buildings

MDPI

Article

Modularity-Driven Keyword Co-Occurrence Network for Mining Statistical Associations in Construction Safety Accidents

Shu Liu [1], Weidong Yan [1], Jian Ma [2,*], Guoqi Liu [3] and Rui Zhang [1]

[1] School of Civil Engineering, Shenyang Jianzhu University, Shenyang 110168, China; liushu@stu.sjzu.edu.cn (S.L.); ywd6441@sjzu.edu.cn (W.Y.); zhangrui@stu.sjzu.edu.cn (R.Z.)
[2] School of Management, Shenyang Jianzhu University, Shenyang 110168, China
[3] School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang 110168, China; liuguoqi@sjzu.edu.cn
* Correspondence: majian@sjzu.edu.cn

Abstract

To address the limitations of traditional construction safety accident analysis, which relies on manually defined causal relationships, requires extensive data annotation, and struggles to identify latent risks from Chinese unstructured texts, this study proposes an unsupervised and data-driven framework, termed CESA-Miner, for mining statistical association patterns among construction safety accidents. The proposed framework adopts a modularity-driven keyword optimization strategy to automatically identify a stable set of risk-related features. Based on this, an accident risk weighted co-occurrence network is constructed, where statistical associations are represented through keyword co-occurrence patterns and network community structures. Community detection algorithms are then applied to identify accident clusters and their underlying relationships. Using a dataset of 1368 official construction accident reports, the results show that the network modularity increases from 0.173 to 0.683, indicating significantly improved structural quality and community separability. In the absence of explicit ground truth, structural quality is evaluated using network modularity as a proxy metric. Compared with conventional clustering-based and embedding-based approaches, the proposed method yields a more structurally distinct network community organization and offers a complementary structure-aware perspective for characterizing accident relationships. The framework enables large-scale intelligent analysis of accident texts without requiring manual annotation, providing data-driven support for latent risk identification and statistical pattern analysis in construction safety.

Keywords: construction safety; accident analysis; adaptive keyword extraction; risk co-occurrence network; text mining; complex network analysis; latent feature association mining; statistical risk pattern mining

Academic Editor: Jaewook Jeong

Received: 13 March 2026
Revised: 5 April 2026
Accepted: 6 April 2026
Published: 7 April 2026

1. Introduction

Construction safety accident reports are typically presented as unstructured text. Analyzing such reports essentially involves extracting and examining risk information as well as the underlying association structures embedded in accident narratives. Early studies mainly relied on theoretical frameworks, such as accident causation models [1], the Suraji model [2], the Swiss Cheese Model (SCM) [3], and the 2–4 model [4], to interpret accident mechanisms from a causal perspective. In parallel, some researchers have investigated intrinsic relationships among accidents from a data-driven perspective [5–8]. For instance, Jesus A. Carrillo-Castrillo identified relationships between accident causes and construction

stages based on investigation reports [5], while Javier Irizarry developed a multi-layer risk network model incorporating causal factors and accident types to analyze accident evolution in complex construction environments [6]. However, these approaches often depend on predefined causal assumptions or expert knowledge, which may introduce subjectivity and limit their applicability to large-scale unstructured text data.

With the continuous accumulation of construction safety data, data-driven approaches based on computational techniques have been increasingly adopted for accident analysis. Methods such as decision trees [9], random forests [10], and the Apriori algorithm [11] have been applied for classification and pattern mining, while probabilistic graphical models, including Bayesian networks, have been utilized for risk prediction and causal inference [12]. Moreover, recent advances in natural language processing (NLP), such as association rule mining, topic modeling (e.g., LDA), and unsupervised text analysis, have been widely employed to extract semantic patterns and co-occurrence-based structural patterns from construction accident reports [13–18]. Despite their effectiveness in handling large-scale data, many of these methods still rely on manual annotation or predefined structures, which may limit model generalization and practical efficiency [19].

As a text modeling approach that does not require extensive labeled data, keyword co-occurrence analysis has been widely used to explore co-occurrence-based structural patterns in textual data [20–22]. Co-occurrence relationships can capture statistical associations between terms and provide a foundation for constructing complex networks. However, most existing approaches treat keyword extraction and network construction as two independent stages. Typically, keywords are first extracted using methods such as TF-IDF or TextRank, and then a network is constructed based on their co-occurrence relationships. This sequential strategy, characterized by a separation between semantic feature extraction and structural modeling, may result in network representations that do not fully reflect the underlying semantic organization of the text.

It is worth noting that although unsupervised methods eliminate the need for manual annotation, their outcomes are not inherently free from bias. The keyword extraction process may still be influenced by term frequency distributions and corpus characteristics, potentially introducing statistical deviations. Therefore, improving the alignment between semantic representation and structural modeling under unsupervised conditions remains a key challenge.

In recent years, complex network-based methods have been explored for modeling construction safety accidents; however, several limitations persist. On the one hand, many network models rely on predefined causal relationships or specific application scenarios, limiting their generalizability to broader construction safety analysis tasks [6,11,23,24]. On the other hand, although techniques such as LDA, association rule mining, and large language models have been applied to accident text analysis [7,15–18], systematic studies on keyword co-occurrence network modeling for Chinese construction safety accident texts remain limited, and the relationship between semantic representation and structural properties has not been fully examined. In addition, existing studies often overlook an important issue: how semantic feature selection influences network structure quality, and how structural information can, in turn, guide semantic representation.

From the perspective of complex network theory, modularity is a widely used metric for evaluating community structure quality, reflecting the density of intra-community connections and the sparsity of inter-community links. In the context of construction safety accident text modeling, if keyword selection effectively captures co-occurrence-based structural patterns, the resulting co-occurrence network is expected to exhibit clear community structures with relatively high modularity. Accordingly, modularity can serve as a structural indicator for assessing the effectiveness of semantic feature selection. Based

on this idea, optimizing the keyword set by maximizing modularity provides a structural feedback signal for selecting a stable keyword subset, rather than for optimizing semantic representation directly.

Based on the above analysis, this study proposes a structure-aware semantic network co-optimization framework for construction safety accident analysis. The proposed framework is built upon a keyword co-occurrence network and introduces modularity as a global optimization objective to enable adaptive keyword selection and structural refinement. This design facilitates the identification of latent associations among accidents under unsupervised conditions.

The main contributions of this study are as follows:

(1) A structure-driven keyword optimization mechanism is proposed, in which network modularity is used as a global objective to achieve coordinated optimization between semantic feature selection and network structure;

(2) A weighted keyword co-occurrence network model for Chinese construction accident texts is constructed, enabling data-driven accident association modeling without relying on predefined causal relationships;

(3) A community-structure-based analysis method is developed to identify latent associations among accidents and to reveal potential data-driven latent risk association patterns between different types of accidents, providing a new perspective for systematic construction safety risk analysis.

2. Methods

2.1. Construction Accident Latent Association Mining Framework (CESA-Miner)

To explore latent associations among construction safety accidents, this study proposes a construction accident latent association mining framework (CESA-Miner). CESA-Miner is a structure-aware semantic network co-optimization framework that integrates natural language processing, complex network modeling, and validation mechanisms to uncover potential association patterns among construction safety accidents. In contrast to traditional approaches that treat keyword extraction and network construction as two independent steps, the proposed framework introduces a network structure parameter search and selection process, where topological characteristics of the network are used to guide semantic feature selection. In this way, coordinated optimization between semantic representation and structural modeling can be achieved.

Specifically, the method is built upon the risk-related semantic information extracted from accident texts. A keyword co-occurrence network is constructed to represent the latent relationships among accidents in a structured form. Meanwhile, modularity-based trend analysis is employed as a structural reference to adaptively refine the keyword set. This process not only enables the network representation of semantic information in accident texts but also improves the discriminative capacity of keyword features through structural constraints, allowing the constructed network to better reflect statistical co-occurrence patterns among accidents. These associations are derived from statistical co-occurrence patterns and do not imply causal relationships or actual risk propagation mechanisms.

The overall workflow of the study consists of two stages:

(1) Text mining stage: The accident descriptions are used as input, with words as the basic analytical units. Through text preprocessing, keyword extraction, and keyword filtering, a keyword set representing accident risk features is constructed.

(2) Network analysis stage: Construction safety accidents are treated as network nodes, and co-occurrence relationships of keywords between accident texts are used as edges to construct an undirected weighted accident association network. Network structure

analysis, community detection, and validation are further conducted to identify potential clustering structures and association patterns among accidents.

Based on the above methodology, a network model for construction safety accidents is established. The main workflow includes text data preprocessing, accident keyword extraction and filtering, keyword co-occurrence matrix construction, accident network modeling, network structure analysis and community detection, as well as a validation module, as illustrated in Figure 1. Unlike conventional linear pipelines, the proposed framework introduces a structure-aware optimization mechanism, where network structural quality (e.g., modularity) is used as feedback to guide keyword selection. This establishes an implicit feedback loop between semantic feature extraction and network topology optimization.

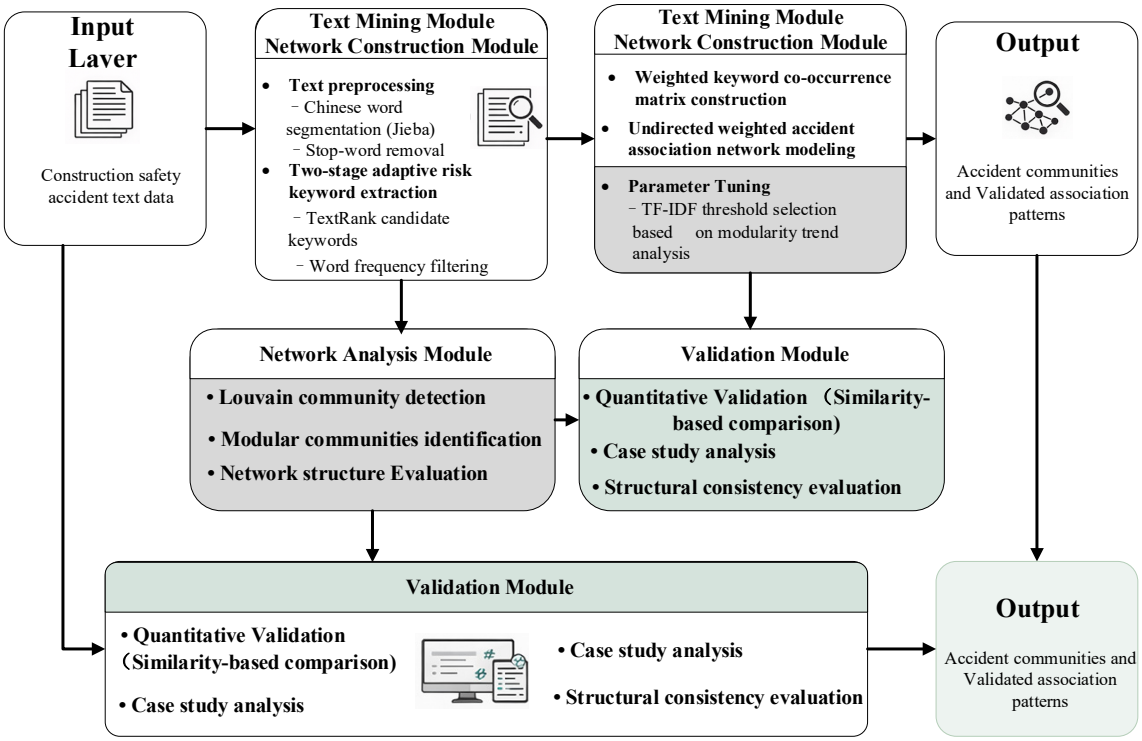


Figure 1. Construction engineering safety accident network model. Note: The framework represents a parameter search and selection process, not an iterative closed-loop optimization. Modularity is used to evaluate and select a stable keyword subset, rather than to feed back into iterative updates.

It should be emphasized that the main contribution of this study does not lie in the application of individual algorithms (e.g., TextRank, TF-IDF, or the Louvain algorithm), but rather in the development of a structure-aware co-optimization mechanism between semantic representation and network structure. By leveraging network structural trend analysis to globally optimize the keyword set, a feedback relationship is established between semantic feature selection and network structure quality, thereby overcoming the limitations of traditional sequential processing pipelines.

2.2. Data Collection

The data used in this study were collected from multiple official and publicly available sources in China, including construction safety accident bulletins released by emergency management departments, housing and urban–rural development authorities, and local government agencies. These reports are typically issued by regulatory institutions during accident investigation or notification processes and contain information such as accident descriptions, causes, and related background details, which provide a reliable basis for analysis.

After systematic collection, cleaning, and deduplication, a dataset consisting of 1368 independent accident cases was constructed. Each record includes not only structured attribute information but also a complete accident description text, which serves as the primary input for subsequent semantic analysis. To improve data organization and reproducibility, the dataset structure is standardized, and the main fields are presented in Table 1.

Table 1. Structure of the construction safety accident dataset.

Field Name	Data Type	Example	Description
Accident_ID	String	A001	Unique identifier of the accident
Date	Date	12 May 2021	Date of the accident
Location	String	Shenyang, Liaoning Province	Accident location
Weather	String	Cloudy	Weather condition
Season	String	Summer	Seasonal information
Project_Type	String	High-rise building	Type of construction project
Company_Qualification	String	First-grade qualification	Qualification of the contractor

Among these fields, the accident description text (Description) serves as the primary input for natural language processing and keyword extraction in subsequent analyses, while the other structured attributes are mainly used for data organization and supplementary analysis.

To enable standardized management and efficient utilization of the data, a construction safety accident data management system was established to encode, structure, and categorize the original reports, thereby improving data consistency and retrievability. It should be noted that although the data are derived from official public reports, certain limitations and uncertainties may still exist. For example: (1) some accidents may be underreported or incompletely documented; (2) reporting standards may vary across regions and time periods; and (3) the description texts may exhibit inconsistencies and variations in expression.

To mitigate the potential impact of these issues, the following strategies were adopted: (1) integrating data from multiple sources to improve coverage; (2) using a relatively large sample size (N = 1368) to reduce the influence of individual biases; and (3) focusing on textual semantic information as the primary analytical basis rather than relying solely on structured labels, thereby reducing the effect of structural bias on the results.

In addition, it should be emphasized that the constructed dataset is a static historical dataset, primarily used to explore latent semantic associations among accidents. Therefore, temporal dynamics (e.g., policy changes or seasonal variations) and spatial heterogeneity (e.g., regional differences) are not explicitly incorporated into the model in this study, and these aspects will be further discussed in the Discussion Section.

Given that accident reports are mainly presented in unstructured text form, their latent information needs to be extracted through appropriate techniques. Accordingly, natural language processing (NLP) methods are introduced to analyze accident description texts, extract risk-related semantic features, and provide a foundation for the subsequent construction of the keyword co-occurrence network.

2.3. Text Preprocessing

To improve the quality of text representation and reduce the impact of noise in unstructured construction safety accident reports, a multi-stage text preprocessing pipeline is developed in this study. The pipeline includes text normalization, tokenization, stop-word removal, and semantic standardization.

First, the raw accident reports are cleaned by removing irrelevant symbols, punctuation marks, and redundant whitespace. On this basis, domain-specific knowledge in construction safety is incorporated to normalize terminology. Expressions with similar meanings are standardized into unified terms. For example, terms such as "scaffold" and "scaffolding system" are mapped to a consistent representation to reduce inconsistencies in semantic expression. The standardization rules are constructed based on a vocabulary list derived from the domain literature and expert knowledge. Second, a Chinese word segmentation tool suitable for processing Chinese corpora is applied to ensure that technical terms can be properly identified and segmented. Common stop words and low-information terms are then removed to reduce noise and improve the quality of the textual data. Furthermore, to address synonymy and semantic variation, a hybrid standardization strategy combining dictionary mapping and semantic similarity matching is introduced. Specifically, domain-specific terms are first normalized using a curated domain dictionary. Then, pretrained language models are used to generate vector representations of words or phrases, and cosine similarity is calculated to identify semantically similar expressions. When the similarity exceeds a predefined threshold, the terms are merged into a unified representation. This approach helps mitigate lexical variability and improves semantic consistency across the corpus. Finally, the processed text is converted into structured token sequences, which serve as input for subsequent analyses, including TF-IDF-based weighting, TextRank-based keyword extraction, and the construction of the keyword co-occurrence network.

2.4. Two-Stage Adaptive Risk Keyword Extraction Method

This study proposes a two-stage adaptive risk keyword extraction method for construction safety accident texts, aiming to automatically identify core keywords that can stably represent accident risk features from unstructured text. Considering the characteristics of accident reports, including high diversity in expression, apparent semantic redundancy, and uneven keyword distribution, a two-stage mechanism consisting of "candidate generation" and "stability-based filtering" is introduced on the basis of traditional unsupervised keyword extraction methods. This mechanism improves keyword quality from both semantic relevance and cross-text consistency perspectives.

In contrast to conventional single-stage keyword extraction methods that rely on a single ranking criterion, the proposed approach incorporates a post-filtering step to further refine the initially extracted results. This allows the method to maintain semantic relevance while improving the representativeness and robustness of keywords across the corpus. Specifically, a graph-based ranking method, TextRank, is first applied to construct a word co-occurrence graph from the internal structure of each document. By leveraging local semantic relationships, a candidate keyword set is generated.

Subsequently, global term frequency statistics are introduced to analyze the distribution of candidate keywords across the entire corpus. Keywords that exhibit stable occurrence patterns are selected, while terms with sporadic occurrences or limited representativeness are filtered out. This process helps reduce the influence of incidental terms and local noise.

The main advantage of the proposed two-stage mechanism lies in the complementary roles of its components. On the one hand, TextRank captures local semantic structures within individual texts, making it suitable for unstructured accident descriptions. On the other hand, the frequency-based filtering imposes consistency constraints from a global statistical perspective, ensuring that the final keyword set maintains both semantic relevance and statistical stability. By integrating local semantic modeling with global statistical constraints, the proposed method enables adaptive optimization of keyword extraction under an unsupervised setting.

The resulting set of core accident keywords not only better reflects the semantic characteristics of individual texts but also exhibits higher consistency and comparability across samples, thereby providing a reliable foundation for the subsequent construction of the keyword co-occurrence network and risk association analysis.

2.4.1. Initial Keyword Screening Based on TextRank

TextRank is an unsupervised keyword extraction algorithm based on the graph ranking principle. Its fundamental idea is to represent words in a text as nodes in a graph and construct a word network based on co-occurrence relationships within a fixed window. The importance of each node is then determined through iterative computation over the graph. After constructing the lexical network, the importance score of each node is computed using the following formula [25]:

$$\text{Textrank}(\text{node}_i) = (1-d) + d \times \sum_{j=1}^{n} \frac{\text{Textrank}(\text{node}_j)}{\text{OutDegree}(\text{node}_j)} \quad (1)$$

where d represents the damping factor and Textrank scores are obtained through iterative updates. To ensure the stability and reproducibility of the results, the key parameters of the TextRank algorithm are set as follows: the co-occurrence window size is 2, the maximum number of iterations is 100, the convergence threshold is 0.0001, and the damping factor is 0.85.

Under fixed parameter settings, the node importance scores are iteratively updated, and words are ranked according to their TextRank scores. The top-ranked terms are selected as the candidate keyword set.

2.4.2. Secondary Keyword Filtering Based on Term Frequency

After obtaining the candidate keyword set, a term frequency-based filtering mechanism is introduced to further improve the representational capability of keywords for accident text features. Specifically, the entire accident corpus is used as the statistical scope to calculate the occurrence frequency of each candidate keyword across different accident texts and to analyze its distribution at the corpus level. Based on this analysis, a frequency-based filtering strategy is applied to refine the candidate keywords. The selection rule retains keywords that exhibit relatively high occurrence frequency across multiple accident texts (e.g., appearing in at least a certain number of documents or ranking among the higher-frequency terms), thereby ensuring cross-text consistency and representativeness of the selected keywords.

This filtering step helps reduce the influence of incidental expressions and text-specific characteristics and removes low-frequency, unstable, or less informative terms. As a result, a more robust set of core accident keywords is obtained, which is subsequently used for the construction of the keyword co-occurrence network and further risk association analysis.

2.5. TF–IDF Adaptive Optimization Algorithm Based on Modularity Maximization

This study proposes a TF-IDF-based adaptive optimization method guided by modularity maximization, aiming to automatically identify a stable and representative feature subset from the candidate keyword set, thereby improving the representation of keywords in accident semantic association modeling. By introducing a network structure evaluation metric, the proposed method compares keyword sets generated under different TF-IDF thresholds, enabling adaptive adjustment of the keyword selection process.

Specifically, the candidate keywords are first ranked according to their TF-IDF values. The TF-IDF measure evaluates the discriminative capability of a term by considering both

its term frequency (TF) within an individual document and its inverse document frequency (IDF) across the entire corpus. The formulation of TF-IDF is given as follows [26–28]:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \quad (2)$$

Term frequency (TF) represents the occurrence frequency of a term within an individual document, while inverse document frequency (IDF) reflects the rarity of the term across the corpus. A higher TF-IDF value indicates that the term has a stronger capability to represent the semantic characteristics of a specific document.

In the keyword selection stage, a fixed threshold is not adopted. Instead, a set of candidate thresholds (or Top-K candidate sets) is defined to generate multiple keyword subsets under different selection criteria. For each subset, a keyword co-occurrence network is constructed, and the corresponding modularity value is calculated. Modularity is used as a structural evaluation indicator to assess the topological separation and clustering stability of the co-occurrence network, rather than as a measure of semantic correctness or validity. The identified associations are derived from statistical co-occurrence patterns and should not be interpreted as reflecting causal relationships or risk propagation processes.

It should be noted that modularity is employed in this study as an indicator of internal structural consistency. Its role is to characterize the aggregation patterns of keyword co-occurrence relationships from a network perspective, rather than to serve as a direct measure of semantic correctnessity. Therefore, modularity is employed as a data-driven structural selection criterion to identify a stable keyword subset, not to validate semantic correctness or causal significance, among multiple candidates, rather than to determine an absolute semantic optimum. It should be clarified that modularity serves only as a topological stability criterion for parameter selection, not as evidence of semantic superiority or causal validity. This design avoids self-fulfilling optimization within the proposed framework.

2.6. Construction of the Accident Risk Weighted Co-Occurrence Matrix

To quantitatively characterize the associations between accidents, this study takes individual accident samples as the basic analytical units and constructs an accident–accident keyword co-occurrence matrix. This process transforms semantic relationships embedded in unstructured texts into a structured and computable representation, thereby providing a unified data foundation for subsequent network modeling. Suppose that the study contains a total of N independent accident samples, denoted as the set $U = \{u_1, u_2, \ldots, u_N\}$. The filtered effective keyword set is denoted as $K = \{k_1, k_2, \ldots, k_M\}$. Based on the co-occurrence distribution of keywords across accident samples, a co-occurrence matrix is constructed, and its formulation is defined as follows:

(1) The rows and columns of the matrix correspond one-to-one to the independent accident samples, where both the i-th row and the j-th column represent the i-th and j-th accident samples, respectively.

(2) The off-diagonal elements represent the number of shared effective keywords between two accident samples. The magnitude of these values reflects the textual similarity and potential association strength between the corresponding accidents.

(3) The diagonal elements are set to zero to eliminate the influence of self-co-occurrence on the network structure.

The above formulation maps semantic relationships between accidents into structured relationships based on keyword co-occurrence, thereby enabling the quantitative representation of potential associations.

It should be noted that the edge weight defined in this study is based on the number of shared keywords, which can be regarded as a discrete feature-based similarity measure. Compared with cosine similarity, BM25, or embedding-based semantic similarity methods, this approach is relatively simple in representation. However, it offers several advantages in the present context. First, the keyword set has undergone a two-stage filtering process and TF-IDF-based optimization, ensuring strong semantic discriminative capability; thus, keyword overlap can partially reflect similarity in underlying risk characteristics between accidents. Second, co-occurrence frequency provides good interpretability, as it directly indicates the number of shared risk factors between accident cases, which is valuable for safety risk analysis and management decision-making. Furthermore, given the relatively limited dataset size and the domain-specific nature of textual expressions, embedding-based methods may introduce additional model-dependent biases. In contrast, the co-occurrence-based statistical approach avoids reliance on pretrained models, thereby maintaining transparency and stability in the analysis process. Therefore, the use of shared keyword counts as edge weights ensures a balance between methodological simplicity and the ability to effectively support the construction of accident association networks and subsequent community structure analysis.

The edge definition based on shared keyword count is adopted for its high computational efficiency and strong interpretability in engineering practice. It directly quantifies the similarity of risk characteristics between accidents in a transparent and easy-to-understand manner. Advanced representations such as semantic embeddings and deep similarity metrics will be explored as future work to further enhance semantic modeling capability.

2.7. Undirected Weighted Accident Association Network Modeling and Community Detection

Based on the accident–keyword co-occurrence matrix, the semantic relationships between texts are transformed into a complex network topology, thereby constructing an undirected weighted network for construction safety accidents. Through network structure analysis and community detection methods, latent association patterns and clustering characteristics among accidents can be identified, providing a structured basis for subsequent accident risk pattern analysis.

2.7.1. Mathematical Definition of the Network

The construction safety accident association problem is represented as a complex network. The undirected weighted accident network is defined as:

$$G = (V, E, W) \tag{3}$$

where V denotes the set of nodes, with each node uniquely corresponding to an individual accident event; E represents the set of edges, where an undirected edge is formed between two nodes if the corresponding accidents share common keywords; and W denotes the set of edge weights.

The edge weight is defined as the number of shared keywords between two accident texts, which is used to quantify the strength of semantic association. A higher weight indicates a higher degree of similarity in risk characteristics or accident types between the corresponding accidents. In this network, edges represent semantic similarity relationships derived from shared risk-related keywords. Therefore, the network exhibits both undirected and weighted properties. On the one hand, undirected edges reflect the symmetry of associations between accidents; on the other hand, edge weights capture variations in association strength, allowing the network to retain differences in semantic relationships among accident texts.

2.7.2. Network Visualization Layout Method

To visually present the structural characteristics of the network, the Fruchterman–Reingold force-directed layout algorithm is employed for spatial arrangement. This algorithm simulates physical forces between nodes, where connected nodes exert attractive forces, while all nodes exert repulsive forces on each other. Through iterative adjustments, the system reaches a balanced state in which strongly connected nodes are positioned closer together, while weakly connected nodes are placed farther apart.

This layout approach facilitates the visualization of overall network topology, the identification of central nodes, and the observation of potential community structures. It has been widely applied in the visualization of complex networks [29].

2.7.3. Community Detection Algorithm

To identify latent structural groups in the network, the Fast Unfolding Algorithm [30], also known as the Louvain community detection algorithm, is adopted to partition the accident association network. This algorithm aims to maximize modularity through an iterative optimization process, in which nodes and communities are progressively merged to improve the quality of the community structure.

Specifically, each node is initially assigned to an individual community. Then, through a local optimization strategy, each node is moved to a neighboring community if such a move leads to an increase in modularity. After completing one round of node-level optimization, the detected communities are aggregated into new nodes to construct a reduced network. This process is repeated iteratively until no further improvement in modularity can be observed.

The Louvain algorithm provides a balance between computational efficiency and community partition quality and has been widely applied in large-scale complex network analysis [30].

It should be noted that, in addition to the Louvain algorithm, other methods such as the Leiden algorithm and the Infomap algorithm are also commonly used for community detection. Compared with Louvain, the Leiden algorithm improves community connectivity from a theoretical perspective, while Infomap partitions networks based on information flow compression. However, given the network scale and weighted structure in this study, the Louvain algorithm is considered appropriate in terms of computational efficiency, stability, and interpretability.

To further examine the robustness of the community detection results, additional experiments are conducted by comparing different algorithms (e.g., Leiden and Infomap). The results show that these methods produce largely consistent community structures, with stable core accident clusters and key connections. This suggests that the proposed framework provides reliable community identification results under different algorithmic settings.

2.7.4. Network Structural Evaluation Metrics

To evaluate the effectiveness of community detection results, modularity (Q) is adopted as the primary metric for assessing network community structure. Modularity measures the difference between the density of connections within communities and that between different communities. The formulation is given as follows:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j) \tag{4}$$

where A_{ij} denotes the adjacency matrix element of the network, taking a value of 1 if there is a link between nodes i and j and 0 otherwise; k_i and k_j represent the degrees of nodes i and j, respectively; m is the total number of edges in the network; c_i and c_j indicate the

community indices of nodes i and j; and $\delta(c_i, c_j)$ is an indicator function, which equals 1 if the two nodes belong to the same community and 0 otherwise.

The value of modularity typically ranges from -1 to 1. A modularity value of $Q > 0$ indicates the presence of some degree of community structure within the network, while $Q > 0.3$ is generally considered to reflect a significant and interpretable community structure [31].

2.8. Adaptive Network Structure Optimization Strategy

To achieve coordinated improvement between keyword selection and network structure optimization, this study proposes a parameter search-based adaptive optimization strategy to automatically identify an appropriate keyword feature set. Specifically, after keyword extraction and TF-IDF weighting, a candidate keyword pool is first constructed based on keyword importance, and multiple keyword subsets are generated under different selection criteria. A hierarchical filtering strategy based on TF-IDF ranking is adopted, in which keywords are sorted in descending order of their TF-IDF values, and a series of candidate thresholds is defined to construct keyword sets of varying sizes.

In practice, candidate keyword subsets are constructed by selecting the Top-K keywords according to the TF-IDF ranking, where K varies within a predefined range (alternatively, a Top-p percentile strategy can be applied, such as selecting the top 10%, 20%, . . ., 60% of keywords). For each candidate keyword subset, a corresponding accident–accident co-occurrence network is constructed, and its modularity value (Q) is computed as the structural evaluation metric. By traversing all candidate subsets, a set of modularity values is obtained, and the keyword subset that yields the highest modularity is selected as the final result. This process can be formulated as:

$$K^* = \arg\max_{K \in \kappa} Q(G_K) \tag{5}$$

where K denotes the space of candidate keyword subsets and G_K represents the accident network constructed based on the keyword subset K.

It should be noted that this optimization process is essentially a parameter search-driven adaptive selection mechanism, rather than manual tuning based on empirical experience. The keyword selection threshold is determined through systematic traversal of the predefined candidate space, and the final result is guided by network structural characteristics (i.e., modularity), which supports reproducibility and reduces subjectivity.

In addition, to reduce potential bias associated with relying on a single metric, supplementary analyses are conducted by considering variations in network density and the number of detected communities. These additional indicators are used to support the rationality of the selected threshold and to further examine the stability of the optimization results.

Through this strategy, the influence of high-frequency but less informative terms is reduced, while keywords with stronger discriminative capability are retained. As a result, the constructed accident association network exhibits clearer structural organization and more interpretable community patterns.

3. Results

3.1. Dataset Overview

The dataset used in this study was collected from the official sources described above and consists of 1368 construction safety accident reports. Each record corresponds to an individual accident event and includes both unstructured textual descriptions and structured contextual attributes, such as time, location, and project-related information.

After data collection, all accident records were organized and stored in a structured database through a dedicated data management platform. This platform enables standardized classification, management, and retrieval of accident information, providing a consistent data foundation for subsequent analysis, modeling, and validation.

Within the dataset, accident description texts serve as the primary analytical source, as they contain detailed information on accident processes, causal factors, and response measures. Since these descriptions are written in natural language, systematic text preprocessing and key semantic feature extraction are required prior to further analysis.

Following the preprocessing procedure described in Section 2.2, all accident texts were transformed into structured keyword representations. This transformation preserves core semantic information while enabling data standardization and vectorization, thereby providing essential input for subsequent keyword co-occurrence analysis and accident network construction.

To provide an overview of the dataset and the textual characteristics after preprocessing, the main statistical information of the collected accident reports is summarized in Table 2.

Table 2. Statistical characteristics of the accident dataset.

Item	Description	Value
Data source	Official Chinese government websites reporting construction safety accidents	Government public reports
Total accident reports	Number of collected accident records	1368
Data format	Accident descriptions and contextual attributes	Text + structured attributes
Initial keywords extracted	Keywords obtained using the TextRank algorithm	5100
Filtered keywords	Keywords retained after frequency filtering	1647
Keyword frequency threshold	Minimum occurrence frequency for keyword retention	$\geq$2

As shown in Table 2, the collected dataset contains 1368 accident reports, from which 5100 initial keywords were extracted using the TextRank algorithm. After applying a frequency filtering threshold of two occurrences, 1647 keywords were retained for subsequent co-occurrence analysis and accident network construction.

Example of the Keyword Extraction Process

To improve the interpretability of the proposed multi-stage text processing pipeline, a representative accident report from the dataset is selected to illustrate the keyword extraction and filtering procedure.

(1) Original accident text.

After slight simplification, the original accident report is as follows: “On August 25, a sidewalk under construction in Guiyang collapsed, resulting in two construction workers being buried. Rescue personnel arrived promptly and carried out rescue operations. The trapped workers were successfully rescued and sent to the hospital for treatment, with no life-threatening injuries.”

(2) Tokenization.

After processing using the Jieba segmentation tool, the following tokens are obtained: [construction, sidewalk, collapse, rescue, personnel, burial, hospital, treatment, life-threatening].

(3) TextRank keyword extraction.

The initial keyword set extracted by TextRank is: [collapse, construction, rescue, personnel, burial, hospital, life-threatening].

(4) Frequency-based filtering.

After applying the frequency threshold (≥2), the filtered keywords are: [collapse, rescue, burial, personnel].

(5) TF-IDF-based filtering.

With a TF-IDF threshold of 6, the final keyword set is: [rescue, life-threatening].

This example illustrates the step-by-step transformation from raw accident text to the final keyword set. It can be observed that the number of keywords is reduced through the multi-stage process, while key semantic elements related to the accident are retained, providing structured input for subsequent accident network construction.

3.2. Initial Accident Network Construction and Community Structure Analysis

Based on the co-occurrence relationships of accident keywords, an undirected weighted network model for construction safety accidents is constructed. First, an accident–keyword co-occurrence matrix is established according to the results of text analysis, where both rows and columns represent individual accident events, and each element indicates the number of shared keywords between two accident texts. This matrix provides a statistical representation of the associations among accident events and serves as the foundation for network construction.

In the constructed network, each node represents an independent accident event, while edges indicate keyword co-occurrence relationships between accident texts. The edge weight reflects the frequency of shared keywords between two accidents. Through this process, unstructured accident texts in natural language are transformed into a structured representation suitable for complex network analysis. Based on the full dataset, keywords are extracted using the TextRank algorithm and further filtered using frequency-based criteria, resulting in a large-scale accident association network with 1368 nodes and 228,561 edges.

To further characterize the network structure, the key statistical metrics of the initial network are summarized in Table 3.

Table 3. Statistical characteristics of the initial accident network.

Metric	Value
Nodes	1368
Edges	228,561
Edge Density	0.244
Average Degree	334
Maximum Degree	678
Modularity (Q)	0.173

To further examine the structural properties of the network, visualization and community detection are performed using Gephi (version 0.10.1). The Fruchterman–Reingold force-directed layout is applied to enhance the visualization of node distribution and connectivity patterns. Community detection is conducted using the Fast Unfolding (Louvain) algorithm, with modularity (Q) used as the primary metric to evaluate the quality of community partitioning. A total of four communities are identified in the initial network, and the overall topology and community structure are illustrated in Figure 2.

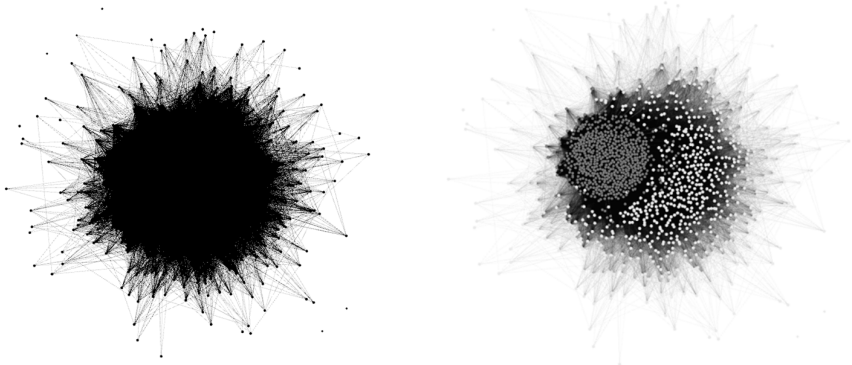
Despite the relatively high connectivity of the network, the community structure appears visually indistinct, which is characteristic of a “hairball-like” network. This phenomenon suggests that the initial feature selection and preprocessing may not sufficiently distinguish structural patterns. The underlying reasons can be summarized as follows:

(1) High-frequency general terms introduce redundant edges: For example, terms such as “accident” and “investigation” appear frequently across the corpus, leading to

a large number of co-occurrence links. While these connections increase network density, they contribute limited discriminative information and may reduce the separability of communities.

(2) Skewed node degree distribution: Some nodes become highly connected due to the influence of high-frequency terms, forming locally dense substructures that may obscure other community patterns.

(3) Limited discriminative capability of initial keyword selection: The initial network is constructed based on TextRank and frequency filtering, without further optimization of feature importance, which may limit the clarity of community partitioning.



(**a**) Network Structure of Construction Safety Accidents

(**b**) Community Structure of the Construction Safety Accident Network

Figure 2. Construction safety accident network.

These observations indicate that the structural clustering patterns in the initial network mainly reflect statistical co-occurrence relationships rather than well-defined accident categories or risk mechanism groupings. This is supported by both the visualization in Figure 3 and the quantitative metrics in Table 2.

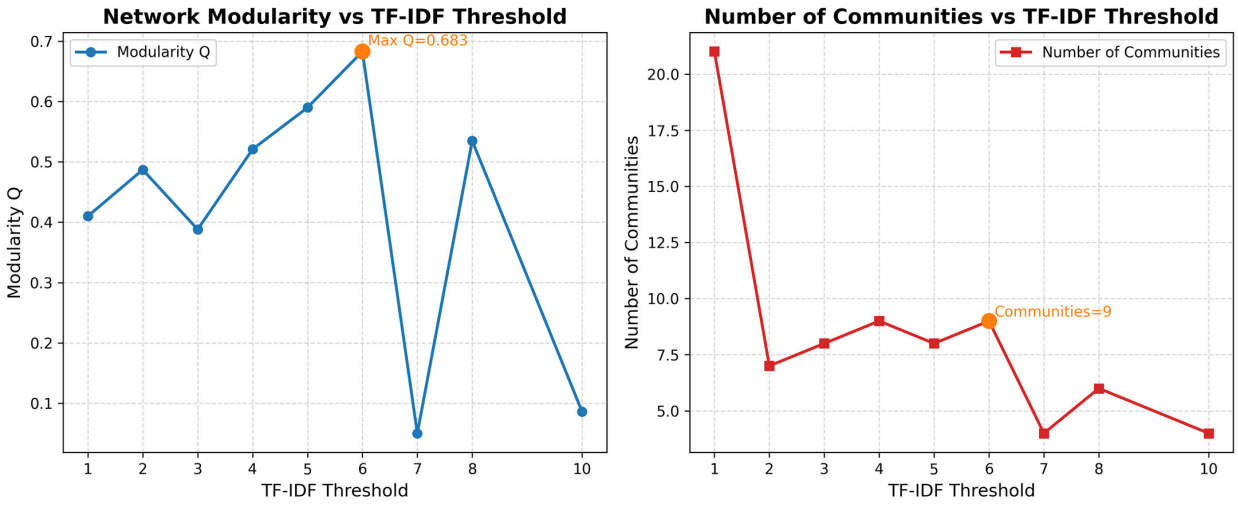


Figure 3. Sensitivity analysis of network structure with respect to TF-IDF thresholds.

To address these issues, a TF-IDF-based keyword weighting and filtering strategy is introduced in the subsequent analysis. By constructing networks under different candidate thresholds through systematic parameter search, the variation in modularity is examined. The results show that modularity exhibits a trend of increasing initially and then decreasing as the TF-IDF threshold increases. When the threshold is approximately 6, the network demonstrates a relatively clearer community structure, suggesting a balance between noise

reduction and information retention. This provides a basis for further network optimization and analysis.

3.3. TF-IDF Optimization

To reduce the influence of high-frequency general terms on the network structure observed in the initial accident network, further refinement of the keyword set is conducted to improve structural distinguishability and community partitioning. In this study, the TF-IDF (term frequency–inverse document frequency) method is applied to evaluate keyword importance, and different threshold values are used to filter keywords, thereby enabling structural optimization of the accident network.

Based on the keyword set obtained after frequency-based filtering, TF-IDF weights are calculated for all keywords. A series of threshold values (1, 2, 3, 4, 5, 6, 7, 8, and $\geq$10) is then applied to select keywords under different conditions. For each threshold, a corresponding accident co-occurrence network is reconstructed. Structural metrics, including the number of nodes, number of edges, modularity, and number of communities, are calculated for each network to examine how the network properties vary with different TF-IDF thresholds. The structural characteristics of the accident networks under different threshold settings are summarized in Table 4.

Table 4. Network structure and community partition under different TF-IDF thresholds.

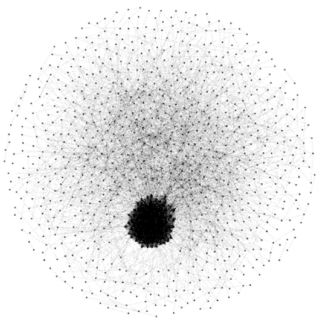
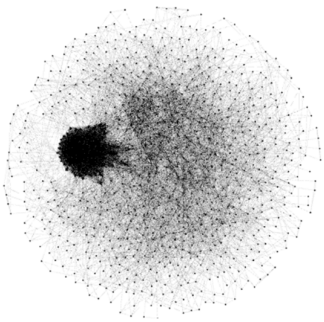
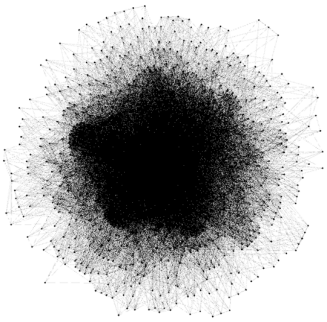
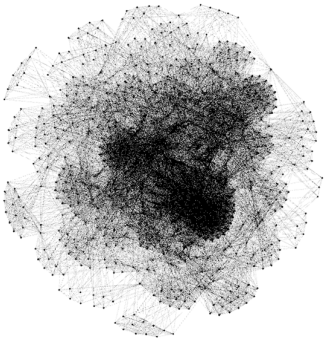
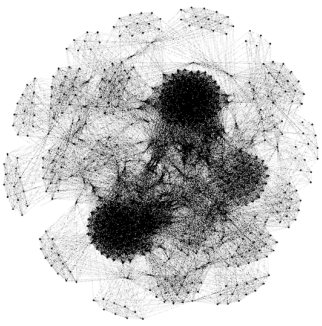
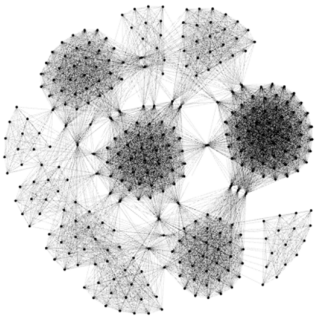
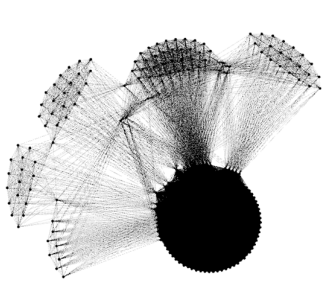
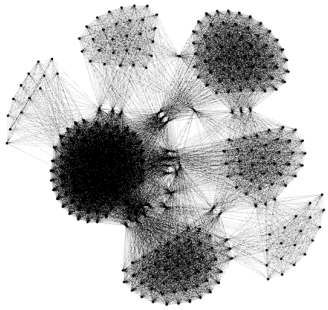
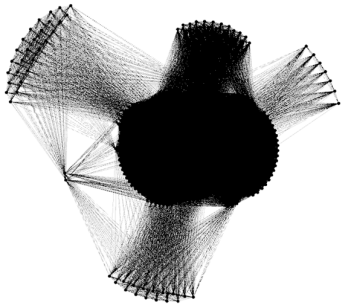
		
TF-IDF = 1	TF-IDF = 2	TF-IDF = 3
Number of Keywords: 986	Number of Keywords: 262	Number of Keywords: 126
Node = 921, Edge = 6878	Node = 850, Edge = 10,117	Node = 1005, Edge = 28,870
Modularity (Q) = 0.410	Modularity (Q) = 0.487	Modularity (Q) = 0.388
Number of Communities: 21	Number of Communities: 7	Number of Communities: 8
		
TF-IDF = 4	TF-IDF = 5	TF-IDF = 6
Number of Keywords: 44	Number of Keywords: 24	Number of Keywords: 11
Node = 633, Edge = 14,085	Node = 531, Edge = 13,825	Node = 308, Edge = 6502
Modularity (Q) = 0.521	Modularity (Q) = 0.590	Modularity (Q) = 0.683
Number of Communities: 9	Number of Communities: 8	Number of Communities: 9

Table 4. *Cont.*

		
TF-IDF = 7	TF-IDF = 8	TF-IDF $\geq$ 10
Number of Keywords: 6	Number of Keywords: 7	Number of Keywords: 6
Node = 513, Edge = 83,992	Node = 319, Edge = 11,791	Node = 581, Edge = 116,973
Modularity (Q) = 0.050	Modularity (Q) = 0.535	Modularity (Q) = 0.086
Number of Communities: 4	Number of Communities: 6	Number of Communities: 4

As shown in Table 4, the number of retained keywords decreases progressively as the TF-IDF threshold increases. This indicates that terms with lower discriminative capability are gradually removed during the filtering process, which reduces redundant co-occurrence relationships in the network. At lower threshold levels, the network still contains a relatively large number of less informative keywords, leading to dense connections and less distinguishable community structures. As the threshold increases, the network structure becomes more clearly organized, and the modularity value shows an increasing trend.

When the TF-IDF threshold is set to 6, the modularity reaches 0.683, representing the most structurally stable configuration within the tested range. This threshold yields consistent community partitions without implying semantic correctnessity. By comparing the structural metrics across different threshold settings, the threshold value of 6 is selected for subsequent analysis. Under this condition, a set of 11 core keywords is obtained, which provides a statistical representation of frequently co-occurring risk-related features in the accident texts and supports the formation of the network community structure. The corresponding keyword set is listed in Table 5.

Table 5. Keyword set corresponding to an optimized network structure.

TF-IDF = 4					
observed	villagers	at that time	deceased	family members	project
construction site	this person	incident	comprehensive	operation/work activities	loud noise
rescued	rescue unsuccessful	construction work	arrived promptly	nearby	accident site
affected	worker	staff	sudden	actively	clearing
one person	doctor	entrance	continue	ensure	
safety supervision	colleagues	public security	various	planning	
injured	people	head	fracture	building	
injuries	multiple locations	severe	unit	responsibility	
TF-IDF = 5			**TF-IDF = 6**		**TF-IDF = 8**
this incident	sent to	death	life-threatening	construction activity	intervention
hospital	search and rescue	trapped	already	follow-up matters	handling
collapse	all	compensation	medical treatment	inspection	related
requirement	expected	one case	collapse		orderly
handling	work stoppage	local	safety hazard		work
control	situation	properly manage	post-incident handling		established
two (persons)	two people	structural collapse	rescue		safety
stable	this year	constructed	surrounding area		

Overall, the TF-IDF-based filtering process reduces the influence of high-frequency general terms and improves the clarity of the network structure. This provides a more stable basis for subsequent accident risk association analysis.

3.4. Sensitivity Analysis

To further examine the rationality of the TF-IDF threshold selection and the stability of the proposed method, a sensitivity analysis is conducted on network structural metrics under different threshold settings. The analysis focuses on the variation in modularity (Q) and the number of detected communities with respect to the TF-IDF threshold, in order to evaluate the robustness of the network structure.

The results of the sensitivity analysis are presented in Figure 3. The left panel illustrates the variation in modularity with increasing TF-IDF thresholds, while the right panel shows the corresponding changes in the number of communities. The maximum values of both metrics occur at a TF-IDF threshold of 6 and are highlighted in the figure.

The results indicate that when the TF-IDF threshold is within the range of 5–6, the modularity remains at a relatively high level (0.590 and 0.683, respectively), and the number of communities shows only minor variation. This suggests that the network structure maintains a certain degree of stability and consistency within this interval. These observations imply that network performance is not dependent on a single parameter value but rather remains stable within a threshold range.

In contrast, when TF-IDF ≤ 3, a large number of low-discriminative keywords are retained, resulting in dense but less informative connections, lower modularity values, and less distinguishable community structures. When TF-IDF ≥ 7, the number of retained keywords decreases substantially, leading to a sparser network, reduced modularity, and less stable community partitioning.

Therefore, the threshold value of 6 is chosen as a representative value within the structurally stable interval (5–6), not as a globally stable or semantically superior parameter. This result suggests that the proposed optimization strategy reduces reliance on subjective parameter selection and supports the robustness and reproducibility of the method.

It should be noted that the proposed approach is based on static textual data, and the resulting network reflects data-driven statistical associations. Future work may consider incorporating temporal analysis or dynamic network modeling to better capture the evolution of accident risks over time.

3.5. Optimized Network Structure

Under the TF-IDF threshold of 6, a total of 11 core risk-related keywords are retained, and the construction safety accident association network is reconstructed accordingly. Compared with the initial network, the optimized network exhibits a clearer and more stable community structure, with improved structural distinguishability. The overall network structure is shown in Figure 4.

As illustrated in Figure 4, the optimized network is partitioned into nine distinct communities. Nodes within each community are relatively densely connected, while the boundaries between different communities are more distinguishable. This observation is consistent with the modularity value (Q = 0.683), indicating that the network exhibits a certain level of structural separation and clustering characteristics. In terms of community size, Community 0 and Community 1 contain the largest number of nodes, which may correspond to relatively common combinations of accident patterns in construction scenarios, whereas smaller communities reflect more localized risk characteristics under specific conditions.

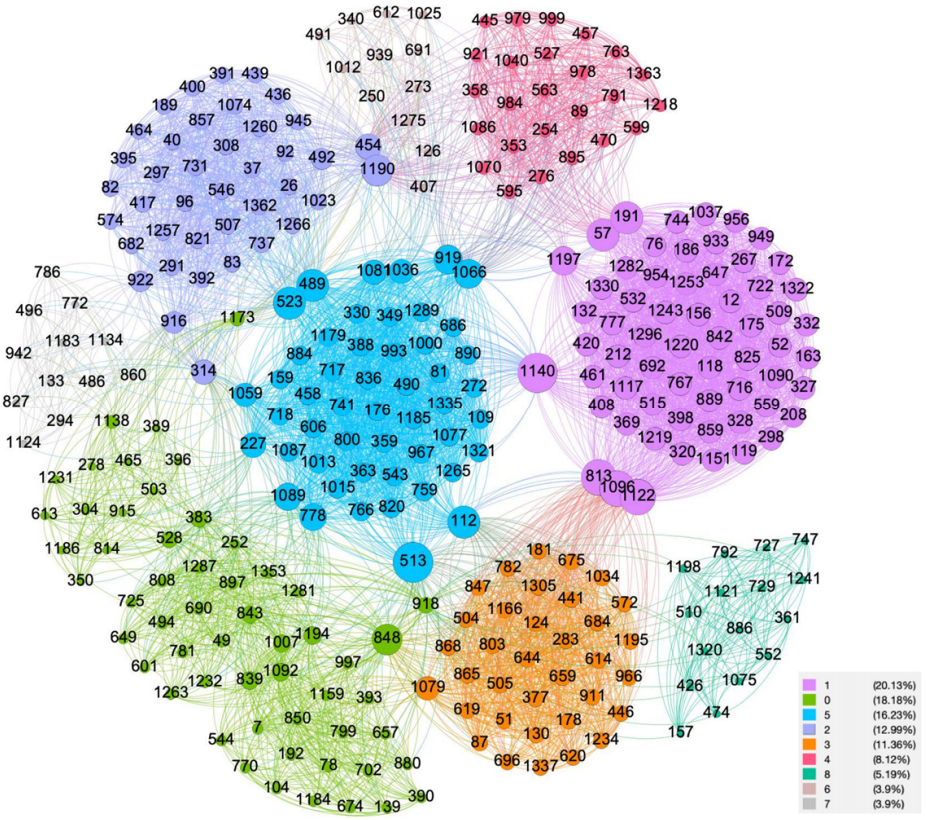


Figure 4. Community structure of the accident network (TF-IDF = 6).

Further examination of the network topology shows that some nodes have relatively high connectivity across multiple communities, which can be regarded as bridging nodes. These nodes connect different communities and play a role in linking otherwise separated substructures. It should be emphasized that such bridging relationships represent statistical associations at the keyword level, rather than direct causal relationships between accidents.

Through these bridging nodes, it can be observed that different types of accidents (e.g., falls from height and structural collapse) may share certain keyword-level similarities. This may be related to common working environments, management factors, or modes of risk representation. However, such associations only indicate co-occurrence patterns in feature representation and should not be interpreted as direct pathways of data-driven statistical associations across different accident types.

In addition, some communities are connected through a small number of key nodes, forming relatively close inter-community structures (e.g., the connection between Community 1 and Community 5, as shown in Figure 5). This pattern reflects the presence of shared keyword features between different accident subgroups and provides statistical evidence for identifying commonly co-occurring features.

It should be noted that the constructed accident association network is essentially a statistical network based on textual co-occurrence. Therefore, the analysis focuses on identifying association patterns rather than causal mechanisms. The identification of causal relationships in accident data-driven latent risk associations would require additional evidence, such as temporal data, causal inference methods, or detailed accident investigation frameworks. Accordingly, the analysis of the optimized network is intended to reveal clustering characteristics and association patterns among accidents, providing a data-driven reference for safety management, rather than a strict model of data-driven latent risk associations.

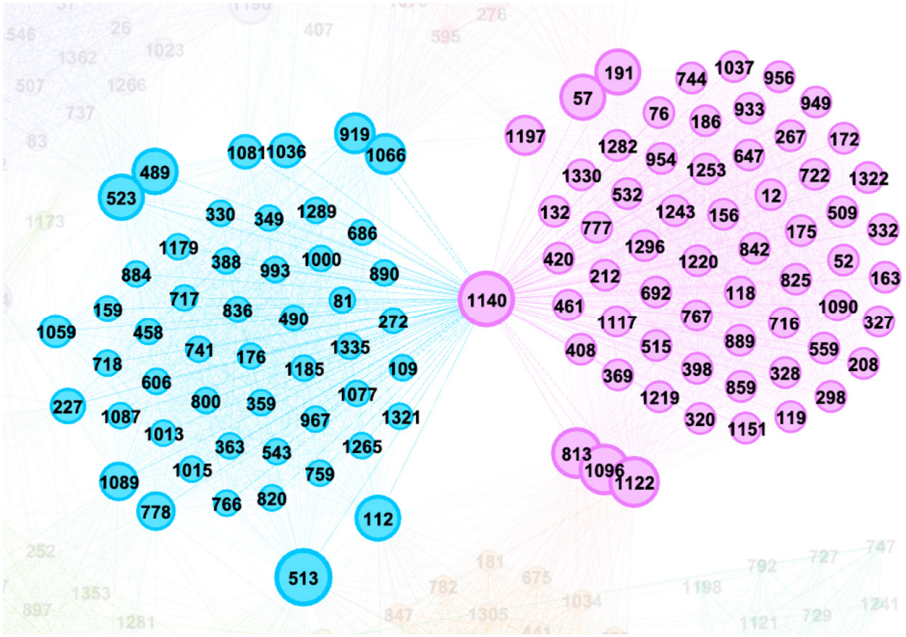


Figure 5. Relationship between Community 1 and Community 5.

To reduce reliance on manual visual interpretation of community structures and bridging nodes, an automated identification strategy based on network topological metrics is further introduced. This strategy enables the generation of structured outputs without manual inspection:

(1) Automatic community detection: Community labels are directly obtained using the Louvain algorithm, enabling automated identification of accident clusters.

(2) Automatic identification of bridging nodes: Nodes that connect different communities are identified based on quantitative metrics, such as betweenness centrality and cross-community connectivity, without relying on visual inspection.

(3) Standardized output generation: Structured results are automatically generated, including accident ID, community assignment, centrality scores, and bridging node indicators, which can be directly utilized by safety management practitioners.

These procedures rely on network topological indicators to enable an automated analysis workflow, reducing dependence on manual visualization and improving the practical applicability and operational feasibility of the proposed framework.

3.6. Comparative Evaluation with Text Clustering Methods

3.6.1. Experimental Procedure

To further evaluate the effectiveness of the proposed keyword co-occurrence network approach in capturing latent structural patterns of construction safety accidents, several conventional text clustering methods are selected for comparison. These include K-means clustering based on TF-IDF representations, hierarchical clustering based on TF-IDF representations, and K-means clustering based on Sentence-BERT (SBERT) embeddings.

The experiments are conducted on the same dataset described in previous sections, consisting of 1368 construction safety accident reports. Each record includes an accident title and a description, which are concatenated into a unified text for analysis. During preprocessing, the Jieba tool is used for Chinese word segmentation, converting raw text into token sequences.

For feature representation, both TF-IDF vectorization and SBERT-based semantic embeddings are employed. Clustering is then performed using the K-means algorithm based on these representations. As shown in Figure 6, to ensure comparability with the results of

the keyword co-occurrence network, the number of clusters for all clustering methods is set to nine, consistent with the number of communities identified in the network analysis.

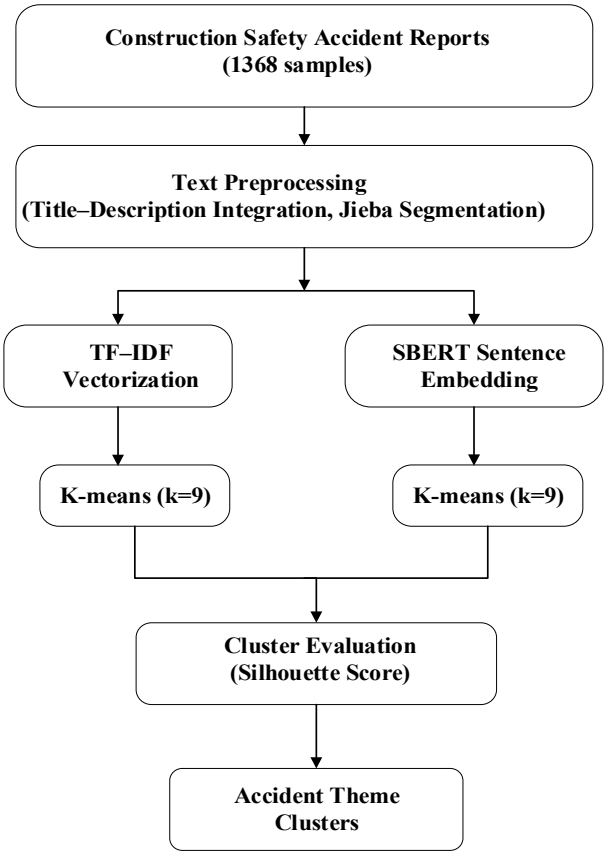


Figure 6. Experimental procedure of TF-IDF + K-means clustering.

In contrast, the keyword co-occurrence network approach constructs a weighted graph based on keyword co-occurrence relationships and identifies modular structures using community detection algorithms. To quantify the quality of network partitioning, modularity (Q) is adopted as the evaluation metric. For vector-based clustering methods, the Silhouette Score is used to measure intra-cluster compactness and inter-cluster separation.

It should be noted that different methods are based on distinct data representations. Vector space methods focus on semantic similarity between texts, whereas the network-based approach emphasizes structural relationships among keywords. Therefore, the evaluation metrics reflect different aspects of structural characteristics and are not intended for direct numerical comparison.

To provide a more unified perspective for comparison, Normalized Mutual Information (NMI) is further introduced to measure the consistency between clustering results obtained from different methods.

It should be emphasized that clustering metrics (Silhouette Score) and network modularity are methodologically inconsistent and not directly comparable. The Silhouette Score measures vector-space semantic similarity, while modularity evaluates topological community separation in the co-occurrence network. This comparison aims to show that the proposed network method provides a complementary structure-aware perspective, rather than claiming numerical superiority over clustering approaches. Normalized Mutual Information (NMI) is used to measure result consistency across different methods.

3.6.2. Clustering Results

The experimental results of the different methods are presented in Table 6.

Table 6. Comparison of clustering methods and the proposed keyword co-occurrence network approach.

Method	Feature Representation	Number of Classes	Metric	Result
TF-IDF + K-means	TF-IDF	9	Silhouette	0.0557
TF-IDF + Hierarchical	TF-IDF	9	Silhouette	0.0449
Sentence-BERT + K-means	SBERT	9	Silhouette	0.1053
Keyword Co-occurrence Network	Network	9	Modularity (Q)	0.683
NMI (Network vs. SBERT)	—	9	NMI	0.0616
NMI (Network vs. TF-IDF)	—	9	NMI	0.0689
NMI (TF-IDF vs. SBERT)	—	9	NMI	0.152

As shown in Table 4, clustering methods based on TF-IDF representations exhibit relatively weak overall performance. Specifically, TF-IDF + K-means achieved a Silhouette Score of 0.0557, while TF-IDF + hierarchical clustering yielded a score of 0.0449. The low Silhouette Scores indicate that, in the vector space constructed from term frequency features, different accident texts exhibit considerable overlap, resulting in poor separation between clusters. When semantic vectors generated by Sentence-BERT were used as text representations, clustering performance improved moderately, with the Silhouette Score increasing to 0.1053. This suggests that, compared to traditional bag-of-words representations, deep semantic embeddings can more effectively capture co-occurrence-based structural patterns between accident texts, thereby enhancing cluster compactness and separability to some extent. However, the score remains relatively low, indicating that semantic overlap still exists among accident texts in the current dataset.

In contrast to vector-space clustering methods, the proposed keyword co-occurrence network constructs connections based on the co-occurrence of key terms extracted from accident texts, forming an association network among accident concepts. In the resulting network, community detection identifies modules with a modularity value of $Q = 0.683$. The high modularity indicates that nodes within the same community are densely connected, whereas connections between different communities are relatively sparse, demonstrating a well-defined structural partitioning of the network. To provide a more intuitive comparison of how different methods capture the structural characteristics of accident texts, the visualization results for each approach are presented in Figure 7.

Figure 7 presents two-dimensional visualizations of the clustering results along with their corresponding quantitative metrics. As observed in Figure 7a,b, the clustering results based on TF-IDF representations exhibit noticeable overlap among samples, which is consistent with their low Silhouette Scores. In Figure 7c, the SBERT-based clustering shows a tendency toward local grouping, corresponding to its relatively higher Silhouette Score, although cluster boundaries remain less distinct. In Figure 7d, the keyword co-occurrence network reveals several relatively well-defined community structures, consistent with the higher modularity value ($Q = 0.683$), indicating clearer structural partitioning at the network level.

By jointly examining visualization results and quantitative metrics (Silhouette, modularity, and NMI), a consistent pattern can be observed across different methods: vector-based approaches provide limited separation in terms of semantic similarity, while the keyword co-occurrence network offers a more distinguishable structure from the perspective of relational organization among keywords.

It should be emphasized that different methods characterize different structural aspects of the data. Vector-based clustering primarily reflects semantic similarity between texts, whereas the keyword co-occurrence network focuses on structural relationships among keywords. Therefore, the proposed approach is not intended to replace conventional clustering methods, but rather to provide a complementary analytical perspective from the

viewpoint of structural associations. The combined use of visualization and quantitative metrics helps reduce reliance on subjective interpretation and provides a more objective basis for comparison.

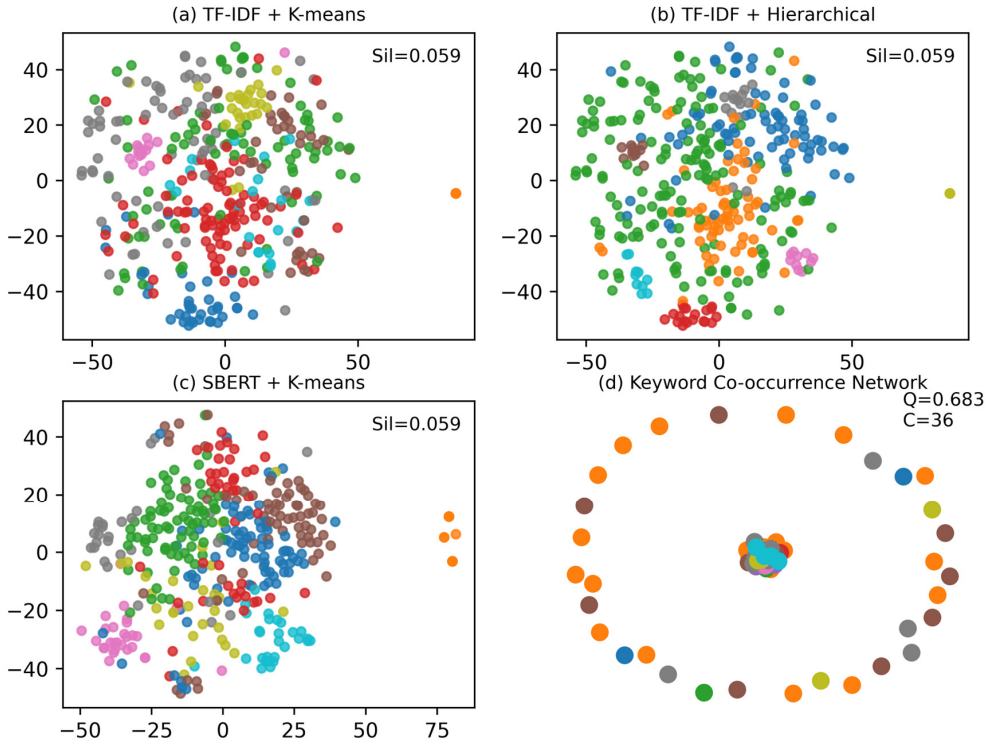


Figure 7. Distribution of accident counts across clusters in the K-means clustering results.

3.7. Case Study: Regional Accident Network Validation

To examine the applicability and transferability of the proposed keyword co-occurrence network approach under different data conditions, a regional case study is conducted using accident data from Southeast China. The study area includes Shanghai, Zhejiang, Jiangsu, Anhui, Jiangxi, Fujian, Guangdong, Guangxi, Hainan, and Taiwan, as well as the Hong Kong and Macao Special Administrative Regions. Based on official accident reporting channels, a total of 82 construction safety accident records from the past ten years (January 2015 to June 2024) are collected to construct a regional dataset.

It should be noted that the size of this regional dataset is relatively limited and is used primarily for external validation rather than model training or parameter optimization. The purpose is to examine whether the proposed method can still identify stable structural patterns under different data distributions.

To ensure comparability, the same data processing procedure as in the main experiment is applied. Accident texts are first standardized and preprocessed, followed by Chinese word segmentation using the Jieba tool. Keywords are then extracted using the TextRank algorithm, resulting in 328 initial candidate keywords. After applying a frequency filtering threshold (occurrence ≥ 2), 66 keywords are retained to construct the regional keyword co-occurrence matrix and the corresponding accident association network, as shown in Figure 8.

Considering potential differences in data distribution between the regional dataset and the full dataset, TF-IDF-based filtering is further applied to refine the keyword set. Multiple network structures are constructed under different threshold settings, and their modularity (Q) and community characteristics are compared. Representative results are summarized in Table 7.

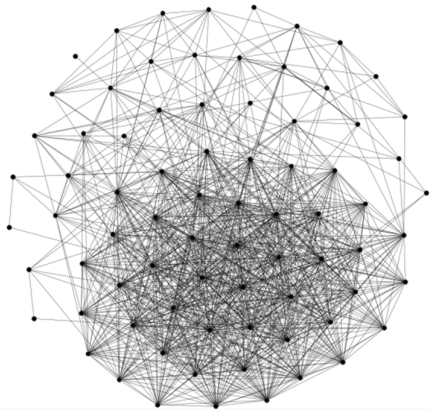
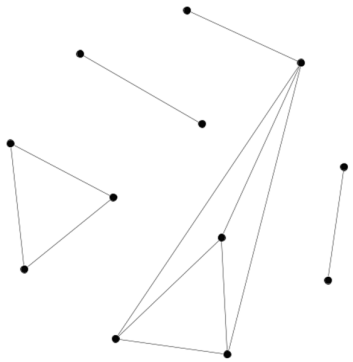
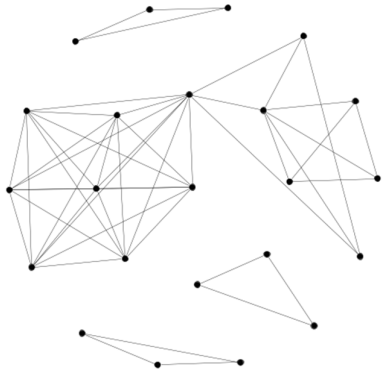
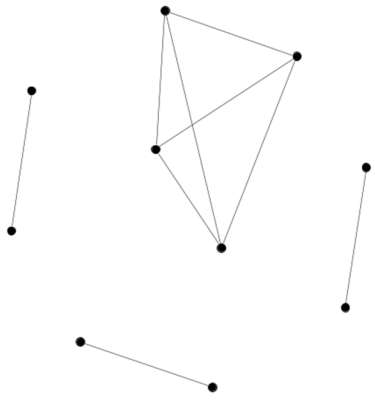
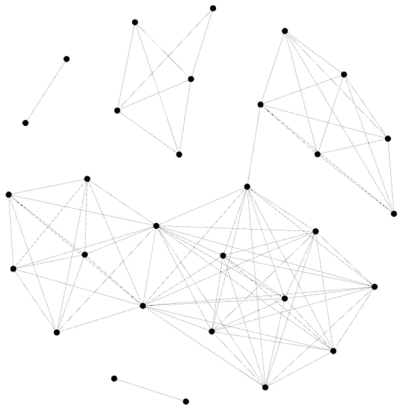


Figure 8. Accident network of construction safety accidents in Southeast China over the past ten years.

Table 7. Network structures under different TF-IDF thresholds with relatively clear community patterns.

	
TF-IDF = 8	TF-IDF = 10
Number of Keywords: 5	Number of Keywords: 6
Node = 12, Edge = 12	Node = 23, Edge = 49
Modularity (Q) = 0.583	Modularity (Q) = 0.519
Number of Communities: 4	Number of Communities: 5
	
TF-IDF = 13	TF-IDF = 18
Number of Keywords: 4	Number of Keywords: 7
Node = 10, Edge = 9	Node = 30, Edge = 91
Modularity (Q) = 0.519	Modularity (Q) = 0.514
Number of Communities: 4	Number of Communities: 6

As shown in Table 7, the modularity values remain around 0.5 under different threshold settings, indicating that the network exhibits a certain level of community structure even under small-sample conditions. When TF-IDF = 8, the modularity reaches the highest value ($Q = 0.583$), although the network size is relatively small. In contrast, when TF-IDF = 18, the network maintains a comparable modularity ($Q = 0.514$) while including a larger number of nodes and edges, thus preserving more structural information. Considering both structural stability and information completeness, TF-IDF = 18 is selected as a representative threshold for regional network analysis, and the corresponding network structure is shown in Figure 9.

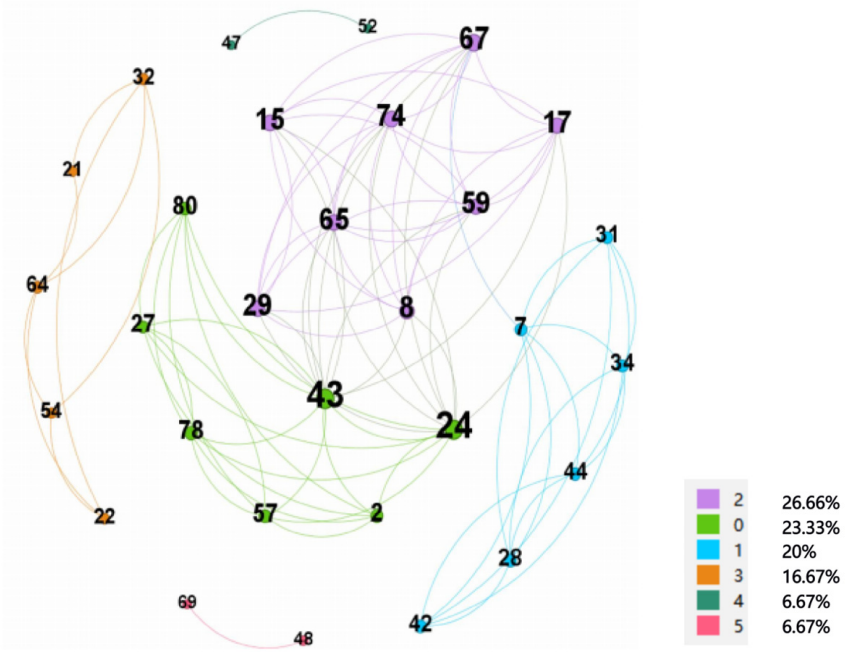


Figure 9. Optimized accident network structure (TF-IDF = 18).

To further assess the statistical stability of the regional network, a bootstrap-based resampling analysis is conducted. Specifically, multiple random samples (80% sampling ratio) are drawn from the original dataset, and the accident network is reconstructed for each sample, with modularity values calculated accordingly. The results show that the variation in modularity remains within a relatively narrow range (approximately ± 0.03), suggesting a certain level of structural stability despite the limited sample size.

From a structural perspective, the regional accident network also exhibits multiple communities, each corresponding to different combinations of accident characteristics. In some communities, different accident types (e.g., structural collapse and falls from height) show overlap at the keyword level, forming connections within the network. This suggests that different accident types may share certain risk-related contexts or management-related characteristics.

It should be emphasized that these associations represent statistical relationships derived from textual data, rather than causal relationships or identifying statistical co-occurrence patterns of risk features. Accordingly, the connections in the network are interpreted as co-occurrence structures of risk-related features, rather than causal linkages or interactive mechanisms. In addition, the current model is constructed based on static accident report data and does not explicitly incorporate temporal or spatial variables (e.g., time dynamics or regional heterogeneity). Therefore, the analysis focuses on structural association patterns rather than dynamic evolution processes. Future work may integrate

temporal analysis and spatial modeling approaches (e.g., temporal or spatial networks) to further enhance the representation of dynamic and regional risk characteristics.

Overall, the regional case study indicates that the proposed keyword co-occurrence network approach can identify relatively stable structural patterns across datasets with different scales and distributions, suggesting its applicability under varying data conditions. The inclusion of robustness analysis and methodological boundary clarification further supports the reliability and interpretability of the results. This provides empirical evidence that the proposed framework maintains structural consistency across datasets with heterogeneous distributions. These results further demonstrate the robustness of the proposed framework across different data scales and distributions.

4. Discussion

From a methodological perspective, the proposed accident–keyword co-occurrence network framework extends conventional paradigms of accident analysis. Unlike approaches that rely on predefined causal relationships, this study adopts a data-driven strategy by constructing an accident association network based on keyword co-occurrence patterns extracted from textual reports, thereby characterizing structural association patterns among accidents.

Experimental results based on 1368 official accident reports show that risk-related keywords form a relatively clear community structure within the network (Modularity $Q = 0.683$), indicating that the method is capable of capturing observable structural patterns among accidents. It should be emphasized that the proposed framework identifies statistical co-occurrence patterns and structural associations derived from textual data, rather than inferring causal relationships, risk propagation processes, or underlying mechanisms. Accordingly, the findings should be interpreted as a descriptive characterization of structural patterns, rather than causal inference.

4.1. Comparison with Traditional Accident Analysis Frameworks

Traditional construction accident analysis approaches—such as the Swiss Cheese Model, accident causation chain theories, and the 2–4 model—typically rely on expert knowledge to infer causal relationships. While these methods provide valuable theoretical insights, they may be influenced by subjective judgment and face limitations in scalability when applied to large-scale accident datasets.

In contrast, the proposed method follows a data-driven paradigm that does not require predefined accident categories or causal pathways. Instead, it extracts keywords from accident texts and constructs a co-occurrence network to identify structural association patterns directly from the data. Compared with traditional causality-oriented frameworks, the present approach focuses on identifying similarities and relational patterns among accidents from a structural perspective. Through community detection, it reveals clustering patterns of accidents with similar risk characteristics, offering a complementary perspective based on network structure.

To enhance the interpretability of the identified network relationships, this study further aligns the constructed accident network with classical accident causation theories (e.g., the Swiss Cheese Model and the 2–4 model), establishing a conceptual mapping between network structures and underlying causation mechanisms. This integration provides a bridge between data-driven findings and domain knowledge:

(1) Network communities correspond to clusters of similar defense layer failures in the Swiss Cheese Model, such as on-site operational deficiencies, inadequate supervision, environmental abnormalities, or organizational management issues;

(2) Bridging nodes represent points where multiple layers of defense may simultaneously exhibit weaknesses, reflecting critical positions of multi-factor coupling across human, equipment, environment, and management dimensions;
(3) Keyword co-occurrence relationships correspond to multi-factor interaction mechanisms described in accident causation theories, providing statistical evidence that is consistent with established theoretical perspectives.

Through this mapping, the structural associations observed in the network can be interpreted from the perspective of traditional accident causation mechanisms, addressing the limitations of purely statistical interpretations and enhancing the domain relevance of the results.

As shown in Table 8, the corresponding mapping relationships are summarized as follows:

Table 8. Mapping between network structures and the Swiss Cheese Model.

Network Structural Element	Corresponding Mechanism in Swiss Cheese Model
Community	Clusters of similar defense layer failures
Bridging node	Intersection of multiple failures/multi-factor coupling points
Keyword co-occurrence	Coupling of human–equipment–environment–management factors

The experimental results indicate that the identified keywords cover multiple dimensions of risk factors, including human, equipment, material, method, and environment. This observation is consistent with the multi-factor interaction perspective emphasized in classical accident theories, which indirectly supports the interpretability of the proposed method.

It should be noted that the grouping of different accident types within the same community reflects similarity in risk characteristics or semantic representation. Such structural associations are statistical in nature and should not be interpreted as evidence of direct causal relationships between accidents.

4.2. Effectiveness of Network Construction and Optimization Methods

The experimental results demonstrate that keyword selection strategies significantly influence the structure of the accident network. In the initial network without filtering, the network constructed using TextRank-extracted keywords exhibits low modularity ($Q = 0.173$), and the community structure is relatively indistinct. This is mainly because high-frequency generic terms introduce a large number of redundant co-occurrence relationships, resulting in overly dense network connections and weakened structural discrimination.

After introducing the TF-IDF-based filtering mechanism, the network became topologically more stable and consistent. Combined with the sensitivity analysis in Section 3.5, it can be observed that when the TF-IDF threshold ranges between 5 and 6, the network modularity remains at a relatively high level (0.590–0.683), and the community structure is stable. When the threshold is low, low-discriminative keywords are retained, leading to redundancy in the network structure; when the threshold is too high, the number of keywords decreases significantly, resulting in a sparse network and a decline in modularity. This phenomenon indicates that an appropriate keyword filtering strategy can preserve essential semantic information while effectively reducing noise interference, thereby improving the clarity and interpretability of the network structure.

In addition, some nodes in the network act as connectors across multiple communities, exhibiting high centrality. These nodes reflect potential associations between different risk features. However, their structural role should be interpreted as "association bridging"

rather than "identifying statistical co-occurrence patterns of risk features," and thus they should not be directly considered as causal intermediaries or transmission mechanisms.

4.3. Practical Implications for Construction Safety Management

From an engineering application perspective, the keyword co-occurrence network provides a structured analytical tool for construction safety analysis. Through community detection, different accidents can be organized into groups with similar risk characteristics based on shared features, offering safety managers a systematic basis for risk classification. Compared with traditional vector-space-based clustering methods, the proposed approach does not rely directly on overall textual semantic similarity. Instead, it characterizes structural relationships among accidents through keyword co-occurrence, which enables the identification of organizational patterns among risk factors. Based on this, safety managers can understand potential accident scenarios from the perspective of combinations of risk factors, thereby supporting the development of targeted safety management strategies. Furthermore, regional case analysis shows that even with relatively small datasets, the optimized network after keyword filtering can still form meaningful community structures. This suggests that the proposed method has a certain level of applicability and transferability across different data conditions.

In practical applications, this approach can be integrated into construction safety management information systems to enable automated processing and analysis of accident texts. For example, automatic keyword extraction and network construction can be used to transform unstructured accident data into structured representations, which can then be combined with visualization techniques to assist risk identification and decision-making.

4.4. Research Limitations and Future Directions

Although the proposed method demonstrates effectiveness in accident structure analysis, several limitations remain. First, keyword co-occurrence relationships reflect statistical associations rather than strict causal relationships; therefore, the interpretation of network results should avoid confusing correlation with causation. Second, under certain parameter settings (e.g., high TF-IDF thresholds), the network structure may deteriorate (e.g., reduced modularity or excessive sparsity), indicating that the model is still sensitive to keyword filtering strategies.

In addition, the constructed network is based on static textual data and does not consider the temporal evolution of accident risks. Spatial attributes such as geographic location and project type are also not incorporated, which limits the model's ability to capture regional heterogeneity and dynamic characteristics in complex construction environments.

Future research can be extended in the following directions: (1) incorporating temporal data to construct dynamic accident networks to capture risk evolution; (2) integrating spatial information to build spatial–semantic coupled networks for improved regional analysis; (3) combining causal inference methods to enhance causal interpretability; and (4) leveraging deep semantic models (e.g., BERT or RoBERTa) to improve keyword extraction and semantic representation, thereby further enhancing model performance and analytical accuracy.

5. Conclusions

This study proposes the CESA-Miner framework, which integrates a two-stage adaptive keyword extraction strategy, TF-IDF filtering, and keyword co-occurrence network modeling to achieve structured analysis of construction accident texts. The optimized network yields a more structurally stable community configuration with modularity increasing from 0.173 to 0.683, identifying groups of accidents with shared statistical features.

Compared with traditional document-similarity-based clustering methods, the proposed approach characterizes structural relationships among accidents from the perspective of keyword associations, providing a complementary analytical viewpoint. In addition, the automated identification rules and mapping mechanism to accident causation theory reduce the operational burden on safety managers while enhancing the interpretability of risk associations, thereby integrating data-driven methods with traditional safety theories. It should be emphasized that the network is constructed based on statistical co-occurrence, and the identified associations reflect structural correlations rather than causal relationships. This study represents exploratory data analysis, and the results are primarily intended to support understanding of the structural features of accident data. Overall, the keyword co-occurrence network offers a feasible approach for analyzing large-scale accident texts. Future work may incorporate dynamic network modeling and deep semantic representations to further enhance analytical capability.

Author Contributions: Conceptualization, W.Y. and G.L.; Methodology, S.L., J.M. and G.L.; Software, S.L. and G.L.; Formal analysis, S.L. and R.Z.; Investigation, R.Z.; Data curation, J.M.; Writing—draft, S.L.; Writing—review & editing, W.Y. and G.L.; Visualization, J.M. and R.Z.; Supervision, W.Y. and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the 2025 University Basic Scientific Research Project of Liaoning Provincial Department Education, grant number LJ212510153048. The APC was funded by the corresponding author.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Demirkesen, S. Investigating linear models of accident causation: A review study in the construction safety context. *Sigma J. Eng. Nat. Sci.* **2021**, *38*, 1939–1949.
2. Surry, J. *Industrial Accident Research: A Human Engineering Appraisal*; Department of Industrial Engineering, University of Toronto: Toronto, ON, Canada, 1971.
3. Reason, J. The contribution of latent human failures to the breakdown of complex systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **1990**, *327*, 475–484. [CrossRef]
4. Fu, G.; Han, M.; Jia, L.; Wan, J.; Fu, H. Accident causation theory and its development: A review. *Safety* **2023**, *44*, 1–8+89. [CrossRef]
5. Carrillo-Castrillo, J.A.; Trillo-Cabello, A.F.; Rubio-Romero, J.C. Construction accidents: Identification of the main associations between causes, mechanisms and stages of the construction process. *Int. J. Occup. Saf. Ergon.* **2017**, *23*, 240–250. [CrossRef]
6. Zhou, Z.; Irizarry, J.; Guo, W. A network-based approach to modeling safety accidents and causations within the context of subway construction project management. *Saf. Sci.* **2021**, *139*, 105261. [CrossRef]
7. Li, Z.; Zhang, Y. High-Altitude Fall Accidents in Construction: A Text Mining Analysis of Causal Factors and COVID-19 Impact. *Modelling* **2025**, *6*, 124. [CrossRef]
8. Wang, Y.; Zou, P.X. Decoding Construction Accident Causality: A Decade of Textual Reports Analyzed. *Buildings* **2025**, *15*, 3859. [CrossRef]
9. Mistikoglu, G.; Gerek, I.H.; Erdis, E.; Usmen, P.M.; Cakan, H.; Kazan, E.E. Decision tree analysis of construction fall accidents involving roofers. *Expert Syst. Appl.* **2015**, *42*, 2256–2263. [CrossRef]
10. Tixier, A.J.P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Application of machine learning to construction injury prediction. *Autom. Constr.* **2016**, *69*, 102–114. [CrossRef]
11. Assaad, R.; El-adaway, I.H. Determining critical combinations of safety fatality causes using spectral clustering and computational data mining algorithms. *J. Constr. Eng. Manag.* **2021**, *147*, 04021035. [CrossRef]
12. Feng, X.; Lu, Y.; He, J.; Lu, B.; Wang, K. Bayesian-Network-Based Predictions of Water Inrush Incidents in Soft Rock Tunnels. *KSCE J. Civ. Eng.* **2024**, *28*, 5934–5945. [CrossRef]
13. Yoon, Y.G.; Ahn, C.R.; Yum, S.G.; Oh, T.K. Establishment of safety management measures for major construction workers through the association rule mining analysis of the data on construction accidents in Korea. *Buildings* **2024**, *14*, 998. [CrossRef]

14. Kumar, A.V.; Swapna, G.; Nikitha, G.; Sheshanthika, C.; Sreeja, G.; Akanshha, G. Construction Site Accident Analysis Using Textmining and Natural Language Processing Techniques. Available online: https://siiet.ac.in/wp-content/uploads/2023/11/23.CONSTRUCTION-SITE-ACCIDENT-ANALYSIS-USING-TEXTMINING-AND-NATURAL-LANGUAGE-PROCESSING-TECHNIQUES_compressed.pdf (accessed on 5 April 2026).
15. Ma, Z.; Chen, Z.S. Mining construction accident reports via unsupervised NLP and Accimap for systemic risk analysis. *Autom. Constr.* **2024**, *161*, 105343. [CrossRef]
16. Yoo, J.W.; Park, J.; Park, H. Enhancing safety of construction workers in Korea: An integrated text mining and machine learning framework for predicting accident types. *Int. J. Inj. Control. Saf. Promot.* **2024**, *31*, 203–215. [CrossRef]
17. Wang, J.; Wu, S.; Qu, Z.; Shen, J.; Memon, S.A.; Skitmore, M.; Ma, L. Classifying OSHA construction accident reports: Leveraging ensemble learning and lightweight large language models (l-LLMs). *Eng. Constr. Archit. Manag.* **2025**, *31*, 1–28. [CrossRef]
18. Yao, H.; She, J.; Zhou, Y. Risk assessment of construction safety accidents based on association rule mining and Bayesian network. *J. Intell. Constr.* **2024**, *2*, 1–16. [CrossRef]
19. Zang, H.; Li, M.; Jin, Z.; Huang, J. Unveiling construction accident causation: A scientometric analysis and qualitative review of research trends. *Front. Built Environ.* **2025**, *11*, 1602297. [CrossRef]
20. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
22. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
23. Poh, C.Q.; Ubeynarayana, C.U.; Goh, Y.M. Safety leading indicators for construction sites: A machine learning approach. *Autom. Constr.* **2018**, *93*, 375–386. [CrossRef]
24. Abbasianjahromi, H.; Mohammadi Golafshani, E.; Aghakarimi, M. A prediction model for safety performance of construction sites using a linear artificial bee colony programming approach. *Int. J. Occup. Saf. Ergon.* **2021**, *28*, 1265–1280. [CrossRef] [PubMed]
25. Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain, 25–26 July 2004; pp. 404–411. Available online: https://aclanthology.org/W04-3252.pdf (accessed on 5 April 2026).
26. Ramos, J. Using tf-idf to determine word relevance in document queries. *Proc. First Instr. Conf. Mach. Learn.* **2003**, *242*, 29–48.
27. Lowe, R.; Pow, N.; Serban, I.V.; Pineau, J. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czech Republic, 2–4 September 2015. [CrossRef]
28. Salton, G.; Wong, A.; Yang, C.S. A vector space model for automatic indexing. *Commun. ACM* **1975**, *18*, 613–620. [CrossRef]
29. Fruchterman, T.M.J.; Reingold, E.M. Graph drawing by force-directed placement. *Softw. Pract. Exp.* **1991**, *21*, 1129–1164. [CrossRef]
30. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
31. Newman, M.E.J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef] [PubMed]