. These tools were used to extract key information from unstructured reports. The LDA topic model enabled the identification of dominant accident-related themes, providing a clearer understanding of prevalent risk factors that can inform managerial decision-making.
- Through the application of the improved Apriori algorithm, the study successfully identified latent correlations among multiple accident-inducing factors. The association rules revealed critical links—such as the relationship between inadequate safety training, poor site management, and equipment failure—that contribute to fall-related incidents. These insights can support proactive risk management strategies,

allowing stakeholders to identify and mitigate hazards in the early stages of construction activities.

This research demonstrates the potential of integrating advanced data mining techniques with domain-specific knowledge to enhance the analysis and management of construction safety risks. The methods and findings presented herein provide both theoretical insight and practical value for improving the prevention of fall-from-height accidents in the construction industry.

Future research could expand the dataset by incorporating data from additional regions or countries to improve the generalizability of the findings. Additionally, integrating real-time monitoring systems with the proposed framework could enhance the dynamic identification and mitigation of fall-related risks during construction activities.

Author Contributions: Conceptualization, J.W.; methodology, R.S.; software, R.S.; validation, R.S.; formal analysis, J.W.; investigation, J.W.; resources, J.W.; data curation, R.S.; writing—original draft preparation, R.S.; writing—review and editing, J.W.; visualization, R.S.; supervision, R.S.; project administration, R.S.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: During the preparation of this manuscript, the author used an AI language editing tool for linguistic polishing of the text. The author has reviewed and edited the AI-generated output and takes full responsibility for the content of this publication.

Conflicts of Interest: Author Rongjian Sun is employed by the Haikou Rural Revitalization Investment and Development Co., Ltd. The remaining author 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.

References

1. Kong, F.; Ahn, S. Use of Knowledge Graphs for Construction Safety Management: A Systematic Literature Review. *Information* **2024**, *15*, 390.
2. Taherpour, F.; Biratu Dadi, G.; Keshavarz, M.; Kheiri, P. Assessing the Potential of Machine Learning in Construction Safety: A Systematic Review. In Proceedings of the Construction Research Congress 2024, Des Moines, Iowa, 20–23 March 2024; pp. 945–955.
3. Bai, X.-P.; Zhao, Y.-H. A novel method for occupational safety risk analysis of high-altitude fall accident in architecture construction engineering. *J. Asian Archit. Build. Eng.* **2021**, *20*, 314–325.
4. Rafindadi, A.D.U.; Napiah, M.; Othman, I.; Mikić, M.; Haruna, A.; Alarifi, H.; Al-Ashmori, Y.Y. Analysis of the causes and preventive measures of fatal fall-related accidents in the construction industry. *Ain Shams Eng. J.* **2022**, *13*, 101712.
5. Wang, B.; Wang, Y. Big data in safety management: An overview. *Saf. Sci.* **2021**, *143*, 105414.
6. Mohy, A.A.; Bassioni, H.A.; Elgendi, E.O.; Hassan, T.M. Innovations in safety management for construction sites: The role of deep learning and computer vision techniques. *Constr. Innov.* **2024**, *26*, 551–578.
7. Chen, H.; Yang, M.; Tang, X. Association rule mining of aircraft event causes based on the Apriori algorithm. *Sci. Rep.* **2024**, *14*, 13440.
8. Niu, H.; Yang, X.; Zhang, J.; Guo, S. Risk coupling analysis of causal factors in construction fall-from-height accidents. *Eng. Constr. Archit. Manag.* **2024**, *32*, 6045–6067.
9. Zhang, X.; Zhang, W.; Jiang, L.; Zhao, T. Identification of critical causes of tower-crane accidents through system thinking and case analysis. *J. Constr. Eng. Manag.* **2020**, *146*, 04020071.
10. Liu, C.; Yang, S. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Syst. Appl.* **2022**, *207*, 117991.

11. Sun, J.; Lei, K.; Cao, L.; Zhong, B.; Wei, Y.; Li, J.; Yang, Z. Text visualization for construction document information management. *Autom. Constr.* **2020**, *111*, 103048.
12. Zhong, B.; Xing, X.; Love, P.; Wang, X.; Luo, H. Convolutional neural network: Deep learning-based classification of building quality problems. *Adv. Eng. Inform.* **2019**, *40*, 46–57.
13. Na, X.U.; Ling, M.A.; Liu, Q.; Li, W.; Deng, Y. An improved text mining approach to extract safety risk factors from construction accident reports. *Saf. Sci.* **2021**, *138*, 105216.
14. Qiu, Z.; Liu, Q.; Li, X.; Zhang, J.; Zhang, Y. Construction and analysis of a coal mine accident causation network based on text mining. *Process Saf. Environ. Prot.* **2021**, *153*, 320–328.
15. Ouassif, K.; Ziani, B. Predicting university major selection and academic performance through the combination of Apriori algorithm and deep neural network. *Educ. Inf. Technol.* **2025**, *30*, 333–346.
16. Ding, X.; Wan, H.; Shi, G.; Hong, C.; Liu, Z. Predicting hazard degree levels of metro operation accidents based on ordered constraint Apriori-RF method. *Int. J. Transp. Sci. Technol.* **2025**, *18*, 245–260.
17. Faraji Abdolmaleki, S.; Bello Bugallo, P.M. Advancing Sustainability Assessment of Renewable Energy: A Study on Indicator-to-Framework Methods and Associated Rule Mining. *Sustain. Dev.* **2025**, *33*, 4393–4427.
18. Bao, C.; Tang, Y.; Yang, B.; Wang, X.; Chen, J.; Xiong, H.; Lou, B.; Zhou, S.; Cao, X.; Gu, G. Mining analysis of traffic accident features based on Fp-growth algorithm and Apriori algorithm. In Proceedings of the Fourth International Conference on Computer Vision, Application, and Algorithm (CVAA 2024), Chengdu, China, 11–13 October 2024; pp. 859–863.
19. Kondylidis, N.; Tzelepi, M.; Tefas, A. Exploiting tf-idf in deep convolutional neural networks for content based image retrieval. *Multimed. Tools Appl.* **2018**, *77*, 30729–30748.
20. Wu, Z.; Fang, G. The Research on the Improvement of FP-growth Algorithm. *Artif. Intell. Technol. Res.* **2024**, *2*, 10.18686/aitr.v2i1.3860.
21. Maskat, K.; Isa, M.R.M.; Khairuddin, M.A.; Kamarudin, N.D.; Ismail, M.N. Enhancing The Server-Side Internet Proxy Detection Technique in Network Infrastructure Based on Apriori Algorithm of Machine Learning Technique. *JOIV Int. J. Inform. Vis.* **2025**, *9*, 504–509.
22. Alsubaey, M.; Asadi, A.; Makatsoris, H. A Naive Bayes approach for EWS detection by text mining of unstructured data: A construction project case. In Proceedings of the 2015 SAI intelligent systems conference (IntelliSys), London, UK, 10–11 November 2015; pp. 164–168.
23. Xi, J.; Zhao, Z.; Li, W.; Wang, Q. A traffic accident causation analysis method based on AHP-apriori. *Procedia Eng.* **2016**, *137*, 680–687.
24. Singhal, A.; Gopalakrishnan, K.; Khaitan, S.K. Predicting Budget from Transportation Research Grant Description: An Exploratory Analysis of Text Mining and Machine Learning Techniques. *J. Soft Comput. Civ. Eng.* **2017**, *1*, 89–102.
25. Fan, C.; Mostafavi, A.; Gupta, A.; Zhang, C. A system analytics framework for detecting infrastructure-related topics in disasters using social sensing. In Proceedings of the Workshop of the European Group for Intelligent Computing in Engineering, Lausanne, Switzerland, 10–13 June 2018; pp. 74–91.
26. Liu, M.; Le, Y.; Hu, Y.; Xia, B.; Skitmore, M.; Gao, X. System dynamics modeling for construction management research: Critical review and future trends. *J. Civ. Eng. Manag.* **2019**, *25*, 730–741.
27. Wang, C.; Zheng, X. Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. *Evol. Intell.* **2020**, *13*, 39–49.
28. Abdirad, H.; Mathur, P. Artificial intelligence for BIM content management and delivery: Case study of association rule mining for construction detailing. *Adv. Eng. Inform.* **2021**, *50*, 101414.
29. Chen, S.; Xi, J.; Chen, Y.; Zhao, J. Association mining of near misses in hydropower engineering construction based on convolutional neural network text classification. *Comput. Intell. Neurosci.* **2022**, *2022*, 4851615.
30. Du, W.; Ge, C.; Yao, S.; Chen, N.; Xu, L. Applicability analysis and ensemble application of BERT with TF-IDF, TextRank, MMR, and LDA for topic classification based on flood-related VGI. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 240.
31. Chen, Y.; Lu, G.; Wang, K.; Chen, S.; Duan, C. Knowledge graph for safety management standards of water conservancy construction engineering. *Autom. Constr.* **2024**, *168*, 105873.

32. Kim, H.-J.; Baek, J.-W.; Chung, K. Optimization of associative knowledge graph using TF-IDF based ranking score. *Appl. Sci.* **2020**, *10*, 4590.
33. Kim, S.-W.; Gil, J.-M. Research paper classification systems based on TF-IDF and LDA schemes. *Hum.-Centric Comput. Inf. Sci.* **2019**, *9*, 30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.