buildings

MDPI

Article

Analyzing OSHA Construction Accident Reports Using BERTopic Topic Modeling for Thematic Insights

Yuntao Cao [1], Ziyi Qu [2], Shujie Wu [2], Yuting Chen [3], Martin Skitmore [4], Xingguan Ma [1] and Jun Wang [2,*]

[1] School of Municipal and Environmental Engineering, Shenyang Jianzhu University, Shenyang 110168, China
[2] School of Management Engineering, Qingdao University of Technology, Qingdao 266520, China
[3] Civil Engineering Technology and Construction Management, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[4] Faculty of Society and Design, Bond University, Robina, QLD 4226, Australia
* Correspondence: wangjun.gt@gmail.com

Abstract

Hazards at construction sites can lead to severe accidents, posing significant risks to worker safety, financial stability, and public confidence in industry safety standards. As a result, understanding and preventing these accidents has become increasingly critical. Although previous studies have examined historical accidents through detailed reports, few have systematically applied automated natural language processing (NLP) techniques to uncover hidden topics and patterns in large datasets without manual intervention. This study addresses this gap by applying topic modeling to 22,623 accident reports from the Occupational Safety and Health Administration (OSHA) spanning 2004 to 2023. The results demonstrate that BERTopic substantially outperforms the traditional LDA model across multiple accident datasets, achieving higher topic coherence and topic diversity. Leveraging contextual embeddings, BERTopic identifies nuanced risk scenarios, occupation–accident patterns, and temporal trends that earlier text-mining approaches often overlooked. The findings also generate actionable managerial insights, including peak accident periods, vulnerable worker groups, and scenario-specific risk factors. Overall, this study provides a clearer and more data-driven understanding of construction accident mechanisms through advanced topic modeling. Applying BERTopic for topic extraction and content analysis introduces a novel and effective approach to analyzing construction accident reports. The insights derived provide valuable guidance for decision-makers in risk mitigation and accident prevention, while helping to rebuild public confidence in safety standards. Moreover, the approach's reproducibility and potential for broader safety applications contribute to fostering a safer construction environment.

Keywords: topic modeling; BERTopic; LDA; construction industry; accident report analysis

Check for updates

Academic Editor: Ahmed Senouci

Received: 15 November 2025
Revised: 10 December 2025
Accepted: 18 December 2025
Published: 19 December 2025

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The construction industry is one of the largest global sectors, significantly contributing to GDP and representing one of the fastest-growing fields within the labor market [1,2]. However, the dynamic, complex, and often temporary nature of construction projects consistently ranks this industry as one of the most hazardous [3–7]. Approximately one in six fatal workplace accidents occurs on construction sites, with an estimated 60,000 fatal accidents occurring globally each year [8]. These accidents result in significant loss of life, environmental damage, financial losses, and other direct and indirect consequences.

Compared to other industries, the construction sector experiences a disproportionately high frequency of occupational accidents [9–12]. The most common accident types are falls, caught-in or -between accidents, struck-by events, and electrocution—collectively referred to as the "Construction Fatal Four" [13–15]. In many countries, the construction industry reports the highest rate of fatal accidents [16], and the U.S. construction industry had 1075 fatalities in 2023, according to the Bureau of Labor Statistics' Census of Fatal Occupational Injuries [17]. As a result, safety has become a critical concern in construction projects, highlighting the urgent need for improved health and safety practices within the industry [14,18].

Understanding the mechanisms underlying various types of construction accidents is important. Learning from past accidents and transforming these lessons into preventive strategies remains a key approach to maintaining workplace safety and enhancing the overall safety performance of the construction industry [19–21]. The Learning from Incidents (LFI) process has been widely recognized as an effective method for enhancing safety performance and preventing future accidents [14,20,22]. In this context, the term 'incident' includes both actual accidents and near-miss events [20,23]. To facilitate effective LFI, it is important to extract meaningful information from occupational incident reports [24,25].

Various theories and frameworks have been proposed to analyze the causes of construction accidents. These include heterogeneous correlation networks [26], which explore relationships between key causes and accident attributes; causation system models [27], which identify critical accident causes; fuzzy hybrid frameworks [28], which examine interrelationships among key causes; and analytic hierarchy process (AHP) models [29], which analyze and prioritize the factors contributing to accidents.

However, these approaches face several limitations. First, they struggle to efficiently process large volumes of construction accident data [30]. Second, they lack systematic automation [31], limiting their ability to perform large-scale, reproducible analyses. Third, they focus on surface-level patterns, overlooking deeper semantic insights within accident reports [32]. As a result, their capacity to provide comprehensive analyses and a deep understanding of accident reports is constrained.

Recent research has begun to explore the use of NLP and machine learning (ML) techniques to analyze large volumes of structured and unstructured text data. Through ML methods, NLP addresses the interface between human language and computers, enabling systems to understand and manipulate natural language to perform specific tasks [30,33]. Research applying these techniques in construction health and safety has primarily focused on developing models for classifying accident reports into different categories [34–36]. However, these models often lack deeper content analysis.

In contrast, topic modeling offers deeper insights and a more comprehensive understanding when analyzing construction accident reports [37]. Topic modeling can automatically extract hidden topics from large volumes of reports, revealing deeper semantic insights within the text [38]. Additionally, it can dynamically update as new data are incorporated, providing the real-time topic analysis needed for tracking emerging safety trends and issues [39].

This study employs the BERTopic model, which integrates BERT (Bidirectional Encoder Representations from Transformers) with contextual topic modeling, to analyze accident descriptions with minimal manual intervention. Unlike traditional frequency-based models, BERTopic captures richer bidirectional context [40], enabling more accurate interpretation of accident narratives. With this capability, the study automatically extracts key information such as accident scenarios and potential causes at a level of detail that earlier methods could not achieve. To address the limitations of previous OSHA-based text-mining research, this study employs BERTopic, which offers substantially better clustering performance than

LDA. By combining contextual embedding with large-scale OSHA construction narratives, the proposed approach uncovers nuanced risk scenarios, occupation and scenario patterns, and temporal trends that traditional LDA or keyword-driven analyses often overlook. This work also advances existing research by linking topics with demographic groups, high-risk occupations, and peak incident times, offering a more comprehensive understanding of accident mechanisms. As an open-source and fully reproducible framework, BERTopic ensures methodological transparency and supports future applications. The insights generated offered decision-makers actionable guidance for targeted risk mitigation and safety planning, ultimately contributing to accident reduction and strengthening confidence in construction safety practices.

In this study, scientometric analysis is adopted to identify latent topics, extract keyword patterns, and examine the temporal evolution of accident types using BERTopic. These quantitative insights reveal the structural and dynamic characteristics of construction accident reports. To complement this scientometric perspective, a qualitative review is conducted to interpret the underlying mechanisms behind each topic, including typical scenarios, contributing factors, vulnerable worker groups, peak accident times, and safety recommendations. Together, this mixed-methods design integrates data-driven topic discovery with domain-specific interpretation, providing a comprehensive understanding of construction accident patterns.

The remainder of this paper is structured as follows: the next section reviews the literature on construction accident report analysis, focusing on traditional and topic modelling approaches. The research methods are outlined in Section 3. The fourth section discusses the model selection process, followed by an analysis of accident types based on the topic modeling results. The fifth section presents a comprehensive discussion of the findings. Finally, concluding remarks, study limitations, and suggestions for future research directions are provided in the last section.

2. Literature Review

2.1. Traditional Analysis of Construction Accident Reports

Accident reports provide detailed accounts of accidents, and the recurrence of similar events underscores the importance of learning from past accidents to prevent future ones [41–44]. This principle also applies to the construction industry. For example, Chi et al. [45] developed a framework to analyze 255 fatal electrocution accidents in Taiwan. Their analysis, using Cramer's V and Phi statistics, revealed that unskilled workers faced higher risks, particularly in smaller companies, with most accidents involving improperly installed or damaged equipment. Similarly, Chi and Han [46] examined 9358 construction accidents in the U.S., sourced from OSHA, incorporating systems theory into the domino theory and using Chi-square and Fisher's exact tests to validate the link between construction risks and accident causation. Raviv et al. [47] employed K-means clustering to evaluate 51 tower crane accidents and 161 near misses, identifying technical failures as the most hazardous risk factor. Na et al. [48] applied a text mining method using a custom lexicon and the TF-H (term frequency–entropy) term importance measure to analyze 221 metro construction accident reports in China, identifying 37 key risk factors and primary accident causes.

Although these studies provide valuable insights into construction safety, they possess limitations. Traditional statistical analyses and predefined frameworks often struggle to capture the complexity, contextual nuances, and emerging risk factors inherent in construction safety. Moreover, reliance on specific, limited-sized datasets hinders the generalizability of findings across diverse contexts and evolving industry practices. These conventional approaches may limit the ability to uncover deeper, more complex patterns

and relationships [49]. In contrast, advanced methodologies, such as text mining and topic modelling, can analyze large volumes of unstructured text, revealing complex and hidden relationships, as well as emerging trends. These sophisticated algorithms detect special patterns and context-specific details that traditional methods may overlook, providing a more comprehensive understanding of the construction sector's safety risks and contributing factors [50,51].

2.2. Topic Modeling on Construction Accident Reports

Topic modeling is highly regarded for its interpretability, as it simplifies complex information into latent topics and patterns that traditional classification methods may overlook, facilitating the understanding of large datasets [50,52]. By extracting underlying topics from extensive text data, topic modeling provides a clearer perspective on complex information, which is particularly valuable in fields such as construction safety. For instance, Liu et al. [53] applied text mining, information entropy, TF-IDF, and LDA models to analyze Chinese construction accident reports, identifying key causal factors and improving safety management insights. Ertek and Kailas [54] developed a custom data analytics framework to analyze wind turbine accident reports, utilizing LDA and association mining to identify key themes related to fatalities and injuries. Min et al. [55] used LDA to analyze social networking service data on occupational accidents, highlighting concerns such as workers' compensation and injuries. Suh [56] applied text mining and LDA to identify five sectoral patterns in OSHA accident reports, addressing industry-specific safety issues. Kuhn [50] employed structural topic modeling to analyze over a million aviation safety reports, uncovering hidden patterns and issues within the data. Seo and Song [57] further applied LDA-based topic modeling to construction safety accident case data, focusing on incident records, post-incident actions, and preventive measures.

Although the LDA topic model has been extensively used and is valuable for identifying hidden topics in large text datasets, it has several limitations. LDA requires the pre-definition of the number of topics, which can be challenging to determine, and it relies on the bag-of-words model, which fails to capture contextual relationships within the text. Furthermore, LDA often requires extensive data preprocessing and tends to produce less coherent and interpretable topics, with diminished relevance and accuracy in matching topics to documents [58–60].

In contrast, research has demonstrated that transformer-based models, such as BERTopic, outperform LDA in several key areas [61,62]. BERTopic does not require the pre-definition of topic numbers and captures contextual relationships within text through embeddings, resulting in more coherent and relevant topics with simplified data preprocessing. These advantages make BERTopic a more flexible and effective choice, particularly when dealing with complex or evolving datasets [61,63,64].

2.3. Knowledge Gap

Despite BERTopic's advantages in extracting meaningful topics from large textual datasets, its use in analyzing construction industry accident reports remains limited. Most previous studies have relied on traditional approaches such as LDA or frequency-based text mining, which often fail to capture nuanced contextual relationships, temporal trends, and detailed occupation–scenario patterns. This highlights a clear gap for an advanced method that leverages contextual embeddings to generate interpretable topic clusters, uncover hidden patterns, and provide actionable insights for accident prevention. To address this, the present study applies BERTopic to a large-scale set of historical OSHA accident reports, using LDA as a baseline for comparison. The approach aims to identify latent accident patterns, associate topics with demographic and occupational factors, and track temporal

trends, thereby offering a deeper understanding of construction safety and supporting targeted risk mitigation strategies.

3. Research Methods

This section covers data collection and labelling, data processing, topic modelling, model evaluation, topic analysis, and comprehensive discussion, as illustrated in Figure 1.

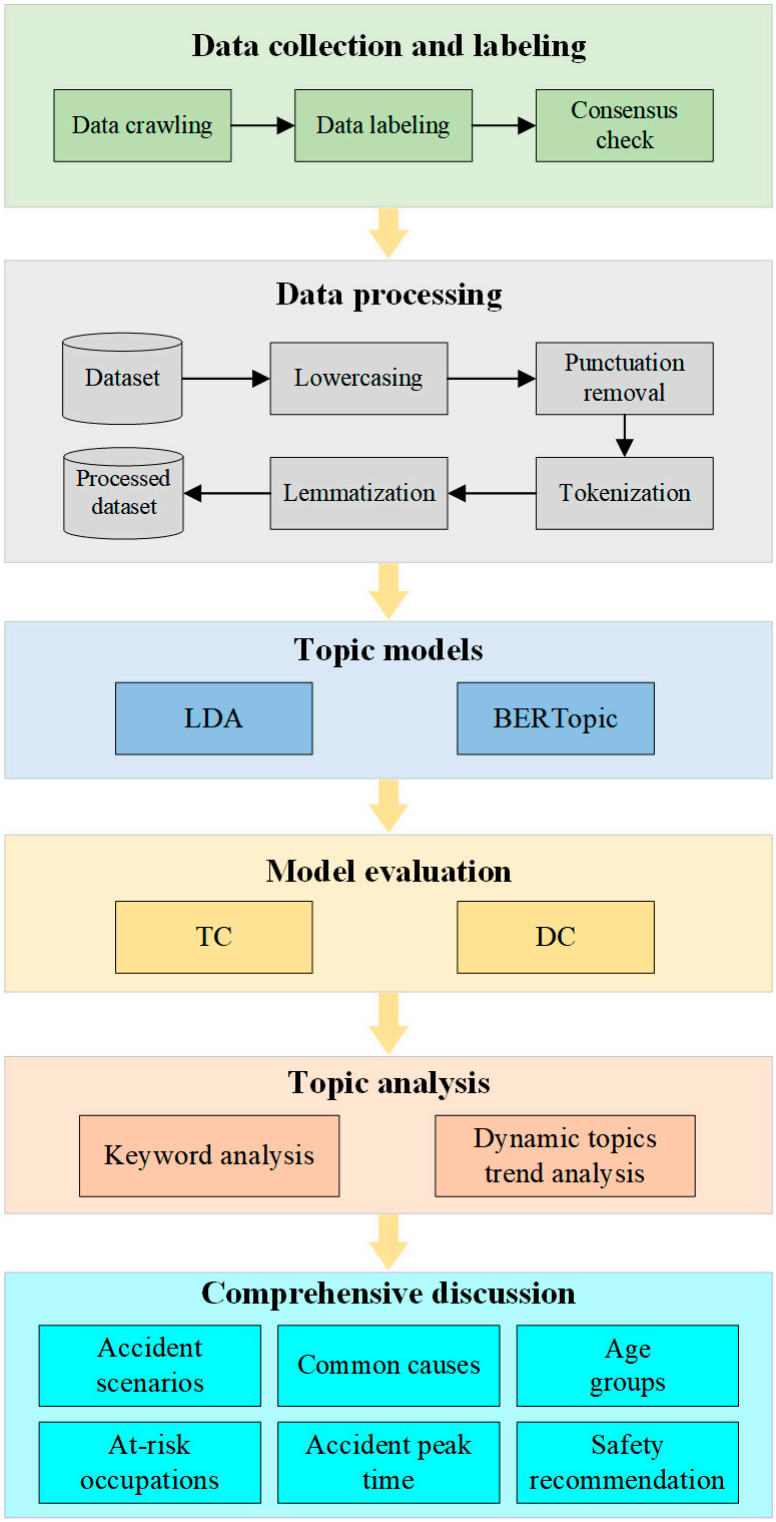

Figure 1. Research methods flowchart.

3.1. Data Collection and Labeling

Safety accident data were collected using a Python-based web scraper from OSHA's Investigation Summary Portal, specifically focusing on the construction industry as classified under the North American Industry Classification System (NAICS) code, where sector 23 represents the construction industry. The dataset spans from 28 May 2004, to 30 January 2023. Each accident was categorized based on the Event or Exposure classifications outlined in the Occupational Injury and Illness Classification (OIICS) manual. These categories include violent acts, transportation incidents, explosions and fires, falls–slips–trips, exposure to harmful substances or environments, contact with objects or equipment, overexertion, bodily reaction, and non-classifiable incidents.

Three experts received rigorous training on the annotation protocol to ensure accuracy and consistency. All annotators were members of the research team with formal training in construction safety management. The independent annotators were graduate students who had completed OSHA construction safety management training. At the same time, the adjudicator had also completed OSHA training and held a doctoral degree with over five years of relevant teaching experience. The manual annotation process consisted of two key stages. First, 1100 accident reports were randomly selected for iterative annotations. Multiple rounds of coding and agreement assessment were performed until inter-coder agreement exceeded the 80% threshold recommended in the literature [65]. After three rounds, an agreement was reached at 98.36%, demonstrating strong reliability. This finding was further supported by a Cohen's Kappa of 0.96 ($p \leq 0.01$), indicating almost perfect agreement, as described by Landis and Koch [66].

The distribution of accidents across categories was as follows: violent acts (10), transportation accidents (1298), explosions and fires (395), falls–slips–trips (10,726), exposure to harmful substances or environments (2619), contact with objects or equipment (6885), overexertion and bodily reaction (569), and non-classifiable (121).

3.2. Data Processing

The 121 non-classifiable accidents were initially excluded from the dataset. Several accident categories, including violent acts, transportation accidents, explosions and fires, and overexertion and bodily reactions, contained relatively few cases. This limited data diversity and potentially hindered BERTopic's ability to generate stable and meaningful topics. Sparse observations in these minority categories may produce poorly defined clusters, overly generalized topics, or embeddings that fail to capture subtle contextual differences among rare events [67,68]. To mitigate these issues, these categories were merged into a single dataset. Consequently, four distinct datasets were prepared for further analysis: falls, slips, and trips (10,726 accidents); contact with objects or equipment (6885 accidents); exposure to harmful substances or environments (2619 accidents); and a combined category including transportation, overexertion, explosion, and violence (2272 accidents).

Each dataset underwent a series of preprocessing steps. The text was converted to lowercase for uniformity, punctuation was removed to ensure focus on the core content, tokenized to facilitate analysis, and words were lemmatized to their base forms to maintain consistency. These preprocessing steps are critical for enhancing both the accuracy and effectiveness of the subsequent topic modeling analysis. During preprocessing, all text was lemmatized and stripped of spaces and punctuation before topic modeling. As a result, certain tokens in the topic tables may appear unconventional (e.g., "instal", "cavein", "sawer"). These tokens correspond to their natural-language forms "install", "cave in", and "saw worker".

3.3. Topic Modeling

Topic modeling identifies underlying topics within a text corpus by classifying documents into latent themes based on content. This method detects hidden trends and patterns using key terms, making it valuable in information retrieval and NLP. Topic modeling can be conducted using supervised, semi-supervised, or unsupervised learning methods [30,69–71]. An unsupervised approach is employed to extract interpretable themes from a collection of documents, allowing the identification of hidden patterns and structures within the data without the need for predefined topic names. This approach effectively organizes large datasets, facilitating the extraction of critical insights and supporting the development of targeted safety interventions to improve workplace safety practices.

3.3.1. LDA

LDA, proposed by Blei et al. [68], is one of the most widely applied topic models, as illustrated in Figure 2. In the LDA framework, each document is viewed as a mixture of topics, and each topic is represented as a distribution over words [72]. The model operates as follows: α, a Dirichlet prior, influences the topic distribution θ_m for each document m. For every document, θ_m is drawn from α, representing the proportions of topics. For each word n in the document, a topic $Z_{m,n}$ is selected based on θ_m, after which a word $W_{m,n}$ is generated from the word distribution φ_k corresponding to the selected topic k. The variable φ_k itself is drawn from another Dirichlet prior, β, which governs the word distribution for each topic.

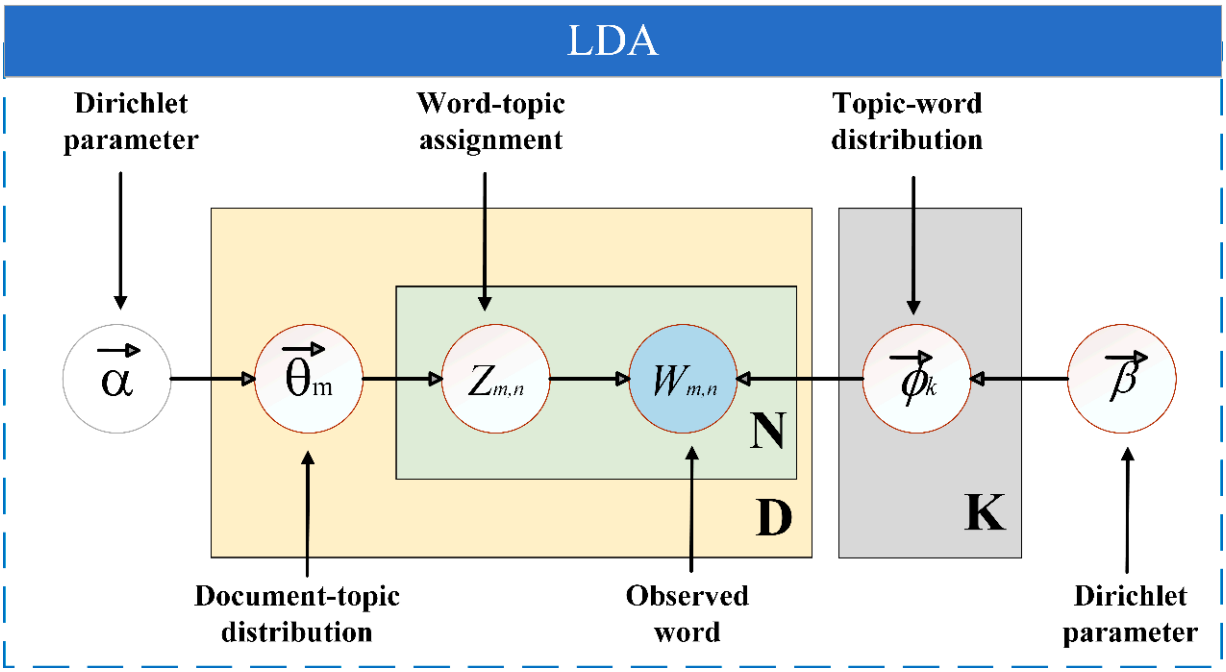

Figure 2. Graphical model representation of LDA.

In this model, $W_{m,n}$ represents the observed variable (the actual words in the document), while the other variables—θ_m, $Z_{m,n}$, and φ_k—are latent and need to be inferred. The diagram utilizes plates to represent repeated processes: **D** represents the number of documents, **N** refers to the total number of words in each document, and **K** denotes the number of topics. Model hyperparameters were optimized using a grid-search procedure to identify configurations that achieve an effective balance between topic coherence and topic diversity. The search ranges and the final parameter settings for each dataset are summarized in Table 1.

Table 1. Overview of LDA hyperparameter settings.

Dataset	n_Topics [1–10]	Alpha [0–1]	Beta [0–1]	Iterations [50–500]	Passes [5–50]
EXPO *	7	0.07	0.10	100	20
TRAN *	6	0.06	0.40	200	15
CONT *	8	0.07	0.50	100	30
FALL *	9	0.08	0.60	300	40

* EXPO: Exposure to harmful substances or environments; TRAN: Transportation, overexertion, explosion, and violence; CONT: Contact with objects or equipment; FALL: Falls–slips–trips.

However, LDA presents several limitations, including the need for subjective decisions regarding the number of topics, an overemphasis on word frequency, and the disruption of text coherence during preprocessing, which can result in less accurate and interpretable topics [60]. These shortcomings have driven the development of more advanced models, such as BERTopic, which addresses these issues by capturing contextual relationships between words and providing greater flexibility in topic modeling [61].

3.3.2. BERTopic

BERTopic employs the BERT model to extract high-dimensional text vectors, which are subsequently clustered to identify topics within the data [73]. By utilizing BERT embeddings and class-based Term Frequency–Inverse Document Frequency (c-TF-IDF), BERTopic generates dense clusters that reveal thematic structures without necessitating predefined categories. This data-driven approach enhances the interpretability of topics, offering a clearer understanding of relationships within large corpora [74,75].

A key advantage of BERTopic is its implementation of HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), a clustering algorithm that automatically determines the optimal number of clusters based on the data. Unlike LDA, which requires a fixed number of topics to be set in advance, HDBSCAN identifies dense regions within the data and interprets them as topics. This enables BERTopic to automatically adapt to the structure of diverse datasets, eliminating the need to predefine the number of topics, thereby enhancing its flexibility and applicability across various types of text corpora [76]. This study was implemented using Python 3.10, with BERTopic 0.16.4, UMAP-learn 0.5.7, HDBSCAN 0.8.39, PyNNDescent 0.5.13, and Sentence-Transformers 3.2.1. To ensure full reproducibility, a global random seed of 42 was applied across all stochastic components of the pipeline.

Figure 3 illustrates the primary steps of the BERTopic process. For each stage, hyperparameters were tuned using a grid search strategy to achieve an optimal balance between topic coherence and topic diversity, while adhering to standard methodological guidelines [77]. The corresponding search ranges and the final hyperparameter configurations for each dataset are summarized in Table 2.

Table 2. Overview of BERTopic hyperparameter settings.

Dataset	UMAP n_Neighbors [1–50]	UMAP n_Components [1–10]	UMAP Min_Dist [0–1]	HDBSCAN Min_Cluster_Size [1–100]	HDBSCAN Min_Samples [1–15]	Count Vectorizer Max_Features [1000–15,000]	MMR Diversity [0–1]
EXPO *	30	2	0.2	10	10	5000	0.9
TRAN *	5	2	0.3	10	10	5000	0.9
CONT *	10	3	0.1	50	2	8000	0.9
FALL *	30	2	0.3	50	2	10,000	0.9

* EXPO: Exposure to harmful substances or environments; TRAN: Transportation, overexertion, explosion, and violence; CONT: Contact with objects or equipment; FALL: Falls–slips–trips.

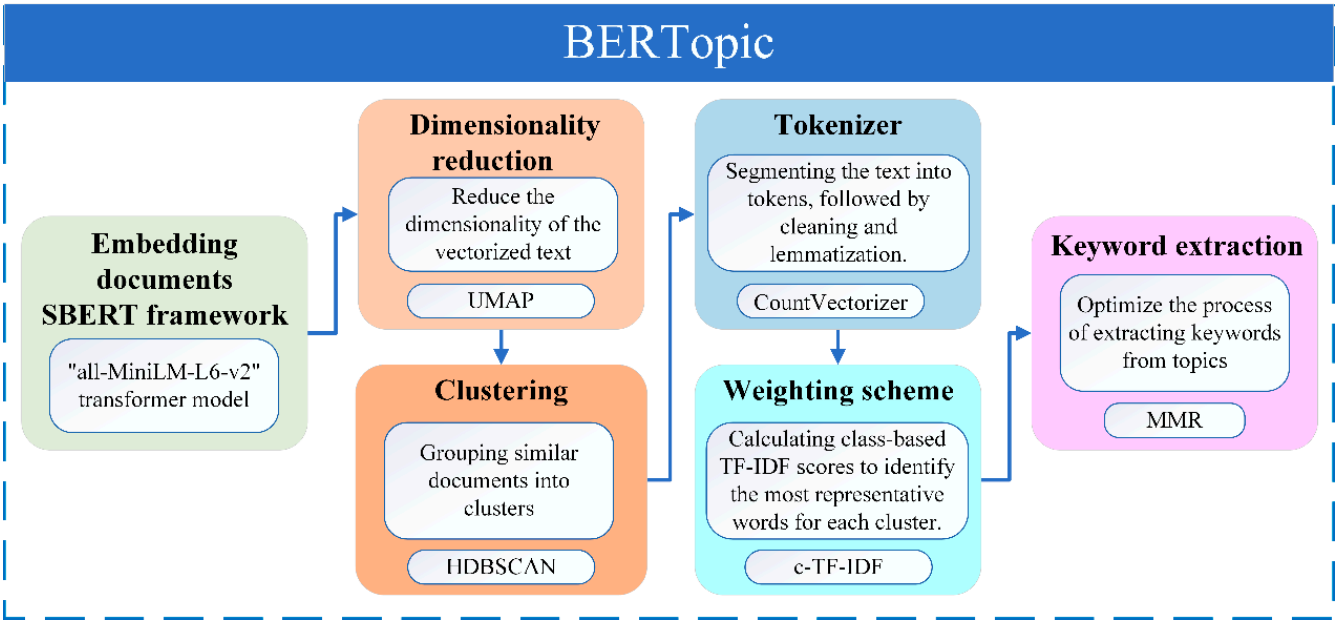

Figure 3. Main steps of the BERTopic method.

1. Embedding: Text data are converted into 384-dimensional vectors using the "all-MiniLM-L6-v2" transformer model, an efficient, pre-trained model suitable for NLP tasks in resource-limited environments [78].
2. Dimensionality reduction: UMAP (Uniform Manifold Approximation and Projection) reduces the dimensionality of these vectors, simplifying clustering while preserving semantic relationships between texts.
3. Clustering: HDBSCAN is applied to the reduced vectors, grouping similar documents and identifying potential topics.
4. Tokenization: The CountVectorizer tokenises the documents, and preprocessing steps, such as lemmatization and stopword removal, ensure that keywords accurately represent each topic, improving clarity and interpretability.
5. Weighting Scheme: BERTopic uses c-TF-IDF to weight keywords within each cluster. Unlike traditional TF-IDF, c-TF-IDF treats all documents in a cluster as a single entity, highlighting terms that are significant to specific topics while rare in others, thereby enhancing topic distinctiveness. The c-TF-IDF formula is shown in Equation (1)

$$\text{c}-\text{TF}-\text{IDF}(t,c) = TF(t,c) \times IDF(t) \quad (1)$$

where t is the term, and c is the class (topic). $TF(t,c) = f_{t,c}/L_c$ represents the term frequency, where $f_{t,c}$ is the number of occurrences of term t in all documents of class c, and L_c is the total number of terms in class c. $IDF(t) = \log(N/(1+DF(t)))$ measures how rare a term is across all classes, where N is the total number of classes (topics) and $DF(t)$ is the number of classes containing term t. The product $TF(t,c) \times IDF(t)$ highlights terms that are frequent within a specific topic but infrequent across other topics, thus enhancing the distinctiveness and interpretability of topic keywords.
6. Keyword extraction with MMR: Maximal Marginal Relevance (MMR) balances relevance and diversity in keyword selection, ensuring that selected keywords are representative and varied, improving overall topic interpretability.

3.4. Model Evaluation

Two metrics were used to evaluate the performance of the topic models: topic coherence (TC) and topic diversity (TD) [63,75]. TC measures how well the words within a topic

co-occur in similar contexts, enhancing the topic's interpretability and meaning [75,76]. It is calculated using the log ratio of co-occurrence and document frequency for the top N words in a topic, with the C_v coherence measure being commonly used due to its strong alignment with human judgment [79,80]. TC scores range from −1 to 1, with higher scores indicating more coherent and meaningful topics [81,82].

TD assesses the variation among topics by calculating the proportion of unique words within the top words across all topics. TD scores range from 0 to 1, where higher values indicate more distinct topics and lower scores suggest overlapping topics [83]. Balancing TC and TD is important, as excessive diversity can reduce coherence, making topics harder to interpret. Therefore, finding an optimal balance between these two metrics ensures that the topics are both coherent and diverse, leading to more meaningful insights. However, when TC and TD yield conflicting results, TC is prioritized to ensure the generation of interpretable and coherent topics [84].

3.5. Topic Analysis

The application of BERTopic in accident report analysis emphasizes two key areas: topic trend analysis within each category and keyword distribution analysis for each topic. Topic trend analysis uncovers how different topics evolve, revealing shifts in risk patterns and emerging issues. The temporal analysis was performed using the Event Date field from the OSHA reports. All entries were first standardized to a uniform date format, and any missing or invalid records were excluded to ensure the accuracy and reliability of subsequent temporal analyses. By calculating the annual number of documents associated with each topic for every dataset, this temporal analysis provides actionable insights, enabling safety professionals to refine their safety strategies, identify and prioritize new hazards, and respond effectively to evolving industry risks.

Keyword distribution analysis identifies significant keywords associated with each topic, and the order of these keywords is determined by their cosine similarity to the topic center vector. Cosine similarity is a measure of the semantic closeness between two vectors, calculated as the cosine of the angle between them. In this context, the topic center vector is obtained by aggregating the vectors of all documents within the topic, while each keyword has its own vector representation. A higher cosine similarity indicates that the keyword is more semantically aligned with the topic, making it more representative of the topic's core content. By ranking keywords according to this measure, the model highlights the most informative and meaningful terms, allowing for a rapid understanding of each topic's essence and capturing deeper semantic relationships that extend beyond simple word frequency. Identifying these patterns enables safety professionals to develop targeted interventions, enhance training programs, and implement more effective prevention strategies.

3.6. Comprehensive Discussion

Analyzing topics generated by BERTopic provides deeper insights into accident scenarios and potential causes. Additionally, the analysis of related original accident reports reveals the age groups, at-risk occupations, and peak accident times for each topic within each category. Finally, specific safety recommendations are suggested through a comprehensive literature review. This section aims to enhance safety standards and strengthen worker protection within the construction industry by identifying critical safety concerns, understanding the profiles of affected workers, and proposing targeted safety recommendations.

4. Results

4.1. Model Selection

The parameters of the BERTopic model for each accident dataset were optimized using a grid search approach within specified parameter ranges. The grid search method systematically explored various parameter combinations to identify the most effective configuration for each dataset, balancing the number of generated topics with the coherence and diversity of the resulting clusters. This approach ensured optimal topic modeling tailored to the unique characteristics of each accident dataset.

The same preprocessed datasets were subsequently input into the LDA model, where the optimal number of topics for each accident type was determined by testing various topic numbers and selecting the configuration with the lowest perplexity. Following this, TC and TD scores were calculated for both the BERTopic and LDA models, with the results summarized in Table 3. The model demonstrating superior performance for each dataset is highlighted in bold.

Table 3. TC and TD scores of both the BERTopic and the LDA model for each dataset.

Dataset	BERTopic-TC	BERTopic-TD	LDA-TC	LDA-TD
EXPO *	**0.7038**	**0.8333**	0.4622	0.6868
TRAN *	**0.5710**	**0.9000**	0.4295	0.7157
CONT *	**0.6113**	**0.7667**	0.4212	0.7552
FALL *	**0.5018**	0.6222	0.4065	**0.7578**

* EXPO: Exposure to harmful substances or environments; TRAN: Transportation, overexertion, explosion, and violence; CONT: Contact with objects or equipment; FALL: Falls–slips–trips.

Table 3 presents the TC and TD scores of BERTopic and LDA across the four datasets, with the better-performing model highlighted in bold. Overall, BERTopic consistently achieves higher TC scores compared to LDA, indicating that the topics it generates are more semantically coherent and better capture the underlying relationships within the data. For instance, in the dataset on exposure to harmful substances or environments, BERTopic attains a TC score of 0.7038, substantially higher than LDA's 0.4622. This trend is observed across all datasets, reflecting BERTopic's superior ability to produce meaningful and interpretable topics.

In terms of TD, BERTopic generally maintains comparable or higher scores relative to LDA, suggesting that it captures a wide range of distinct topics without excessive overlap. For example, in the transportation, overexertion, explosion, and violence dataset, BERTopic achieves a TD score of 0.9000, higher than LDA's 0.7157. An exception is observed in the falls–slips–trips dataset, where LDA exhibits a slightly higher Topic Diversity score of 0.7578 compared to BERTopic's 0.6222, indicating that while LDA produces a broader spread of topics, the semantic coherence of these topics is lower.

These results highlight BERTopic's ability to balance topic interpretability and diversity, making it particularly suitable for analyzing complex accident reports with short, context-dependent texts and specialized terminology. The quantitative comparison demonstrates that BERTopic not only produces more coherent topics but also effectively captures the nuanced semantic structure of each dataset, providing stronger support for the analysis of construction safety incidents compared to traditional LDA modeling.

This overall superiority is attributable to BERTopic's ability to leverage contextual embeddings and density-based clustering, enabling it to capture subtle semantic relationships within accident narratives. Unlike LDA, which relies on bag-of-words assumptions and often struggles with short or sparsely expressed accident descriptions, BERTopic embeds text into a semantic vector space and groups documents based on similarity patterns, resulting in more coherent topic structures.

Given its stronger performance and its ability to better reflect the semantic and contextual properties of accident reports, BERTopic was adopted as the primary topic modeling method for subsequent analysis. The extracted keywords from BERTopic were then analyzed to interpret topic themes and summarize the underlying patterns associated with each accident category.

Using the Falls slips trips dataset as an illustrative example in Table 4, the topic quality of BERTopic and LDA was compared using the key indicators TC and TD. BERTopic achieved a TC of 0.5018 and a TD of 0.6222, demonstrating a balanced ability to generate both coherent and diverse topics. In contrast, LDA produced a lower TC of 0.4065 but a higher TD of 0.7578, indicating that while its topics are more diverse, their semantic consistency is weaker. These results highlight BERTopic's advantage in capturing meaningful structures within accident narratives.

Table 4. Comparison of BERTopic and LDA topic words based on the falls, slips, and trips dataset.

Topic No.	BERTopic Words	LDA_Topic Words
0	['roof', 'ladder', 'fracture', 'foot', 'construction', 'protection', 'head', 'floor', 'concrete', 'instal']	['fall', 'floor', 'ft', 'fracture', 'opening', 'construction', 'work', 'approximately', 'concrete', 'foot']
1	['scaffold', 'fracture', 'platform', 'plank', 'construction', 'foot', 'collapse', 'protection', 'head', 'ground']	['trench', 'excavation', 'manhole', 'sewer', 'deep', 'excavator', 'vault', 'asphyxiate', 'pipe', 'pave']
2	['truss', 'collapse', 'construction', 'instal', 'fracture', 'roof', 'plate', 'set', 'protection', 'top']	['fall', 'floor', 'ft', 'fracture', 'opening', 'construction', 'work', 'approximately', 'concrete', 'foot']
3	['skylight', 'roof', 'unguarded', 'protection', 'concrete', 'floor', 'step', 'head', 'foot', 'blunt']	['fall', 'steel', 'tower', 'beam', 'kill', 'protection', 'work', 'lanyard', 'crane', 'approximately']
4	['shock', 'electric', 'electrical', 'light', 'ladder', 'contact', 'fixture', 'energize', 'power', 'line']	['skylight', 'tree', 'barn', 'fiberglass', 'warehouse', 'corrugate', 'limb', 'unguarded', 'branch', 'panel']
5	['elevator', 'shaft', 'floor', 'door', 'car', 'fracture', 'construction', 'foot', 'hoist', 'platform']	['roof', 'fall', 'fracture', 'construction', 'protection', 'foot', 'work', 'roofer', 'hospital', 'employ']
6	['tower', 'communication', 'cable', 'antenna', 'foot', 'ground', 'instal', 'electrical', 'cell', 'ladder']	['scaffold', 'fall', 'platform', 'work', 'lift', 'elevate', 'kill', 'foot', 'construction', 'aerial']
7	['lanyard', 'harness', 'roof', 'protection', 'lifeline', 'rope', 'fracture', 'wear', 'attach', 'foot']	['head', 'fall', 'kill', 'trauma', 'die', 'blunt', 'force', 'injury', 'struck', 'concrete']
8		['ladder', 'fall', 'fracture', 'foot', 'approximately', 'extension', 'work', 'lose', 'ft', 'hospitalize']

More broadly, BERTopic outperforms LDA in modeling construction accident descriptions related to falls. LDA frequently generates topics dominated by generic, high-frequency terms such as "fall," "fracture," "foot," "construction," and "work," which leads to substantial thematic overlap and limited distinctiveness. As a result, LDA cannot effectively differentiate critical fall scenarios such as roof edge falls, scaffold collapses, skylight openings, elevator shaft incidents, or ladder misuse—scenarios that BERTopic identifies more clearly. By leveraging contextual embeddings and density-based clustering, BERTopic produces more coherent, scenario-specific, and interpretable topics with minimal redundancy. Overall, the comparison reveals that LDA yields less reliable and less informative thematic structures, whereas BERTopic provides clearer, more meaningful, and domain-relevant insights for construction accident analysis.

4.2. Topic Analysis

4.2.1. Falls–Slips–Trips

Falls, slips, and trips represent the leading causes of injuries and fatalities in the U.S. construction industry [85,86]. The topic analysis results for this category are presented in Figure 4, accompanied by the topic similarity matrix in Figure 5. The corresponding topic model outputs are summarized in Table 5. Fall-related accidents slightly increased in 2004, followed by a steady decline until 2016. However, a sharp rise occurred thereafter, peaking in 2019 before declining and stabilizing. Data for 2023 remains incomplete. For the falls–slips–trips dataset, a cross-topic examination reveals several shared structural patterns as well as important distinctions that cannot be captured through isolated topic descriptions.

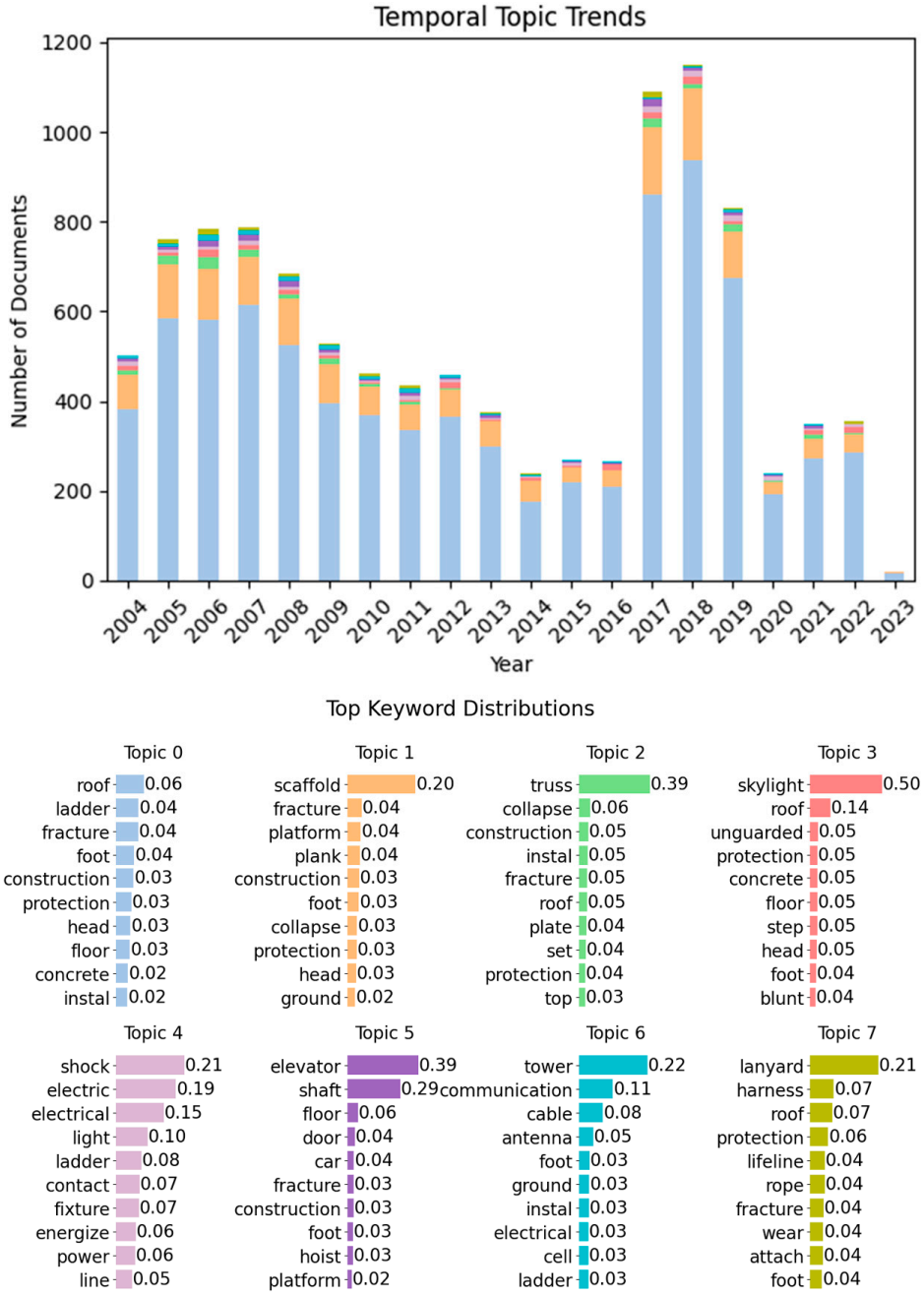

Figure 4. Results of topic analysis of the falls–slips–trips category.

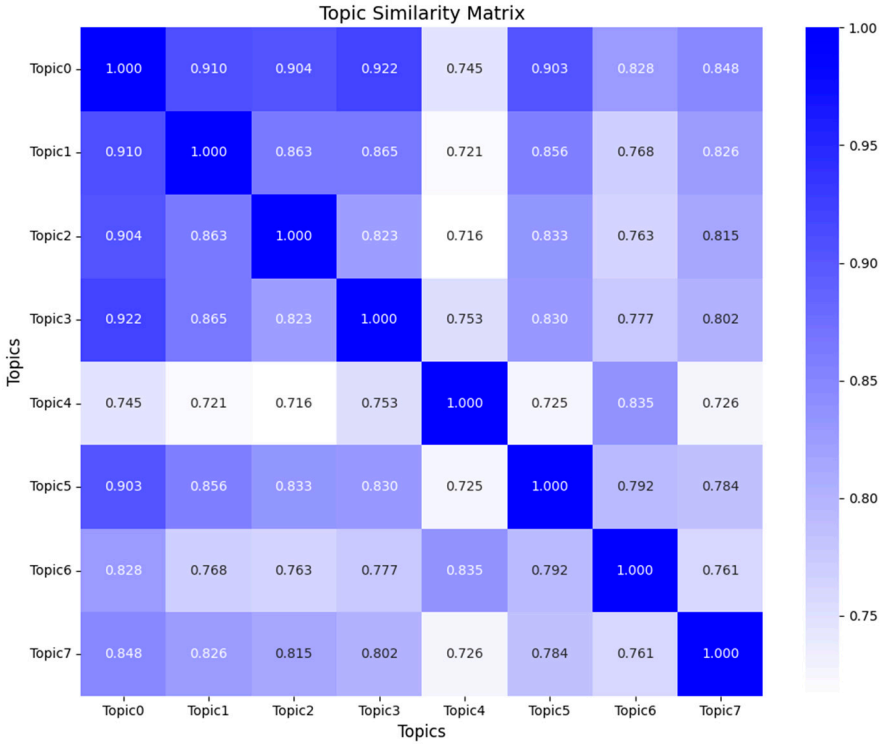

Figure 5. Topic similarity matrix of the falls–slips–trips dataset.

Table 5. BERTopic outputs on the dataset of falls, slips, and trips.

Topic Name	Count	Representation *
0_roof_ladder_fracture_foot	8292	[‘roof’, ‘ladder’, ‘fracture’, ‘foot’, ‘construction’, ‘protection’, ‘head’, ‘floor’, ‘concrete’, ‘instal’]
1_scaffold_fracture_platform_plank	1479	[‘scaffold’, ‘fracture’, ‘platform’, ‘plank’, ‘construction’, ‘foot’, ‘collapse’, ‘protection’, ‘head’, ‘ground’]
2_truss_collapse_construction_instal	317	[‘truss’, ‘collapse’, ‘construction’, ‘instal’, ‘fracture’, ‘roof’, ‘plate’, ‘set’, ‘protection’, ‘top’]
3_skylight_roof_unguarded_protection	172	[‘skylight’, ‘roof’, ‘unguarded’, ‘protection’, ‘concrete’, ‘floor’, ‘step’, ‘head’, ‘foot’, ‘blunt’]
4_shock_electric_electrical_light	141	[‘shock’, ‘electric’, ‘electrical’, ‘light’, ‘ladder’, ‘contact’, ‘fixture’, ‘energize’, ‘power’, ‘line’]
5_elevator_shaft_floor_door	122	[‘elevator’, ‘shaft’, ‘floor’, ‘door’, ‘car’, ‘fracture’, ‘construction’, ‘foot’, ‘hoist’, ‘platform’]
6_tower_communication_cable_antenna	116	[‘tower’, ‘communication’, ‘cable’, ‘antenna’, ‘foot’, ‘ground’, ‘instal’, ‘electrical’, ‘cell’, ‘ladder’]
7_lanyard_harness_roof_ protection	87	[‘lanyard’, ‘harness’, ‘roof’, ‘protection’, ‘lifeline’, ‘rope’, ‘fracture’, ‘wear’, ‘attach’, ‘foot’]

* Due to lemmatization and the removal of spaces and punctuation during preprocessing, some tokens (e.g., instal) may appear unconventional. This token corresponds to install.

The similarity matrix indicates that Topics 0 to 3 form a highly cohesive cluster, with similarity values exceeding 0.82. These topics, although presented through different surface events such as ladder missteps, scaffold failures, truss collapses, and unguarded skylights, share common underlying mechanisms. These mechanisms include unstable walking surfaces, insufficient fall protection, and edge hazards during elevated work. This suggests that safety strategies for these scenarios should prioritize unified engineering measures, including standardized edge protection, enhanced anchorage design, and consistent procedures for setting up ladders and scaffolds. A second grouping can be observed among Topics 4 to 6. These topics exhibit lower similarity to the main cluster but stronger internal

similarity. They capture mixed-mechanism incidents involving electrical contact, tower climbing, and the installation of communication cables. Although categorized as fall-related injuries, these incidents often arise from task-specific hazards that involve simultaneous climbing and interaction with energized components. These patterns highlight the importance of integrated hazard control, which combines fall arrest protection with electrical isolation, rather than addressing each hazard independently. Topic 7 exhibits moderate similarity to both clusters, suggesting that lanyard and harness failures occur across a wide range of working environments, including roofs and towers. This pattern suggests that improper harness use and inadequate anchorage selection are pervasive issues not limited to a specific task or occupation. Strengthening training on harness inspection, tie-off procedures, and anchor-point selection could therefore contribute to safety improvements across multiple scenarios.

4.2.2. Contact with Objects or Equipment

Several key topics emerge in the category of contact with objects or equipment, as illustrated in Figure 6, and the topic similarity matrix is shown in Figure 7. The corresponding topic model outputs are summarized in Table 6. The trend analysis of contact accidents indicates a slight increase in 2004, followed by a decline starting in 2008, which reached its lowest point in 2016. A significant rise occurred in 2017, followed by another decline starting in 2019, which eventually stabilized. A cross-topic analysis of the contact with objects or equipment dataset shows several shared mechanisms that extend beyond individual incident scenarios. The similarity matrix indicates that Topics 0, 1, 4, 5, 6, and 7 form a moderately cohesive cluster, with similarity values ranging from 0.75 to 0.89. Although these topics involve different tools such as saws, grinders, nail guns, and falling tree limbs, they reflect a common pattern linked to high-energy contact between workers and moving objects or materials. The recurring mechanisms include loss of tool control, unexpected object movement, kickback from rotating blades, and inadequate awareness of the line of fire. These findings suggest that prevention strategies emphasize energy-control principles, improved guarding, and consistent tool-handling protocols in a wide range of construction tasks.

Table 6. BERTopic outputs on the dataset of contact with objects or equipment.

Topic Name	Count	Representation *
0_struck_fall_crush_construction	4929	[‘struck’, ‘fall’, ‘crush’, ‘construction’, ‘fracture’, ‘truck’, ‘caught’, ‘lift’, ‘crane’, ‘concrete’]
1_saw_cut_blade_finger	712	[‘saw’, ‘cut’, ‘blade’, ‘finger’, ‘amputate’, ‘hand’, ‘circular’, ‘thumb’, ‘kick’, ‘wood’]
2_trench_collapse_excavation_cavein	655	[‘trench’, ‘collapse’, ‘excavation’, ‘cavein’, ‘bury’, ‘soil’, ‘wall’, ‘foot’, ‘deep’, ‘sawer’]
3_nail_gun_puncture_knee	276	[‘nail’, ‘gun’, ‘puncture’, ‘knee’, ‘shot’, ‘pneumatic’, ‘tool’, ‘construction’, ‘discharge’, ‘carpenter’]
4_tree_fall_limb_struck	108	[‘tree’, ‘fall’, ‘limb’, ‘struck’, ‘branch’, ‘cut’, ‘excavator’, ‘chain’, ‘object’, ‘head’]
5_saw_kick_cut_back	85	[‘saw’, ‘kick’, ‘cut’, ‘back’, ‘neck’, ‘pipe’, ‘laceration’, ‘blade’, ‘face’, ‘struck’]
6_eye_nail_struck_puncture	65	[‘eye’, ‘nail’, ‘struck’, ‘puncture’, ‘construction’, ‘gun’, ‘ricochet’, ‘tool’, ‘left’, ‘face’]
7_grinder_wheel_grind_cut	55	[‘grinder’, ‘wheel’, ‘grind’, ‘cut’, ‘laceration’, ‘angle’, ‘kick’, ‘hand’, ‘use’, ‘portable’]

* Due to lemmatization and the removal of spaces and punctuation during preprocessing, some tokens (e.g., cavein, sawer) may appear unconventional. These correspond to cave in and saw worker.

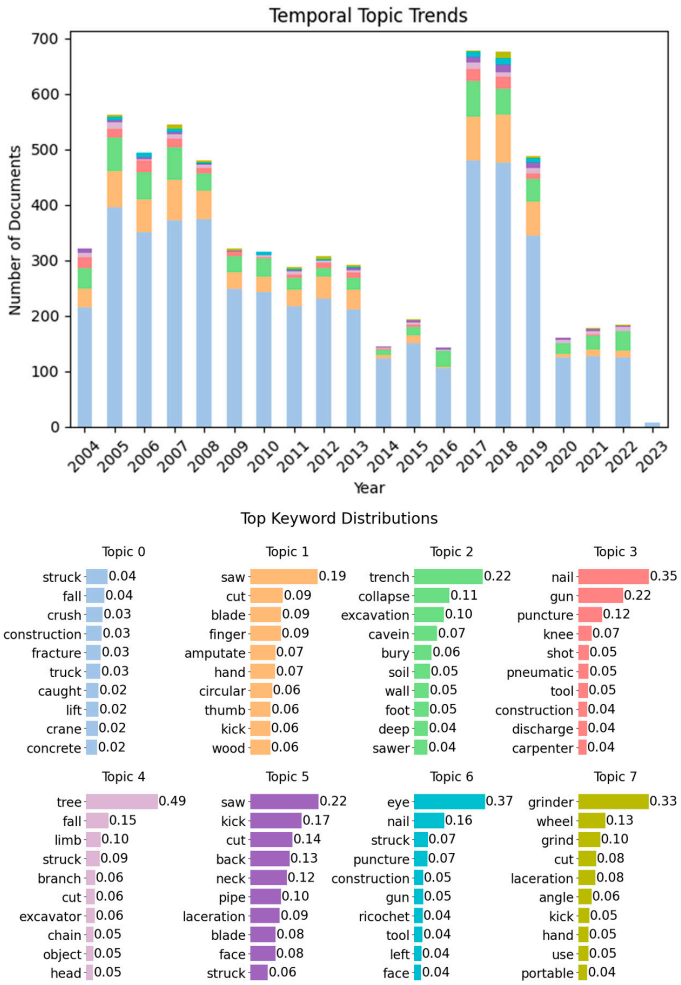

Figure 6. Results of topic analysis of the contact with objects or equipment category.

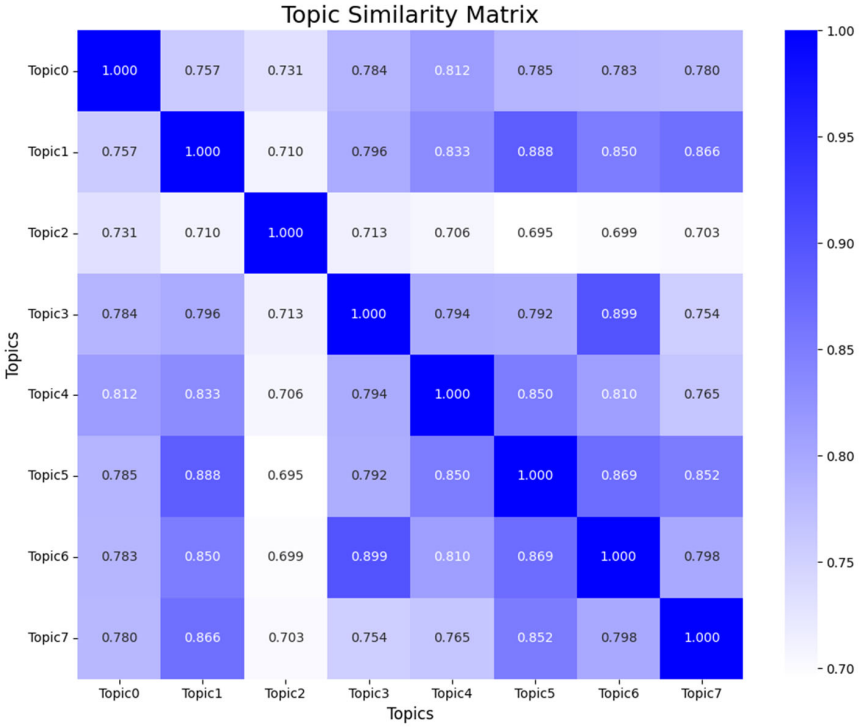

Figure 7. Topic similarity matrix of the contact with objects or equipment dataset.

Topic 2 stands apart from the other topics, showing the lowest similarity values. It represents trench collapses and cave-in incidents that involve large-scale material failure and engulfment hazards rather than sudden mechanical contact. This distinction underscores the necessity for excavation-specific interventions, including soil stability analysis, protective trench systems, and continuous monitoring of environmental conditions. Topic 3 shows an intermediate pattern. Nail gun puncture injuries often involve both projectile force and unstable working positions. This reflects the combined influences of tool mechanics and human factors such as recoil control, alignment, and body positioning. These mechanisms differ from both the tool kickback patterns and the material collapse patterns found in the other topics.

The topic structure also reveals several non-intuitive patterns. For example, saw kickback injuries, as represented by Topics 1 and 5, disproportionately affect workers performing occasional cutting tasks rather than experienced carpenters, indicating gaps in training for intermittent tool use. Grinder-related injuries in Topic 7 frequently occur during brief adjustments rather than planned cutting activities, suggesting that momentary lapses in hazard awareness remain a persistent issue. These findings emphasize the need for micro-training modules that focus on short-duration but high-risk interactions with tools.

From the perspective of safety strategy, the cross-topic relationships indicate that contact injuries occur through both frequent, small-scale interactions with handheld tools and less frequent, large-scale collapse events. Prevention, therefore, requires reinforcing standard controls for tool operation, such as two-hand switches, anti-kickback mechanisms, and blade guards, while also implementing specialized engineering controls for excavation work. The wide distribution of struck-by and crush incidents within Topic 0 indicates persistent exposure to uncontrolled object movement across multiple construction settings, underscoring the need to strengthen communication protocols, exclusion zones, and load-handling procedures.

4.2.3. Exposure to Harmful Substances or Environments

In the category of exposure to harmful substances or environments, several key topics emerge, as illustrated in Figure 8, and the topic similarity matrix is shown in Figure 9. The corresponding topic model outputs are summarized in Table 7. The overall trend indicates a slight increase in accidents starting in 2004, followed by a gradual decrease beginning in 2006, reaching its lowest point in 2016. A significant rise occurred in 2017, followed by a sharp decrease in 2019, with stabilization observed from 2020 onwards. A cross-topic examination of the exposure to harmful substances or environments dataset shows several coherent patterns that become clearer when moving beyond individual topic descriptions. The similarity matrix reveals two main clusters, along with several intermediate topics, suggesting that different surface-level incidents often share common exposure mechanisms.

Table 7. BERTopic outputs on the dataset of exposure to harmful substances or environments.

Topic Name	Count	Representation *
0_electrical_electrocute_line_power	1739	[‘electrical’, ‘electrocute’, ‘line’, ‘power’, ‘electric’, ‘shock’, ‘contact’, ‘burn’, ‘overhead’, ‘energize’]
1_heat_exhaustion_stroke_illness	295	[‘heat’, ‘exhaustion’, ‘stroke’, ‘illness’, ‘heatrelated’, ‘temperature’, ‘day’, ‘construction’, ‘hospital’, ‘roof’]
2_drown_water_boat_river	186	[‘drown’, ‘water’, ‘boat’, ‘river’, ‘dive’, ‘barge’, ‘pond’, ‘pipe’, ‘lake’, ‘swim’]
3_monoxide_carbon_poison_saw	81	[‘monoxide’, ‘carbon’, ‘poison’, ‘saw’, ‘concrete’, ‘fume’, ‘generator’, ‘gas’, ‘exhaust’, ‘ventilation’]

Table 7. *Cont.*

Topic Name	Count	Representation *
4_manhole_oxygen_confine_enter	69	[‘manhole’, ‘oxygen’, ‘confine’, ‘enter’, ‘space’, ‘sewer’, ‘hydrogen’, ‘gas’, ‘asphyxiate’, ‘sulfide’]
5_hot_asphalt_tar_roof	63	[‘hot’, ‘asphalt’, ‘tar’, ‘roof’, ‘burn’, ‘hand’, ‘bucket’, ‘roofer’, ‘arm’, ‘kettle’]
6_covid_infection_infectious_disease	62	[‘covid’, ‘infection’, ‘infectious’, ‘disease’, ‘positive’, ‘test’, ‘symptom’, ‘november’, ‘january’, ‘contract’]
7_chemical_inhalation_vapor_respiratory	44	[‘chemical’, ‘inhalation’, ‘vapor’, ‘respiratory’, ‘exposure’, ‘ammonia’, ‘lung’, ‘fume’, ‘tank’, ‘inhale’]
8_water_hot_steam_burn	32	[‘water’, ‘hot’, ‘steam’, ‘burn’, ‘pipe’, ‘boiler’, ‘valve’, ‘release’, ‘degree’, ‘pressure’]
9_struck_roof_strike_storm	31	[‘struck’, ‘roof’, ‘strike’, ‘storm’, ‘construction’, ‘rain’, ‘instal’, ‘roofer’, ‘site’, ‘frame’]
10_methylene_chloride_chemical_bathtub	17	[‘methylene’, ‘chloride’, ‘chemical’, ‘bathtub’, ‘tub’, ‘bathroom’, ‘paint’, ‘vapor’, ‘respirator’, ‘refinish’]

* Due to lemmatization and the removal of spaces and punctuation during preprocessing, some tokens (e.g., instal) may appear unconventional. This token corresponds to install.

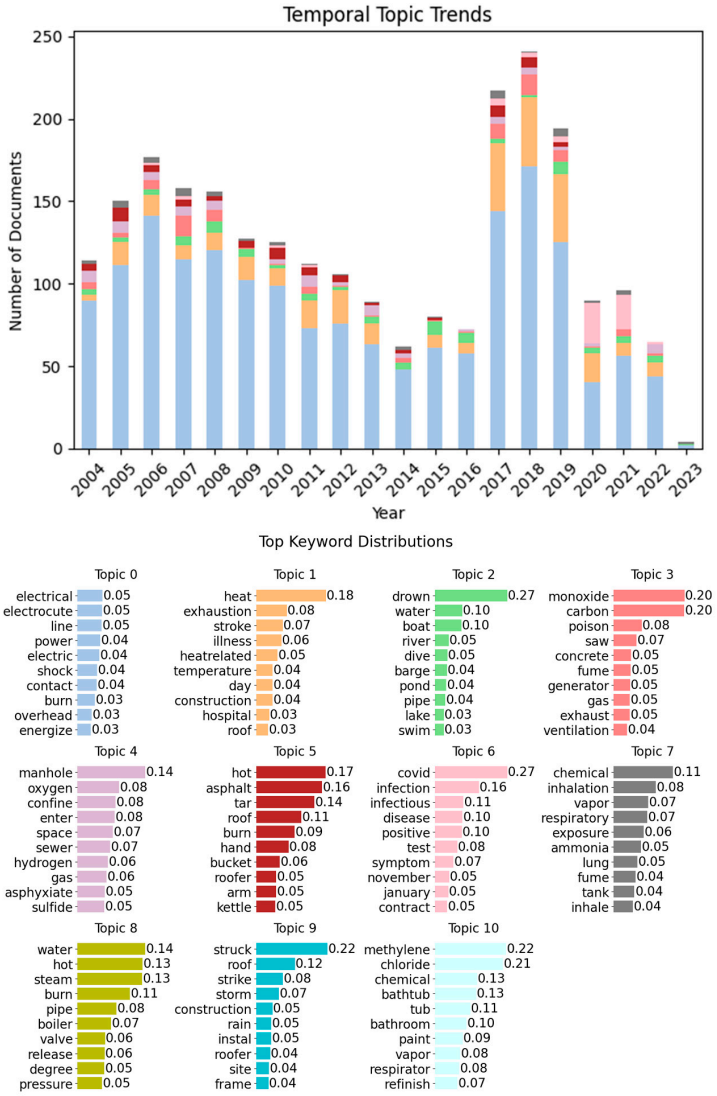

Figure 8. Results of topic analysis for the ‘Exposure to Harmful Substances or Environments’ category.

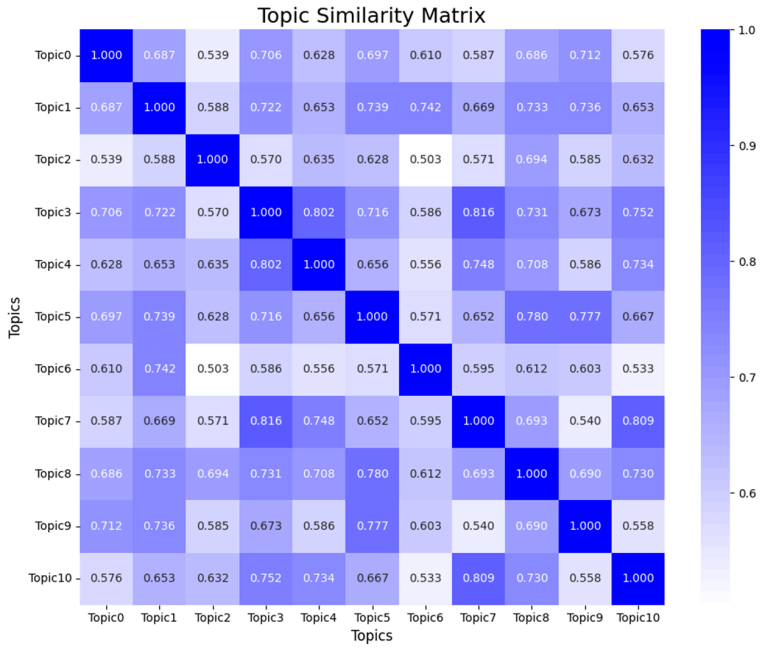

Figure 9. Topic similarity matrix of the exposure to harmful substances or environments dataset.

The first cluster includes Topics 0, 3, 7, and 10, which show relatively high similarity values, mostly above 0.70. Despite differences in context, such as electrical shock, carbon monoxide poisoning, chemical vapor inhalation, and solvent exposure during refinishing work, these topics all involve the accumulation of harmful agents in enclosed or partially enclosed spaces. This pattern suggests the need for enhanced engineering controls, including mechanical ventilation, atmospheric monitoring, and restrictions on the use of fuel-powered tools in indoor or semi-enclosed areas. A second cluster links Topics 1, 5, 8, and 9, representing heat illness, hot asphalt handling, steam-related burns, and storm-related roofing hazards. These topics are more strongly shaped by scheduling practices, material temperatures, and outdoor environmental variation. Their similarity suggests that preventive measures such as adaptive work scheduling, heat stress monitoring, and improved handling protocols for hot materials are broadly applicable across these scenarios.

Topic 4, which involves oxygen-deficient or toxic gas conditions in maintenance holes and sewer systems, occupies an intermediate position. It reflects both chemical accumulation risks and environmental stressors, underscoring that confined space safety necessitates an integrated approach that combines ventilation, atmospheric testing, and rescue readiness. Topic 6, which captures infectious disease exposures, shows low similarity to most other topics. Its distinctiveness indicates that infectious risks depend more on temporal and behavioral factors, making administrative controls and exposure tracking especially important. The similarity structure also reveals several non-intuitive patterns. Occupations such as roofers, pipefitters, and bathtub refinishers share elevated risks due to vapor, heat, or airflow constraints, despite their tasks appearing unrelated. Likewise, the similarity between solvent exposures and carbon monoxide incidents highlights shared airborne hazard pathways across different trades. These connections underscore the need for cross-trade training focused on recognizing atmospheric hazards rather than job-specific scenarios.

4.2.4. Transportation, Overexertion, Explosion, and Violence

Several key topics emerge in the category encompassing transportation, overexertion, explosion, and violence, as illustrated in Figure 10. The topic similarity matrix is shown in Figure 11. The corresponding topic model outputs are summarized in Table 8. This combined dataset provides a relatively balanced foundation for topic analysis across the original accident categories, ensuring that each type is represented in the extracted topics. However, transportation accidents account for a larger proportion of the data, which naturally leads to a higher number of topics associated with this category. The overall trend indicates an increase in accidents starting in 2004, with a peak in 2006. This was followed by a decline, reaching the lowest point in 2014. Accidents then rose steadily until 2018, after which there was a slight decrease, followed by a significant reduction in 2019, after which they stabilized. The similarity matrix reveals a dominant transportation-related cluster formed by Topics 0, 3, and 4, which all exhibit similarity values above 0.70. Although these topics describe different surface-level events, including construction traffic collisions, machinery lane maneuvers, and bulldozer run-over cases, they share underlying mechanisms related to mobile equipment movement, operator blind spots, and uncontrolled vehicle paths. This pattern indicates the need for unified safety controls across equipment types, such as standardized spotter protocols, traffic separation plans, and proximity detection systems, rather than interventions designed only for specific tasks. A second set of closely related topics includes Topics 2 and 6, which involve the ignition of flammable gases and electrical arc flash burns. Their similarity scores, between 0.70 and 0.75, suggest common energy-release mechanisms, even though they occur in different operational contexts, such as pipe cutting and electrical vault work. These relationships underscore the importance of coordinating hazard controls for ignition sources across both mechanical and electrical activities, including lockout procedures, vapor detection, and hot work permitting.

Table 8. BERTopic outputs on the dataset of transportation, overexertion, explosion, and violence.

Topic Name	Count	Representation
0_struck_vehicle_construction_traffic	1285	[‘struck’, ‘vehicle’, ‘construction’, ‘traffic’, ‘road’, ‘highway’, ‘motor’, ‘dump’, ‘run’, ‘fracture’]
1_heart_attack_body_cardiac	590	[‘heart’, ‘attack’, ‘body’, ‘cardiac’, ‘arrest’, ‘cardiovascular’, ‘cause’, ‘collapse’, ‘found’, ‘pm’]
2_burn_gas_tank_ignite	250	[‘burn’, ‘gas’, ‘tank’, ‘ignite’, ‘cut’, ‘pipe’, ‘vapor’, ‘flammable’, ‘use’, ‘line’]
3_machine_lane_operator_driver	67	[‘machine’, ‘lane’, ‘operator’, ‘driver’, ‘side’, ‘back’, ‘traffic’, ‘dump’, ‘vehicle’, ‘left’]
4_bulldozer_dozer_run_crush	33	[‘bulldozer’, ‘dozer’, ‘run’, ‘crush’, ‘struck’, ‘operator’, ‘eject’, ‘bull’, ‘seat’, ‘track’]
5_crane_collapse_scaffold_floor	20	[‘crane’, ‘collapse’, ‘scaffold’, ‘floor’, ‘boom’, ‘injured’, ‘section’, ‘mast’, ‘ft’, ‘truss’]
6_electrical_burn_volt_flash	16	[‘electrical’, ‘burn’, ‘volt’, ‘flash’, ‘arc’, ‘panel’, ‘fuse’, ‘electric’, ‘vault’, ‘fault’]
7_trench_excavation_cavein_manhole	11	[‘trench’, ‘excavation’, ‘cavein’, ‘manhole’, ‘soil’, ‘pit’, ‘bury’, ‘cave’, ‘foot’, ‘wall’]

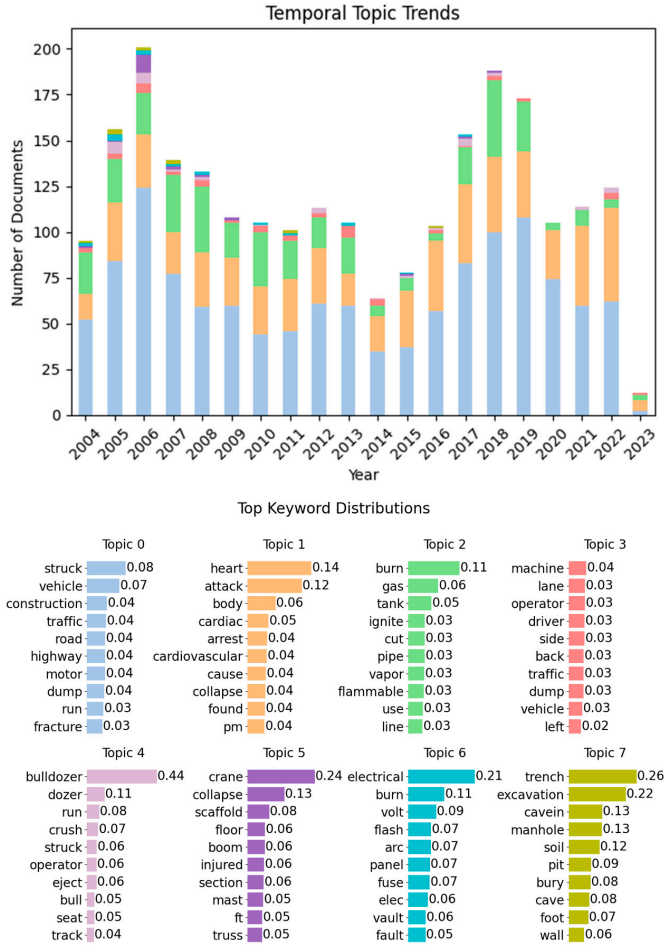

Figure 10. Results of topic analysis of the transportation, overexertion, explosion, and violence categories.

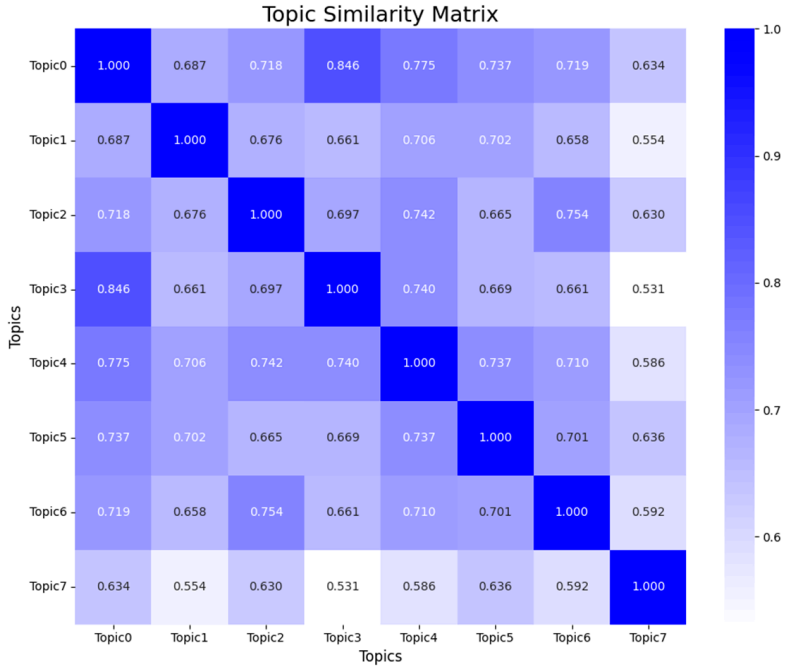

Figure 11. Topic similarity matrix of the transportation, overexertion, explosion, and violence incidents dataset.

Topic 1, which involves sudden cardiac or physiological collapse, forms a semi-independent category that shows moderate similarity to several clusters. Its connections to transportation-related topics likely stem from prolonged physical workload, heat exposure, or psychological stress experienced during the operation of equipment. This pattern suggests that medical emergencies may arise as secondary outcomes of physically demanding or high-pressure tasks, underscoring the importance of monitoring fatigue and conducting early medical screenings. Topic 7, which concerns excavation and trench cave-ins, exhibits the lowest similarity to all other topics. This indicates that soil failure hazards are structurally distinct from transportation, explosion, or physiological hazards. Its isolation reinforces the need for specialized protective systems and competent person inspections that general safety measures cannot replace. Several non-intuitive patterns emerge across the topics. For example, vehicle-related incidents affect not only equipment operators but also ground workers who are struck while performing unrelated tasks, revealing a broader exposure profile than traditionally assumed. Similarly, explosion-related incidents are often associated with pipe cutting, tank maintenance, and vault work, indicating that combustible vapor hazards exist across multiple occupations that are rarely grouped in practice.

5. Discussions

BERTopic utilizes Transformer-based embeddings to capture the contextual meaning, domain-specific terminology, and latent semantic relationships inherent in accident narratives. By representing accident descriptions in a continuous semantic vector space and clustering documents according to their intrinsic similarity, BERTopic generates topics that are more coherent, semantically consistent, and closely aligned with the underlying patterns in the data. This approach enables BERTopic to distinguish topics with greater precision across diverse types of construction safety incidents, revealing subtle semantic differences that traditional bag-of-words models often fail to detect. Furthermore, this enhanced performance can be attributed to BERTopic's ability to leverage both the semantic features of accident texts and the clustering characteristics of word vectors, which together enable the model to identify nuanced relationships between terms and group related incidents more effectively. By integrating these semantic and structural insights, BERTopic provides a clearer understanding of the underlying causes and patterns of different accident types, offering a more interpretable and practically valuable analysis.

Building on this technical advantage, the results and discussion section provides a comprehensive analysis of the identified topics and their corresponding keywords for each accident category. The analysis highlights relevant scenarios, potential causes, affected age groups, at-risk occupations, and peak accident times. Furthermore, the study proposes targeted safety recommendations aimed at improving safety standards and protecting workers. Findings are presented systematically to facilitate a clear understanding and practical application, addressing critical safety concerns within the construction industry [87,88]. By identifying potential hazards and implementing effective preventive measures, the study aims to reduce the incidence of accidents, protect workers' health and safety, and enhance their confidence and job satisfaction. These improvements, in turn, contribute to greater work efficiency and foster a safer, more sustainable working environment across the industry.

For each dataset, the BERTopic model was applied to perform thematic analysis. After generating the topics, all original documents assigned to each topic were extracted and summarized. Each document corresponds to one accident case and contains the variables Age, Occupation, and Time of Accident, which were used to compute topic-level descriptive statistics. Documents with missing or inconsistent values in these variables were removed to ensure analytical reliability. Subsequently, the Excel tool was used to calculate the age

quartiles for each topic and to identify the high-risk occupations and the peak time periods in which accidents most frequently occurred within each topic.

5.1. Falls–Slips–Trips

Table 9 systematically examines common scenarios of falls, slips, and trips by focusing on typical activities such as roof work, scaffold use, truss installation, and working near open shafts. The likely causes of these accidents include equipment instability, inadequate ground protection, poor material quality, and insufficient protective measures. The age groups are divided into quartiles (Q1, Q2, and Q3), showing a broad age range from younger workers in their early 20s (Q1) to older workers around age 50 (Q3), with risks distributed across age levels but particularly elevated among those in early or late career stages due to physical demands and exposure. Occupations such as roofers, carpenters, painters, and electricians are identified as being most vulnerable to these hazards, especially during peak accident times, typically in the morning (e.g., 8–11 am) and early afternoon (e.g., 1–2 pm). Safety recommendations provide targeted prevention strategies, including the use of personal protective equipment, stable platforms, proper ladder handling, and the secure attachment of safety harnesses to reduce the likelihood of accidents.

Table 9. Falls, Slips, Trips: Scenarios, Causes, Worker Profiles, and Safety Recommendations.

Topics	Accident Scenario	Potential Causes	Age Groups *	At-Risk Occupations	Accident Peak Time	Safety Recommendation
Roof Work and Ladder Safety	Roof-related tasks, including ladder use	Unstable ladders, insufficient ground protection	Q1: 31 Q2: 41 Q3: 52	Roofers Carpenters Painters	10–11 AM 8–9 AM 1–2 PM	Highlights the need for PPE (personal protective equipment, such as helmets and non-slip shoes) and safer ladder practices.
Scaffolding and Platform Collapses	Scaffold/ platform work	Instability, inadequate support, poor-quality materials	Q1: 33 Q2: 43 Q3: 53	Carpenters Painters Brick masons and Stonemasons	10–11 AM 9–10 AM 2–3 PM	Emphasizes the importance of sturdy platforms, quality materials, and comprehensive safety management.
Truss Installation Accidents	Accidents during truss installation	Design flaws, improper installation, and insufficient material strength	Q1: 31 Q2: 38 Q3: 47	Carpenters	11 AM–12 PM 9–10 AM 3–4 PM	Stresses the need for strict safety measures during roof and structural installations.
Skylight and Opening Hazards	Falls through unguarded skylights and openings	Lack of protective measures	Q1: 28 Q2: 37 Q3: 49	Roofers Painters	10–11 AM 9–10 AM 11 AM–12 PM	Urges the installation of guardrails or covers in hazardous areas.
Electrical Shock Accidents **	Electric shocks from operating equipment or lines	Live equipment, poor insulation, unsafe contact with wires	Q1: 30 Q2: 41 Q3: 50	Electricians Electricians' apprentices	11–12 PM 10–11 AM 3–4 PM	Emphasizes proper maintenance and management of electrical systems.
Elevator Shaft and Door Safety	Accidents in elevator shafts and door areas	Open shafts, faulty door mechanisms, unstable platforms	Q1: 28 Q2: 40 Q3: 50	Elevator installers and repairers Drywall installers	10–11 AM 9–10 AM 11 AM–12 PM	Stresses the need for strict safety measures during construction and maintenance.
Communication Tower Maintenance Risks	Maintenance or installation of towers, antennas	Installation issues, risks when using ladders	Q1: 25 Q2: 34 Q3: 41	Electricians Electrical power installers and repairers	1–2 PM 10–11 AM 11 AM–12 PM	Highlights the importance of electrical safety when handling cables and equipment.
Safety Harness and Equipment Issues	Accidents involving harnesses during roof work	Improper use of harnesses or lifelines	Q1: 24 Q2: 31 Q3: 38	Roofers Carpenters Structural metal workers	11 AM–12 PM 8–9 AM 12–1 PM	Emphasizes proper usage and attachment of safety harnesses.

* Q1: first quartile; Q2: median; Q3: third quartile; ** The OIICS manual includes provisions stating that falls due to electric shock are classified as falls–slips–trips.

5.2. Contact with Objects or Equipment

Table 10 systematically examines scenarios of injuries resulting from contact with objects or equipment, highlighting accident-prone tasks like working with saws, trench excavation, and handling heavy machinery. The likely causes of these accidents include poor site management, inadequate equipment safeguards, inadequate trench support, and insufficient use of safety features. The age distribution spans three quartiles (Q1, Q2, and Q3),

with younger workers in Q1 (ages 25–33) and older workers in Q3 (up to age 52) showing distinct vulnerability to object and equipment-related hazards, as both groups may encounter unique challenges in handling heavy machinery and sharp tools safely. Occupations such as carpenters, truck drivers, machine operators, and plumbers are identified as particularly susceptible to these injuries, with peak accident times occurring predominantly in the morning (e.g., 9–10 am) and late morning to early afternoon (e.g., 11 am–12 pm). Recommended safety measures focus on load securing, the proper use of guards and barriers, worker training on equipment handling, and the use of protective gear, such as helmets, gloves, and goggles.

Table 10. Contact with Objects or Equipment: Scenarios, Causes, Worker Profiles, and Safety Measures.

Topics	Accident Scenario	Potential Causes	Age Groups *	At-Risk Occupations	Accident Peak Time	Safety Recommendation
Struck-by and Crush Injuries	Struck by falling objects or crushed by machinery	Poor site management, lack of barriers, and unsecured materials	Q1: 30 Q2: 40 Q3: 52	Carpenters Truck drivers, heavy	9–10 AM 11 AM–12 PM 8–9 AM	Proper load securing, crane operation, and use of protective barriers
Hand Injuries from Saws	Injuries from saws, especially circular saws	Lack of guards, improper handling, kickback	Q1: 28 Q2: 36 Q3: 49	Carpenters Cabinet makers and bench carpenters	9–10 AM 11 AM–12 PM 10–11 AM	Use protective guards, focus during operation, and train workers on saw use
Trench Collapse Accidents	Trench or excavation collapse	Lack of trench support, unstable soil, heavy machinery near edges	Q1: 28 Q2: 37 Q3: 48	Excavating and loading machine operators Plumbers, pipefitters, and steamfitters	11 AM–12 PM 9–10 AM 1–2 PM	Ensure trench shoring, regular inspections, and avoid heavy equipment near trenches
Nail Gun Injuries	Injuries from nail gun misfires or accidental discharges	Misuse of tools, lack of safety locks	Q1: 25 Q2: 34 Q3: 47	Carpenters	11 AM–12 PM 2–3 PM 10–11 AM	Use nail gun safety features, wear protective gear, and train workers
Injuries from Falling Tree Limbs	Struck by falling tree limbs or branches during trimming	Inadequate tree stability assessment, improper chainsaw use	Q1: 32 Q2: 40 Q3: 52	Timber cutting and logging operators Excavating and loading machine operators	9–10 AM 11 AM–12 PM 12–1 PM	Proper tree-cutting techniques, use helmets, and maintain a safe perimeter
Saw Kickback Injuries	Injuries from a saw kickback	Improper handling, dull blades, and unsecured materials	Q1: 30 Q2: 40 Q3: 52	Pipelayer Plumbers, pipefitters, and steamfitters	11 AM–12 PM 8–9 AM 9–10 AM	Maintain saws, secure materials, and be aware of kickback risks
Eye Injuries from Sharp Objects	Eye injuries from nails or sharp objects	Lack of eye protection, improper tool handling	Q1: 30 Q2: 40 Q3: 45	Carpenters	10–11 AM 12–1 PM 3–4 PM	Wear safety goggles and ensure proper tool handling
Grinder Injuries	Injuries from grinders (rotating wheel)	Improper use, damaged wheels, and unsecured workpiece	Q1: 33 Q2: 41 Q3: 48	Tile setters Carpenters	11 AM–12 PM 10–11 AM 9–10 AM	Ensure proper training, regular equipment inspections, and use gloves/face shields

* Q1: first quartile; Q2: median; Q3: third quartile.

5.3. Exposure to Harmful Substances or Environments

Table 11 systematically examines scenarios involving exposure to harmful substances or environments, detailing high-risk tasks such as working near live electrical lines, handling hot substances, and entering confined or poorly ventilated spaces. Common causes of these accidents include inadequate ventilation, lack of protective barriers, improper handling of equipment, and failure to monitor environmental hazards such as oxygen levels or chemical fumes. The age groups most impacted are categorized into three quartiles (Q1, Q2, and Q3), with younger workers (Q1, aged 22–29) and older workers (Q3, up to age 58) facing specific risks associated with tasks involving high temperatures, confined spaces, and exposure to toxic chemicals. Occupations such as electricians, plumbers, roofers, and heating, ventilation, and air conditioning (HVAC) workers are particularly vulnerable to these hazards, with peak accident times occurring most often in the late morning (e.g., 10–11 am) and early afternoon (e.g., 2–3 pm), though certain risks also peak later in the day. Safety recommendations emphasize preventive measures, including ensuring proper ventilation, wearing personal protective equipment, undergoing regular

health screenings, completing hazard-specific training, and adhering to strict protocols for handling hazardous substances and working in high-risk environments.

Table 11. Exposure to Harmful Substances or Environments: Scenarios, Causes, Worker Profiles, and Safety Measures.

Topics	Accident Scenario	Potential Causes	Age Groups *	At-Risk Occupations	Accident Peak Time	Safety Recommendation
Electrical Exposure and Electrocution	Contact with live electrical lines, improperly insulated equipment	Lack of de-energizing circuits, faulty insulation, and inadequate grounding	Q1: 29 Q2: 36 Q3: 47	Electricians Electrical power installers and repairers Electricians' apprentices	10–11 AM 9–10 AM 1–2 PM	Follow electrical safety protocols, training, use insulated tools, and PPE
Heat-Related Illnesses	Exposure to high temperatures, risk of heat stroke	Prolonged heat exposure, inadequate hydration	Q1: 29 Q2: 38 Q3: 49	Roofers Carpenters Electricians	11 AM–12 PM 12–1 PM 1–2 PM	Regular breaks, hydration, proper clothing, and monitoring for heat stress
Drowning and Water-Related Accidents	Activities near water bodies, risk of drowning	Lack of barriers, inadequate life vests	Q1: 28 Q2: 40 Q3: 49	Diver Excavating and loading machine operators Plumbers, pipefitters, and steamfitters	10–11 AM 9–10 AM 5–6 PM	Water safety training, life vests, secure work areas near water
Carbon Monoxide Poisoning	Exposure to CO in poorly ventilated areas	Poor ventilation, improper equipment use	Q1: 28 Q2: 35 Q3: 45	Plumber, pipefitter, and steamfitter apprentices HVAC and refrigeration workers Drywall installers	10–11 AM 8–9 AM 2–3 PM	Proper ventilation, CO detectors, and recognizing CO poisoning signs
Oxygen Deficiency in Confined Spaces	Entry into oxygen-deficient spaces	Failure to monitor oxygen levels, lack of ventilation	Q1: 29 Q2: 37 Q3: 51	Plumbers, pipefitters, and steamfitters Pipelayer	9–10 AM 10–11 AM 11 AM–12 PM	Air quality monitoring, ventilation, use breathing apparatus, follow protocols
Exposure to Hot Substances	Handling hot asphalt or tar, risk of burns	Lack of protective gear, accidental spills	Q1: 28 Q2: 43 Q3: 50	Roofers Roofers: helpers	3–4 PM 11 AM–12 PM 12–1 PM	Heat-resistant gloves, protective clothing, and careful handling procedures
Infectious Diseases	Risk of infectious disease spread on-site	Inadequate health screening, lack of PPE	Q1: 43 Q2: 51 Q3: 58	Construction laborers	8–9 AM 12–1 PM 9–10 AM	Health screenings, masks, social distancing, and hygiene practices
Chemical Inhalation and Respiratory Hazards	Inhalation of toxic chemicals or vapors	Poor ventilation, inadequate respiratory protection	Q1: 28 Q2: 37 Q3: 44	Plumbers, pipefitters, and steamfitters Painters Carpenters	10–11 AM 2–3 PM 8–9 AM	Proper ventilation, use of respirators, and chemical hazard training
Burns from Hot Water and Steam	Exposure to hot water or steam	Equipment malfunctions, accidental contact	Q1: 27 Q2: 35 Q3: 41	HVAC and refrigeration workers Mechanical engineering technicians Plumber, pipefitter, and steamfitter apprentices	2–3 PM 10–11 AM 9–10 AM	Regular maintenance, protective gear, and awareness of hot surfaces
Roof-Related Strikes During Storms	construction sites during roof installation or framing, with rain and strong winds	unsecured materials or structural instability during storms	Q1: 27 Q2: 33 Q3: 39	Roofers Carpenter Roof repair: sheet metal	2–3 PM 3–4 PM 4–5 PM	secure equipment, enforce safety protocols during storms, and provide workers with protective gear
Chemical Exposure (Methylene Chloride)	Exposure to hazardous chemicals like methylene chloride	Inadequate ventilation, prolonged exposure to fumes	Q1: 22 Q2: 30 Q3: 37	Painters Glaziers paperhangers and plasterers	4–5 PM 3–4 PM 2–3 PM	Ventilation, respirators, and adherence to safety guidelines

* Q1: first quartile; Q2: median; Q3: third quartile.

5.4. Transportation, Overexertion, Explosion, and Violence

Table 12 systematically examines scenarios of injuries related to transportation, overexertion, explosions, and violence, identifying high-risk tasks such as vehicle operation, handling heavy machinery, and working with flammable substances. Key causes include poor traffic management, inadequate equipment maintenance, excessive physical exertion, improper storage of flammable materials, and interpersonal conflicts on-site. The age groups most affected are presented in three quartiles (Q1, Q2, and Q3), with younger workers, typically around age 25, in Q1 for structural failures, and older workers, up to 58 in Q3, for electrical injuries, indicating distinct vulnerabilities related to physical strain, equipment operation, and situational hazards. Occupations identified as particularly at risk include construction laborers, truck drivers, equipment operators, and electricians, with peak accident times generally occurring in the early morning (e.g., 8:00–10:00 am) and midday (e.g., 10:00 am–12:00 pm). However, certain accidents, such as structural failures,

also peak later in the day. Safety recommendations emphasize preventive measures, including implementing workplace violence protocols, effective traffic management, adequate training, regular equipment maintenance, clear safety signage, and adherence to emergency preparedness procedures for handling explosives and hazardous materials.

Table 12. Transportation, overexertion, explosion, and violence: Scenarios, causes, worker profiles, and safety measures.

Topics	Accident Scenario	Potential Causes	Age Groups *	At-Risk Occupations	Accident Peak Time	Safety Recommendation
Violent Acts	On-site violence, gunshots, explosions	Interpersonal conflicts, accidental firearm discharge, and mishandling explosives	Q1: 30 Q2: 37 Q3: 52	Construction inspectors Construction laborers	9–10 AM 8–9 AM 1–2 PM	Implement workplace violence safety measures and proper handling of explosives
Transportation Accidents	Accidents involving vehicles on-site or on nearby roads	Poor traffic management, lack of safety barriers, vehicle malfunction, and driver error	Q1: 32 Q2: 45 Q3: 56	Truck drivers Paving, surfacing, and tamping equipment operators Highway maintenance worker	10–11 AM 9–10 AM 3–4 PM	Effective traffic management, safety barriers, and clear signage
Overexertion	Medical conditions (heart attacks, cardiac arrest) due to overexertion	Excessive physical effort, inadequate rest, high stress, and pre-existing conditions	Q1: 43 Q2: 51 Q3: 57	Construction laborers Carpenters Electricians	1–2 PM 2–3 PM 11 AM–12 PM	Work-rest schedules, monitor worker health, stress symptom training
Explosions and Fires	Explosions or fires from flammable gases, chemicals, or equipment	Improper storage, equipment malfunction, and unsafe practices	Q1: 26 Q2: 36 Q3: 48	Welders and cutters Plumbers, pipefitters, and steamfitters Painters	11 AM–12 PM 10–11 AM 12–1 PM	Safety protocols, regular maintenance, and emergency preparedness
Equipment Malfunction	Accidents due to equipment malfunctions or mishandling	Poor maintenance, operator error, inadequate training	Q1: 36 Q2: 43 Q3: 52	Equipment operator: heavyCarpenters Construction inspectors	10–11 AM 8–9 AM 9–10 AM	Regular maintenance, operator training, and safety guidelines
Machinery Accidents	Heavy machinery accidents (bulldozers) leading to crushing or run-over injuries	Malfunctioning equipment, operator error, and inadequate safety features	Q1: 36 Q2: 47 Q3: 55	Excavating and loading machine operators Equipment operator: heavyGrader, dozer, and scraper operators	1–2 PM 8–9 AM 10–11 AM	Proper training, equipment maintenance, and use of safety restraints
Structural Failures	Crane or scaffold collapses	Structural failures, improper assembly, overloading, and poor design	Q1: 25 Q2: 30 Q3: 33	Carpenters Construction laborers	6–7 PM 12–1 PM 4–5 PM	Proper design, assembly, and maintenance
Electrical Injuries	Electric shocks, burns from high-voltage equipment	Faulty wiring, improper equipment handling, live wire exposure	Q1: 43 Q2: 57 Q3: 58	Electricians Plumber, pipefitter, and steamfitter apprentices	12–1 AM 10–11 AM 9–10 AM	Electrical safety measures, regular inspections, and equipment training
Excavation and Trenching Accidents	Trench collapses, falls into maintenance holes or pits	Unstable soil, lack of proper shoring, and insufficient safety measures	Q1: 27 Q2: 27 Q3: 27	Excavating and loading machine operators Construction laborers	10–11 AM 11 AM–12 PM 6–7 PM	Proper excavation techniques, safety shoring, and regular inspections

* Q1: first quartile; Q2: median; Q3: third quartile.

6. Conclusions

Analyzing workplace accident reports provides critical insights into past accidents, which can help prevent future ones. This study employs BERTopic, an effective topic modeling approach, to reduce the workload and enhance the efficiency of extracting valuable information from construction accident reports. Specifically, an open-source, off-the-shelf topic model is utilized to extract relevant details from reports within the construction industry, demonstrating its applicability and effectiveness. The research analyzes accident reports published by OSHA, encompassing 22,623 accident descriptions from 2004 to 2023. This study offers several meaningful contributions to the analysis of construction accidents. First, we show that BERTopic performs more effectively than the traditional LDA baseline, yielding higher topic coherence and topic diversity across all datasets. This enhanced performance highlights its ability to generate more interpretable, distinctive, and contextually meaningful topics from accident narratives. Second, by using contextual embeddings, the study identifies several previously overlooked risk scenarios and relationships between occupations and accident situations that earlier LDA-based or frequency-driven approaches did not capture. These insights provide a richer understanding of the mechanisms be-

hind different accident types. Third, the analysis uncovers temporal patterns, including shifts in accident prevalence, peak periods, and groups of workers who face higher risk. These findings can support more targeted safety planning and more effective allocation of resources. Finally, the study provides a comprehensive topic-level summary for each accident category, outlining key scenarios, possible causes, vulnerable age groups, high-risk occupations, and time-related patterns. Together, these results demonstrate the value of contextual topic modeling for extracting structured knowledge from accident records.

From a theoretical perspective, the study contributes to refining the understanding of accident patterns in the construction industry. By revealing connections and distinctions among different types of accidents, it supports the development of a more integrated framework for interpreting construction safety risks. The examination of key terms and temporal dynamics also informs more nuanced views of accident causation, helping to strengthen the conceptual foundations of safety science.

From a practical viewpoint, the findings offer insights that may help improve safety management in construction settings. Identifying critical periods and high-risk scenarios enables organizations and regulators to adopt more targeted preventive measures. Detailed descriptions of accident scenarios, contributing factors, and affected body parts can guide the formulation of safety protocols tailored to specific hazards. These results may also inform the revision of safety standards and enforcement strategies. In addition, the insights gained from topic patterns and temporal trends can support the design of more effective training programs, helping workers better recognize and avoid potential hazards and ultimately improving workplace safety.

Despite the study's strengths, certain limitations need to be acknowledged. A key challenge was the time-intensive process of selecting the optimal parameter combination for BERTopic. Due to the large number of hyperparameters involved, extensive experimentation and fine-tuning were necessary, which prolonged the process. The complexity of balancing factors such as dimensionality reduction, clustering precision, and topic coherence further contributed to this challenge. Additionally, space constraints limited the depth of analysis for each accident type, preventing a more detailed exploration of specific scenarios and contributing factors. While merging categories helps address the issue of insufficient sample size, the resulting topics should be interpreted cautiously, especially when considering practical interventions or broader generalization. Because conceptually distinct accident mechanisms were combined, the interpretability and generalizability of the model outcomes remain limited. In addition to the operational limitations previously discussed, several conceptual and data-related constraints should be acknowledged. First, our analysis relies solely on U.S. OSHA accident reports, which may introduce data-source bias due to regional coverage limitations, potential under-reporting, and variability in reported accident severity. Second, the narrative quality and length of accident descriptions vary considerably, which can affect the topic modeling process and the interpretability of the extracted themes. Third, topic stability may be influenced by the choice of random seeds and hyperparameter settings, potentially resulting in different topic distributions under alternative configurations. Finally, the model has not been externally validated on independent datasets, which limits the generalizability of the findings.

Future research could address these issues by incorporating additional datasets from other regions, standardizing and enriching narrative data, systematically evaluating model robustness under different parameter settings, and performing external validation to enhance the reliability and applicability of the results. Future research should also focus on optimizing the hyperparameter selection process to streamline the tuning of BERTopic, potentially through the use of automated optimization techniques or advanced computational resources. Addressing the space constraints in future studies could enable more

comprehensive and detailed analyses of each accident type, involving deeper investigations into specific accident scenarios, contributing factors, and mitigation strategies. These improvements would enhance both the depth and efficiency of the analysis, yielding more robust and actionable insights into accidents in the construction industry.

Author Contributions: Conceptualization, Y.C. (Yuntao Cao) and Z.Q.; methodology, Y.C. (Yuntao Cao), Z.Q., S.W., Y.C. (Yuting Chen), M.S., X.M. and J.W.; validation, Y.C. (Yuntao Cao), Z.Q., S.W., Y.C. (Yuting Chen), M.S., X.M. and J.W.; formal analysis, Y.C. (Yuntao Cao), Z.Q. and S.W.; data curation, Y.C. (Yuntao Cao) and M.S.; writing—original draft preparation, Y.C. (Yuntao Cao) and Z.Q.; writing—review and editing, S.W., Y.C. (Yuting Chen), M.S., X.M. and J.W.; visualization, Y.C. (Yuntao Cao), Z.Q., S.W. and Y.C. (Yuting Chen); supervision, M.S., X.M. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shandong Province Natural Science Foundation, grant number ZR2023QG168.

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

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Hossain, M.A.; Zhumabekova, A.; Paul, S.C.; Kim, J.R. A review of 3D printing in construction and its impact on the labor market. *Sustainability* **2020**, *12*, 8492. [CrossRef]
2. Alaloul, W.S.; Musarat, M.A.; Rabbani, M.B.A.; Iqbal, Q.; Maqsoom, A.; Farooq, W. Construction sector contribution to economic stability: Malaysian GDP distribution. *Sustainability* **2021**, *13*, 5012. [CrossRef]
3. Bertelsen, S. Construction as a Complex System. In Proceedings of the 11th Annual Conference of the International Group for Lean Construction, Blacksburg, VA, USA, 22–24 July 2003; pp. 11–23.
4. Lee, S.H.; Peña-Mora, F.; Park, M. Dynamic planning and control methodology for strategic and operational construction project management. *Autom. Constr.* **2006**, *15*, 84–97. [CrossRef]
5. Wang, J.; Zhang, S.; Teizer, J. Geotechnical and safety protective equipment planning using range point cloud data and rule checking in building information modeling. *Autom. Constr.* **2015**, *49*, 250–261. [CrossRef]
6. Choi, S.D.; Guo, L.; Kim, J.; Xiong, S. Comparison of fatal occupational injuries in construction industry in the United States, South Korea, and China. *Int. J. Ind. Ergon.* **2019**, *71*, 64–74. [CrossRef]
7. Cao, Y.; Wu, S.; Chen, Y.; Skitmore, M.; Ma, X.; Wang, J. Enhancing Knowledge of Construction Safety: A Semantic Network Analysis Approach. *Buildings* **2025**, *15*, 3036. [CrossRef]
8. International Labour Organization. 2021. Available online: https://www.ilo.org/ (accessed on 13 July 2025).
9. Fang, D.; Wu, H. Development of a Safety Culture Interaction (SCI) model for construction projects. *Saf. Sci.* **2013**, *57*, 138–149. [CrossRef]
10. Ibrion, M.; Paltrinieri, N.; Nejad, A.R. Learning from failures: Accidents of marine structures on Norwegian continental shelf over 40 years time period. *Eng. Fail. Anal.* **2020**, *111*, 104487. [CrossRef]
11. Zhang, M.; Ghodrati, N.; Poshdar, M.; Seet, B.C.; Yongchareon, S. A construction accident prevention system based on the Internet of Things (IoT). *Saf. Sci.* **2023**, *159*, 106012. [CrossRef]
12. Kim, T.; Lee, M.; Shin, Y.; Yoo, W.S. Deep neural network-based probabilistic classifier of occupational accident types on a construction site in Korea. *J. Asian Archit. Build. Eng.* **2025**, *24*, 2763–2772. [CrossRef]
13. Abukhashabah, E.; Summan, A.; Balkhyour, M. Occupational accidents and injuries in construction industry in Jeddah city. *Saudi J. Biol. Sci.* **2020**, *27*, 1993–1998. [CrossRef] [PubMed]
14. Chan, A.P.; Guan, J.; Choi, T.N.; Yang, Y.; Wu, G.; Lam, E. Improving safety performance of construction workers through learning from accidents. *Int. J. Environ. Res. Public Health* **2023**, *20*, 4570. [CrossRef]
15. Occupational Safety and Health Administration; U.S. Department of Labor. Focus Four: Common Hazards in Construction. Available online: https://www.osha.gov/training/outreach/construction/focus-four (accessed on 8 October 2024).
16. Choi, J.; Gu, B.; Chin, S.; Lee, J.S. Machine learning predictive model based on national data for fatal accidents of construction workers. *Autom. Constr.* **2020**, *110*, 102974. [CrossRef]
17. U.S. Bureau of Labor Statistics. Fatal Occupational Injuries by Industry Sectors and Selected Events or Exposures (1), 2023. 2024. Available online: https://www.bls.gov/news.release/cfoi.t02.htm (accessed on 25 April 2025).

18. Manu, P.; Emuze, F.; Saurin, T.A.; Hadikusumo, B. (Eds.) *Construction Health and Safety in Developing Countries*; Routledge: London, UK, 2019.
19. Gadekar, H.; Bugalia, N. Automatic classification of construction safety reports using semi-supervised YAKE-Guided LDA approach. *Adv. Eng. Inform.* **2023**, *56*, 101929. [CrossRef]
20. Zhang, Q.; Chan, A.P.; Yang, Y.; Guan, J.; Choi, T.N. Influence of learning from accidents, safety information flow, and resilient safety culture on construction safety performance. *J. Manag. Eng.* **2023**, *39*, 04023007.
21. Guan, J.; Zixuan, Y.; Chan, A.P.; Choi, T.; Yang, Y. Factors affecting learning from accidents: A cross-industry review. *J. Loss Prev. Process Ind.* **2024**, *89*, 105297. [CrossRef]
22. Stanton, N.A.; Margaryan, A.; Littlejohn, A. Learning from accidents. *Saf. Sci.* **2017**, *99*, 1–4. [CrossRef]
23. Drupsteen, L.; Groeneweg, J.; Zwetsloot, G.I. Critical steps in learning from accidents: Using learning potential in the process from reporting an incident to accident prevention. *Int. J. Occup. Saf. Ergon.* **2013**, *19*, 63–77. [CrossRef]
24. Tirunagari, S.; Hanninen, M.; Stanhlberg, K.; Kujala, P. Mining causal relations and concepts in maritime accidents investigation reports. *Int. J. Innov. Res. Dev.* **2012**, *1*, 548–566.
25. Li, S.; You, M.; Li, D.; Liu, J. Identifying coal mine safety production risk factors by employing text mining and Bayesian network techniques. *Process Saf. Environ. Prot.* **2022**, *162*, 1067–1081. [CrossRef]
26. Cao, D.; Cheng, L. Interaction effect of building construction accident attributes based on complex network. *Process Saf. Prog.* **2024**, *43*, S293–S303. [CrossRef]
27. Zhang, W.; Zhu, S.; Zhang, X.; Zhao, T. Identification of critical causes of construction accidents in China using a model based on system thinking and case analysis. *Saf. Sci.* **2020**, *121*, 606–618. [CrossRef]
28. Mohandes, S.R.; Sadeghi, H.; Fazeli, A.; Mahdiyar, A.; Hosseini, M.R.; Arashpour, M.; Zayed, T. Causal analysis of accidents on construction sites: A hybrid fuzzy Delphi and DEMATEL approach. *Saf. Sci.* **2022**, *151*, 105730. [CrossRef]
29. Rafindadi, A.D.U.; Napiah, M.; Othman, I.; Mikić, M.; Haruna, A.; Alarifi, H.; Al-Ashmori, Y.Y. Analysis of the causes and preventive measures of fatal fall-related accidents in the construction industry. *Ain Shams Eng. J.* **2022**, *13*, 101712. [CrossRef]
30. Hong, W.T.; Clifton, G.; Nelson, J.D. A data-driven conceptual framework for understanding the nature of hazards in railway accidents. *Transp. Policy* **2024**, *152*, 102–117. [CrossRef]
31. Van Dinter, R.; Tekinerdogan, B.; Catal, C. Automation of systematic literature reviews: A systematic literature review. *Inf. Softw. Technol.* **2021**, *136*, 106589. [CrossRef]
32. Elekes, Á.; Englhardt, A.; Schäler, M.; Böhm, K. Toward meaningful notions of similarity in NLP embedding models. *Int. J. Digit. Libr.* **2020**, *21*, 109–128. [CrossRef]
33. Chowdhary, K.R. Natural Language Processing. In *Fundamentals of Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 603–649.
34. Srinivasan, P.; Nagarajan, V.; Mahadevan, S. Mining and Classifying Aviation Accident Reports. In *AIAA Aviation 2019 Forum*; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2019; p. 2938.
35. Shrestha, S.; Morshed, S.A.; Pradhananga, N.; Lv, X. Leveraging Accident Investigation Reports as Leading Indicators of Construction Safety Using Text Classification. In Proceedings of the Construction Research Congress 2020, Tempe, AZ, USA, 8–10 March 2020; American Society of Civil Engineers: Reston, VA, USA, 2020; pp. 490–498.
36. Zhang, J.; Zi, L.; Hou, Y.; Deng, D.; Jiang, W.; Wang, M. A C-BiLSTM approach to classify construction accident reports. *Appl. Sci.* **2020**, *10*, 5754. [CrossRef]
37. Barde, B.V.; Bainwad, A.M. An Overview of Topic Modeling Methods and Tools. In Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 15–16 June 2017; IEEE: New York, NY, USA, 2017; pp. 745–750.
38. Hu, Y.; Boyd-Graber, J.; Satinoff, B.; Smith, A. Interactive topic modeling. *Mach. Learn.* **2014**, *95*, 423–469. [CrossRef]
39. Qiang, J.; Qian, Z.; Li, Y.; Yuan, Y.; Wu, X. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 1427–1445. [CrossRef]
40. Tenney, I. BERT Rediscovers the Classical NLP Pipeline. *arXiv* **2019**, arXiv:1905.05950. [CrossRef]
41. Sohrabi, M.S.; Mahdavi, N.; Kalatpour, O. Suggestion a Template for Accident Reporting Form. *Int. J. Occup. Hyg.* **2017**, *9*, 215–222.
42. Stemn, E.; Bofinger, C.; Cliff, D.; Hassall, M.E. Failure to learn from safety accidents: Status, challenges and opportunities. *Saf. Sci.* **2018**, *101*, 313–325. [CrossRef]
43. Dallat, C.; Salmon, P.M.; Goode, N. Risky systems versus risky people: To what extent do risk assessment methods consider the systems approach to accident causation? A review of the literature. *Saf. Sci.* **2019**, *119*, 266–279. [CrossRef]
44. Pan, X.; Zhong, B.; Wang, Y.; Shen, L. Identification of accident-injury type and bodypart factors from construction accident reports: A graph-based deep learning framework. *Adv. Eng. Inform.* **2022**, *54*, 101752. [CrossRef]
45. Chi, C.F.; Yang, C.C.; Chen, Z.L. In-depth accident analysis of electrical fatalities in the construction industry. *Int. J. Ind. Ergon.* **2009**, *39*, 635–644. [CrossRef]

46. Chi, S.; Han, S. Analyses of systems theory for construction accident prevention with specific reference to OSHA accident reports. *Int. J. Proj. Manag.* **2013**, *31*, 1027–1041. [CrossRef]
47. Raviv, G.; Fishbain, B.; Shapira, A. Analyzing risk factors in crane-related near-miss and accident reports. *Saf. Sci.* **2017**, *91*, 192–205. [CrossRef]
48. Na, X.U.; Ling, M.A.; Liu, Q.; Li, W.A.N.G.; Deng, Y. An improved text mining approach to extract safety risk factors from construction accident reports. *Saf. Sci.* **2021**, *138*, 105216. [CrossRef]
49. Baumer, E.P.; Mimno, D.; Guha, S.; Quan, E.; Gay, G.K. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 1397–1410. [CrossRef]
50. Kuhn, K.D. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transp. Res. Part C Emerg. Technol.* **2018**, *87*, 105–122. [CrossRef]
51. Abdelrazek, A.; Eid, Y.; Gawish, E.; Medhat, W.; Hassan, A. Topic modeling algorithms and applications: A survey. *Inf. Syst.* **2023**, *112*, 102131. [CrossRef]
52. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **2016**, *5*, 1608. [CrossRef] [PubMed]
53. Liu, Y.; Wang, J.; Tang, S.; Zhang, J.; Wan, J. Integrating information entropy and latent Dirichlet allocation models for analysis of safety accidents in the construction industry. *Buildings* **2023**, *13*, 1831. [CrossRef]
54. Ertek, G.; Kailas, L. Analyzing a decade of wind turbine accident news with topic modeling. *Sustainability* **2021**, *13*, 12757. [CrossRef]
55. Min, K.B.; Song, S.H.; Min, J.Y. Topic modeling of social networking service data on occupational accidents in Korea: Latent Dirichlet allocation analysis. *J. Med. Internet Res.* **2020**, *22*, e19222. [CrossRef]
56. Suh, Y. Sectoral patterns of accident process for occupational safety using narrative texts of OSHA database. *Saf. Sci.* **2021**, *142*, 105363. [CrossRef]
57. Seo, J.U.; Song, C.H. Big Data Analytics of Construction Safety Accidents Using Text Mining. *J. Korean Soc. Ind. Converg.* **2024**, *27*, 581–590.
58. Blei, D.; Lafferty, J. Correlated topic models. *Adv. Neural Inf. Process. Syst.* **2006**, *18*, 147.
59. Cao, J.; Xia, T.; Li, J.; Zhang, Y.; Tang, S. A density-based method for adaptive LDA model selection. *Neurocomputing* **2009**, *72*, 1775–1781. [CrossRef]
60. Bystrov, V.; Naboka-Krell, V.; Staszewska-Bystrova, A.; Winker, P. Choosing the Number of Topics in LDA Models–A Monte Carlo Comparison of Selection Criteria. *J. Mach. Learn. Res.* **2024**, *25*, 1–30.
61. Abuzayed, A.; Al-Khalifa, H. BERT for Arabic topic modeling: An experimental study on BERTopic technique. *Procedia Comput. Sci.* **2021**, *189*, 191–194. [CrossRef]
62. Egger, R.; Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef] [PubMed]
63. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
64. Wang, Z.; Chen, J.; Chen, J.; Chen, H. Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics* **2023**, *129*, 7359–7384. [CrossRef]
65. Brennan, R.L.; Prediger, D.J. Coefficient kappa: Some uses, misuses, and alternatives. *Educ. Psychol. Meas.* **1981**, *41*, 687–699. [CrossRef]
66. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
67. Ng, A.; Jordan, M. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. In Proceedings of the Advances in Neural Information Processing Systems 14 (NIPS 2001), Vancouver, BC, Canada, 3–8 December 2001; Volume 14.
68. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
69. Greene, D.; O'Callaghan, D.; Cunningham, P. How Many Topics? Stability Analysis for Topic Models. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, 15–19 September 2014; Proceedings, Part I 14. Springer: Berlin/Heidelberg, Germany, 2014; pp. 498–513.
70. Alnusyan, R.; Almotairi, R.; Almufadhi, S.; Shargabi, A.A.; Alshobaili, J. A Semi-Supervised Approach for User Reviews Topic Modeling and Classification. In Proceedings of the 2020 International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia, 9–10 September 2020; IEEE: New York, NY, USA, 2020; pp. 1–5.
71. Matira, Y.; Setiawan, I. Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation. *Estimasi J. Stat. Its Appl.* **2023**, *4*, 53–63.
72. Zhou, K.; Wang, J.; Ashuri, B.; Chen, J. Discovering the Research Topics on Construction Safety and Health Using Semi-Supervised Topic Modeling. *Buildings* **2023**, *13*, 1169. [CrossRef]
73. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

74. Samsir, S.; Saragih, R.S.; Subagio, S.; Aditiya, R.; Watrianthos, R. BERTopic modeling of natural language processing abstracts: Thematic structure and trajectory. *J. Media Inform. Budidarma* **2023**, *7*, 1514–1520. [CrossRef]
75. Medvecki, D.; Bašaragin, B.; Ljajić, A.; Milošević, N. Multilingual Transformer and BERTopic for Short Text Topic Modeling: The case of Serbian. In Proceedings of the Conference on Information Technology and its Applications, Da Nang, Vietnam, 28–29 July 2023; Springer Nature: Cham, Switzerland, 2023; pp. 161–173.
76. Aryani, D.; Kharisma, I.L.; Sujjada, A.; Kamdan, K. Topic Modeling of the 2024 Election Using the BERTopic Method on Detik. com News Articles. *Inf. J. Ilm. Bid. Teknol. Inf. Komun.* **2024**, *9*, 171–180.
77. Ma, T.; Chen, H.; Zhang, K.; Shen, L.; Sun, H. The rheological intelligent constitutive model of debris flow: A new paradigm for integrating mechanics mechanisms with data-driven approaches by combining data mapping and deep learning. *Expert Syst. Appl.* **2025**, *269*, 126405. [CrossRef]
78. Axelborn, H.; Berggren, J. Topic Modeling for Customer Insights: A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls. Master's Thesis, Umeå University, Umeå, Sweden, 2023.
79. Aletras, N.; Stevenson, M. Evaluating Topic Coherence Using Distributional Semantics. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013), Potsdam, Germany, 19–22 March 2013; pp. 13–22.
80. Syed, S.; Spruit, M. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19–21 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 165–174.
81. Hoyle, A.; Goel, P.; Hian-Cheong, A.; Peskov, D.; Boyd-Graber, J.; Resnik, P. Is automated topic model evaluation broken? the incoherence of coherence. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2018–2033.
82. Bellaouar, S.; Bellaouar, M.M.; Ghada, I.E. Topic Modeling: Comparison of LSA and LDA on Scientific Publications. In Proceedings of the 2021 4th International Conference on Data Storage and Data Engineering, Barcelona, Spain, 18–20 February 2021; pp. 59–64.
83. Miles, S.; Yao, L.; Meng, W.; Black, C.M.; Miled, Z.B. Comparing PSO-based clustering over contextual vector embeddings to modern topic modeling. *Inf. Process. Manag.* **2022**, *59*, 102921. [CrossRef]
84. Newman, D.; Bonilla, E.V.; Buntine, W. Improving Topic Coherence with Regularized Topic Models. In Proceedings of the Advances in Neural Information Processing Systems 24 (NIPS 2011), Granada, Spain, 12–14 December 2011; Volume 24.
85. Wang, J.; Pradhananga, N.; Teizer, J. Automatic Fall Risk Identification Using Point Cloud Data in Construction Excavation. In *Computing in Civil and Building Engineering*; ASCE Library: Reston, VA, USA, 2014; pp. 981–988.
86. Kang, Y.; Siddiqui, S.; Suk, S.J.; Chi, S.; Kim, C. Trends of fall accidents in the U.S. construction industry. *J. Constr. Eng. Manag.* **2017**, *143*, 04017043. [CrossRef]
87. Fonseca, E.D.; Lima, F.P.; Duarte, F. From construction site to design: The different accident prevention levels in the building industry. *Saf. Sci.* **2014**, *70*, 406–418. [CrossRef]
88. Selleck, R.; Cattani, M.; Hassall, M. Proposal for and validation of novel risk-based process to reduce the risk of construction site fatalities (Major Accident Prevention (MAP) program). *Saf. Sci.* **2023**, *158*, 105986. [CrossRef]

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