



# Accidents analysis of mining industry through semantic text representation and dimensionality reduction: An integrated clustering framework

Abid Ali Khan Danish<sup>a,\*</sup>, Snehamoy Chatterjee<sup>a</sup>, Kumar Vaibhav Raj<sup>b</sup>, Rennie Kaunda<sup>c</sup>, Hugh Miller<sup>c</sup>, Aref Majdara<sup>d</sup>

<sup>a</sup> Department of Geological and Mining Engineering and Sciences, Michigan Technological University, MI, USA

<sup>b</sup> National Institute for Occupational Safety and Health, Pittsburgh Mining Research Division, Pittsburgh, PA, USA

<sup>c</sup> Colorado School of Mines, 1600 Illinois St., Colorado School of Mines, Golden, CO, 80401, USA

<sup>d</sup> Washington State University, 14204 NE Salmon Creek Ave, Vancouver, WA, 98686, USA

## ARTICLE INFO

### Keywords:

Sentence transformer embeddings  
Nonlinear dimensionality reduction  
K-means clustering  
Mining safety  
Machine learning for hazards analysis  
Accident narratives clustering

## ABSTRACT

Despite significant improvements in worker health and safety in recent decades, the mining industry still experiences fatal and non-fatal accidents. This underscores the critical need for a more nuanced understanding of accident causation patterns through accident data analysis. While conventional analytical approaches have yielded valuable insights, the extensive information embedded within text-based accident narratives remains underutilized. To address this gap, this study presents an artificial intelligence (AI)-based framework that integrates transformer-based natural language processing, nonlinear dimensionality reduction, and unsupervised machine learning to analyze and cluster accident narratives from the U.S. mining industry. Specifically, this study uses Sentence-BERT (SBERT), a sentence embedding model based on Bidirectional Encoder Representations from Transformers (BERT), to extract the high-dimensional semantical representation of accident narratives. These embeddings are then mapped to a low-dimensional space using the Uniform Manifold Approximation and Projection (UMAP) technique, followed by clustering with the k-means algorithm and subsequent hazard-focused cluster analysis. The primary contribution to AI lies in demonstrating the effectiveness of combining modern sentence embedding techniques with dimensionality reduction and clustering for the semantic analysis of safety-related narratives. From an engineering standpoint, this framework enables the identification of latent accident patterns that can inform hazard detection and guide safety interventions in the mining industry. The resulting clusters reveal diverse accident patterns across mining operations. In clusters associated with underground mining, a high proportion of incidents (ranging from 84 to 98 %) involved no equipment, with distinct injury patterns: torso injuries (67 %) from over-exertion, lower extremity injuries (58 %) from slips/falls, and upper extremity injuries from over-exertion (95 %) and material handling (85 %). Equipment-related clusters revealed strong associations with drilling tools (92 %), loaders (98 %), and bolting equipment (96 %). Clusters associated with strip/quarry/open pit operations exhibited a high frequency of vehicle-related accidents (98 % transportation, 99 % loaders), often resulting in multiple body part injuries. Milling operation clusters show 52–97 % no-equipment accidents, with injury patterns similar to those in underground mining. Additionally, noise-induced hearing loss (96–97 %) was observed in clusters spanning all mining operation types. These findings offer actionable insights for safety professionals and support data-driven, targeted interventions in the mining industry.

## 1. Introduction

Workers' health and safety in the U.S. mining industry has improved significantly over recent decades, as evidenced by declining injury and fatality rates. According to the U.S. Bureau of Labor Statistics (2023), the

mining sector reports one of the lowest nonfatal injury and illness rates (1.3 per 100 full-time workers) among major industry sectors, compared to construction (2.3), manufacturing (2.8), and transportation (4.5) (BLS). This improvement is further demonstrated by National Institute for Occupational Safety and Health (NIOSH) data showing a substantial

\* Corresponding author.

E-mail address: [aadanish@mtu.edu](mailto:aadanish@mtu.edu) (A.A.K. Danish).

<https://doi.org/10.1016/j.engappai.2025.112164>

Received 21 April 2025; Received in revised form 29 June 2025; Accepted 29 August 2025

Available online 3 September 2025

0952-1976/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

decrease in nonfatal injury rates from 3.85 per 100 Full-time equivalent (FTE) in 2000 to 1.43 in 2023. Similarly, occupational fatalities have declined from 85 deaths (30.15 per 100,000 FTE) in 2000 to 40 deaths (15.59 per 100,000 FTE) in 2023, though annual fluctuations persist. Despite these improvements, persistent occupational hazards continue to pose significant challenges. In 2023 alone, the industry recorded 3664 nonfatal lost-time injuries and 40 fatalities. This indicates that occupational hazards continue to pose significant challenges despite advanced safety protocols and regulations (National Institute for Occupational Safety and Health (NIOSH)a) (National Institute for Occupational Safety and Health (NIOSH)b). To address this, it is crucial to determine potential areas of concern by examining the causes of accidents from past accidents and injury data, and to manage these potential hazards through effective accident data analysis. Such analysis can provide hidden trends and patterns in accident data that can help find the hazards associated with several operations and help mitigate future incidents and ensure a safe working environment for workers (Sarkar et al., 2019).

However, current accident data analysis methods face significant limitations. Traditional rule-based and statistical-based Natural Language Processing (NLP) methods, such as bag-of-words or Term Frequency-Inverse Document Frequency (TF-IDF) are often constrained by scalability and their inability to capture nuanced semantic relationships within unstructured text, frequently obscuring less frequent but critically important patterns (Kemajl et al., 2024) (Smetana et al., 2024). As highlighted by Kemajl et al. (2024), simplistic statistical methods like Simple Linear Regression Method (SLRM) are insufficient for the inherent complexity of accident data, reinforcing their limitations in mining accident forecasting and risk assessment. Furthermore, existing approaches lack robust mechanisms to effectively integrate structured metadata with unstructured narratives, limiting their ability to achieve comprehensive analysis of factors contributing to accidents and hindering the development of a holistic understanding of accident causation patterns (Smetana et al., 2024) (Bridgelall and Tolliver, 2024). While recent developments in NLP, such as sophisticated embedding models, can encapsulate rich semantic information, their high dimensionality poses challenges for direct analysis and visualization (Ayesha et al., 2020).

This research addresses these critical limitations in current mining accident analysis through a novel integrated framework. This framework combines semantic enrichment of accident narratives with advanced dimensionality reduction and clustering techniques. It integrates structured metadata with unstructured narratives before embedding extraction, uses Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) for generating sentence-level semantic embeddings, applies Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) for dimensionality reduction while preserving semantic relationships, and employs k-means clustering (Ikotun et al., 2023) for automated pattern discovery. This approach enables the identification of latent accident patterns and causal relationships that are not detectable through traditional statistical-based or word-level embedding methods. The framework's effectiveness in understanding accident occurrence patterns for mining safety management is validated using Mine Safety and Health Administration (MSHA) accident data.

## 2. Literature review

### 2.1. Current challenges in accident data analysis

The contemporary analysis of accident data faces multifaceted challenges that stem from the inherent complexity and heterogeneous nature of available information sources (Ganguli et al., 2021) (V Raj and Tarshizi, 2020). Accident reports typically comprise both structured metadata, such as categorical variables describing accident type and equipment type, and unstructured narrative descriptions that capture contextual details about incident circumstances and can reveal deeper

insights into incident causation (Ma and Chen, 2024) (Rose et al., 2020). Analyzing such unstructured textual data presents significant challenges due to its volume, complexity, and the semantic relationships embedded within natural language descriptions. While the analysis of such accident narratives has evolved significantly with the advent of advanced NLP techniques and machine learning (ML) models, which have been successfully explored across different industries such as construction (Zhong et al., 2020a) (Zhong et al., 2020b) (XU et al., 2021) (Tixier et al., 2016), aviation (Srinivasan et al., 2019) (Rose et al., 2020), transportation (Williams et al., 2016) (Figueres-Esteban et al., 2016), and manufacturing (Sarkar et al., 2019), fundamental limitation persists.

Traditional statistical-based NLP methods, such as bag-of-words, Term Frequency (TF), Document Frequency (DF), and term TF-IDF, have provided valuable insights into accident causation and risk patterns. However, these statistical methods have limitations in capturing the semantic relationships between words, which are crucial for representing the distributed meaning of text. In accident data analysis, understanding the semantic meaning of words or phrases is essential when grouping similar text data, as it facilitates the identification of similar traits or behaviors (Kemajl et al., 2024) (Smetana et al., 2024) (Babić et al., 2020a). Additionally, these frequency-based approaches tend to emphasize commonly reported events while potentially obscuring less frequent but critically important patterns. Furthermore, many existing approaches, including some early ML applications, overlook the integration of metadata with unstructured narrative descriptions, leading to an incomplete understanding of accident causation (Smetana et al., 2024) (Rose et al., 2020). This merging of metadata with accident narratives represents a crucial enrichment strategy for accident analysis, as it enables the integration of structured and unstructured data to enhance the semantic representation of incidents. Accident narratives alone may omit critical structured details, such as the specific nature of the accident or the part of the body involved, which are systematically recorded in metadata fields and are essential for comprehensive analysis (Smetana et al., 2024) (Jiang and Wang, 2024) (Darda'u Rafindadi et al., 2025). Therefore, concatenating or otherwise integrating metadata variables with the narrative embeddings enriches the feature space, allowing the model to leverage both the latent semantic information and explicit categorical attributes, thereby supporting more robust downstream analytical tasks. This enrichment is particularly valuable in the context of mining accident analysis, where the complexity of incidents often arises from the interplay of multiple factors.

Furthermore, recent developments in NLP have introduced sophisticated transformer-based embedding models capable of generating dense, high-dimensional vector representations of textual data. These embeddings encapsulate rich semantic information, enabling more precise analysis of accident reports and facilitating the identification of latent patterns that may not be apparent through traditional methods (Ma and Chen, 2024) (Ahmadi et al., 2024). The use of pretrained models such as SBERT (Reimers and Gurevych, 2019), a modification of the pre-trained Bidirectional Encoder Representations from Transformers (BERT) framework (Devlin et al., 2019), enables the extraction of sentence-level embeddings that are contextually informed, capturing both syntactic and semantic nuances present in accident narratives. By combining these embeddings with structured metadata, comprehensive feature sets can be constructed that reflect both the textual and contextual dimensions of each incident (Ma and Chen, 2024). However, the high dimensionality of these embeddings poses challenges for direct analysis and visualization (Babić et al., 2020b) (Ayesha et al., 2020). Dimensionality reduction techniques, such as UMAP, have emerged as effective solutions for transforming complex, high-dimensional data into lower-dimensional spaces while preserving intrinsic structure and relationships among data points (McInnes et al., 2018). However, these techniques have not been universally optimized for semantic-rich embeddings in mining safety contexts or for integrated metadata enrichment strategies. These gaps collectively hinder the identification of

latent accident patterns and causal relationships for robust accident data analysis and safety management.

## 2.2. Advances in accident data analysis

The application of advanced NLP and ML techniques to accident data analysis represents a paradigmatic shift from traditional statistical approaches toward automated, scalable, and semantically informed methodologies. Early ML applications in mining safety primarily relied on structured data. Onder (2013) applied binary logistic regression to identify high-risk worker categories, while Tixier et al. (2016) employed Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB) to predict injury types, and Amoako et al. (2021) used multiclass logistic regression to MSHA data, to access the role of worker experience. While these models achieved reasonable predictive performances, they lacked interpretability and leveraging of semantic content from accident descriptions critical to understanding accident causation. Subsequent work emphasized the value of accident narratives. Ganguli et al. (2021) demonstrated that models incorporating text features outperformed those using only structured data, though they relied on frequency-based representations. Raj and Tarshizi (V Raj and Tarshizi, 2020) used co-occurrence matrices and Latent Dirichlet Allocation (LDA) for clustering fatality reports, uncovering thematic patterns but lacking semantic depth and metadata integration. These studies underscore the limitations of early methods in capturing latent semantic relationships and the need for more advanced, semantically informed approaches that integrate both structured and unstructured data.

Traditional statistical text analysis methods provided foundational insights but revealed significant limitations. Tan et al. (2017) applied TF analysis and word cloud visualization to large-scale accident records for identifying safety risks, using basic text preprocessing and frequency calculation to highlight prominent safety concerns. Though useful for exploratory analysis, this approach was fundamentally constrained by simple frequency counts without semantic understanding. Li et al. (2018) created domain-specific lexicons and applied DF analysis to identify prevalent safety risk factors through frequency-based ranking. While systematic, this methodology required manual lexicon construction and subjective vocabulary selection, failing to capture semantic relationships between conceptually related but lexically different terms. These frequency-based approaches suffer from the fundamental limitation of not understanding semantic relationships between words, which is crucial for representing distributed meaning in safety analysis (Babić et al., 2020a).

The emergence of word embedding techniques marked a significant advancement in semantic representation capabilities. Word to Vector (Word2Vec) (Zhang, 2022) and Global Vectors for Word Representation (GloVe) (Valcamonico et al., 2022) approaches learn dense vector representations from large corpora to capture semantic and syntactic relationships between words. Zhang et al. (Zhang, 2022) demonstrated the effectiveness of Word2Vec skip-gram algorithms combined with neural networks for construction accident classification, showing improved robustness over traditional methods. However, Word2Vec operates at the word level and requires aggregation strategies for sentence representation, often losing contextual information. GloVe addresses some Word2Vec limitations by incorporating global co-occurrence statistics through matrix factorization, but it shares the fundamental constraint of word-level operation and the inability to generate context-dependent representations. Valcamonico et al. (2022) employed Doc2Vec and Hierarchical Dirichlet Processes for document-level embeddings in accident classification using various ML algorithms. While Doc2Vec provided document-level representations, it remained dependent on

document-level training data and complex hyperparameter tuning, limiting its practical applicability.

Recent transformer-based approaches have revolutionized natural language understanding through contextual embeddings. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) employs deep bidirectional transformers with masked language modeling to generate context-dependent representations. Jing et al. (2022) applied Word2Vec with bidirectional LSTM and attention mechanisms for chemical accident classification, demonstrating effective pattern extraction through correlation analysis of embeddings. Ramos et al. (2022) also employed the BERT text representations combined with numerical variables in multi-layer perceptrons for accident leave prediction, showing the value of integrating textual and structured information. Despite these advances and achieving state-of-the-art results in extracting features from text data (word and sentence embedding vectors) (Devlin et al., 2019) (Vaswani et al., 2017), BERT's network structure does not compute independent sentence embeddings (May et al., 2019) (Zhang et al., 2019) (Qiao et al., 2019). To overcome this limitation, single sentences are processed through BERT and fixed-size vectors are derived by either averaging the outputs or using the output of the special CLS token. However, these methods may not produce optimal sentence embeddings compared to models specifically designed for sentence-level representation (Devlin et al., 2019). Ma and Chen (2024) addressed this limitation by employing CoSENT-fine-tuned BERT models specifically optimized for Chinese sentence embeddings, combined with an Accimap methodology for systemic risk analysis in construction. Their approach used clustering algorithms to identify latent patterns from sentence-level representations. However, their methodology was limited by manual coding requirements for theme identification, lacked comprehensive integration of metadata with narrative content, and remained focused solely on the construction domain.

These transformer-based approaches have paved the way for more sophisticated integration methods using large language models. Smetana et al. (Ma and Chen, 2024) explored highway construction safety analysis using large language models, including text-embedding-ada-002, achieving superior performance in clustering safety-related incidents and automating incident categorization. This methodology demonstrated the effectiveness of state-of-the-art embedding models but was limited to the construction domain without integration between structured metadata and accident narratives. Ahmadi et al. (2024) developed automatic construction accident report analysis using large language models, including GPT-3.5, GPT-4.0, Gemini Pro, and LLaMA 3.1, achieving high classification accuracy for accident attributes through zero-shot learning approaches. While these studies demonstrate the continued evolution toward more sophisticated approaches for accident analysis across safety-critical domains, these methodologies rely on proprietary large models with billions of parameters that require substantial computational resources and dependency on commercial API access with associated ongoing costs, potentially limiting accessibility, reproducibility, and practical deployment for specialized research applications in resource-constrained environments (Zhang et al., 2025) (Wang et al., 2024).

Dimensionality reduction and clustering approaches have evolved to address the challenges of high-dimensional embedding, including the curse of dimensionality that can lead to data sparsity, computational inefficiency from poor algorithm scaling, and visual interpretation (Wang et al., 2024). Earlier methods, such as Principal Component Analysis (PCA) (Shao et al., 2019) and t-distributed Stochastic Neighbor Embedding (t-SNE), have been used to support clustering of text data, with t-SNE demonstrating effectiveness in local structure preservation

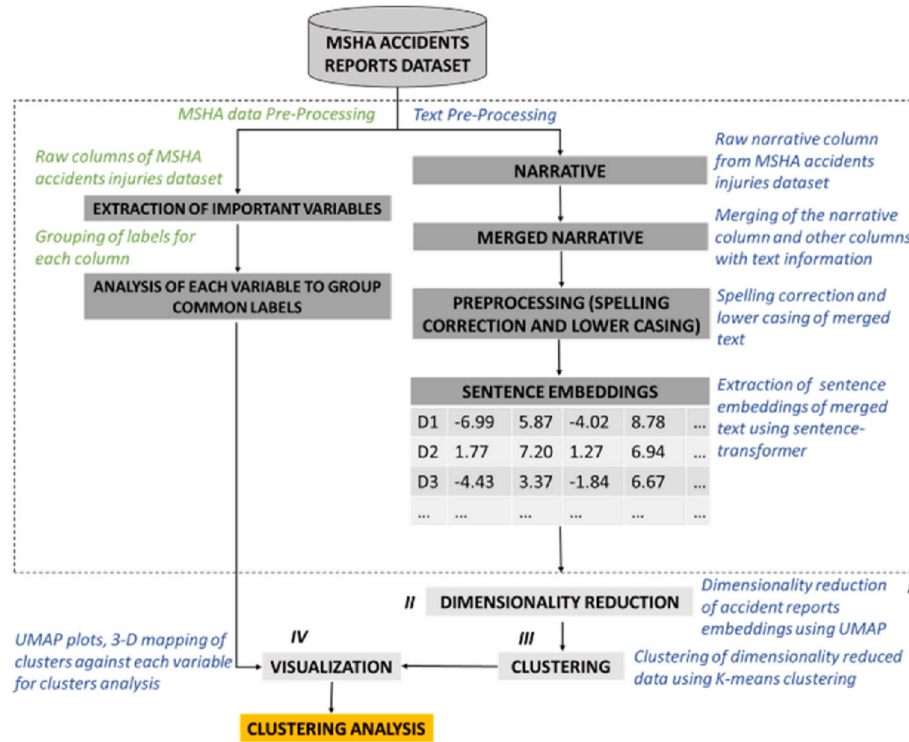


Fig. 1. An SBERT-UMAP framework for mining accidents analysis: An integrated approach to accidents narratives clustering.

but lacking global consistency and scalability (Rose et al., 2020). More recently, UMAP has gained prominence for its ability to preserve both local and global structures, enabling more meaningful clustering of complex embeddings (McInnes et al., 2018). Hozumi et al. (2021) demonstrated UMAP's effectiveness in preserving both local and global data structures during dimensionality reduction, showing improved clustering accuracy compared to traditional methods. However, most existing applications have not specifically addressed semantic-rich embeddings in safety contexts or integrated metadata enrichment strategies.

Based on this comprehensive review of existing methodologies reveals several critical gaps that limit current approaches to mining accident analysis. This research addresses these limitations through a novel integrated framework and proposes three key methodological advancements directly addressing these unresolved challenges in existing literature. First, the semantic enrichment strategy, which enriches accident narratives by merging them with text-based metadata variables before embedding extraction. This pre-embedding data fusion directly integrates the structural data with unstructured data, ensuring a more comprehensive feature set for analysis. Second, the SBERT-UMAP-k-means pipeline represents a novel combination specifically designed for mining accident analysis that addresses both semantic representation and dimensionality reduction challenges simultaneously. SBERT addresses traditional BERT limitations in generating independent sentence embeddings through fine-tuning on sentence pair tasks using siamese network architectures, enabling the extraction of semantically meaningful sentence-level representations that capture contextual dependencies within accident narratives. The integration of SBERT embeddings with UMAP dimensionality reduction addresses the curse of dimensionality while preserving both local and global semantic structures necessary for meaningful clustering. Third, this method enables

more comprehensive analysis of accident patterns and trends in different mining operations, revealing relationships that may not be apparent through traditional techniques such as TF-IDF, Word2Vec, or GloVe.

### 3. Methodology

This study implements a multi-stage methodology to analyze the MSHA accidents and injuries dataset, with the goal of uncovering underlying accident patterns through systematic text analysis. The methodology, as illustrated in.

Fig. 1, encompasses multiple stages of data processing and analysis. The initial phase involves extracting initial incident narratives (Form 7000-1) from the MSHA dataset. Relevant text-based variables are filtered and compiled from the same dataset. The dataset then undergoes a cleaning process to remove NAN values and duplicate text narratives involving multiple injuries or fatalities, as these share the same narrative. Following this, these narratives are merged with the other text-based variables of the dataset to create comprehensive accident descriptions. This merged text data then undergoes spelling correction and lowercase conversion to standardize the input for further processing. Moreover, dense vector representations of this merged text data are obtained using the pre-trained sentence transformer, followed by the dimensionality reduction of these embeddings and grouping these resultant low-dimensional embeddings by the application of the k-means clustering technique. Finally, variables from MSHA data are visualized against each cluster to reveal the distinctive accident patterns and circumstances that characterize each group.

The underlying computational procedure corresponding to this framework is detailed in Algorithm 1.

**Algorithm 1.** SBERT-UMAP-kmeans Framework for Mining Accident Pattern Analysis



```
procedure ACCIDENT_CLUSTERING_FRAMEWORK (MSHA_dataset, text_vars, sbert_model, umap_params)
```

```
  // Step 1: Data Cleaning
```

```
  cleaned_data ← REMOVE_MISSING_AND_DUPLICATES (MSHA_dataset)
```

```
  // Step 2: Domain-Based Label Grouping
```

```
  metadata_vars ← {subunit, mining_equipment, classification, accident_type,  
                  injury_source, nature_of_injury, injured_body_parts, degree_of_injury}
```

```
  grouped_labels ← GROUP_LABELS (metadata_vars)
```

```
  // e.g., mining_equipment: 70 → 6, accident_type: 45 → 10, body_parts: 47 → 5
```

```
  // Step 3: Merge Text Variables into Narrative
```

```
  merged_narratives ← []
```

```
  for each record r in cleaned_data do
```

```
    narrative ← r.accident_narrative
```

```
    var_text ← CONCAT_TEXT_FIELDS(r[text_vars])
```

```
    full_text ← PREPROCESS_TEXT CONCAT (narrative, var_text)
```

```
    merged_narratives.append(full_text)
```

```
  end for
```

```
  // Step 4: Sentence Embedding using SBERT
```

```
  embeddings ← SBERT_ENCODE (merged_narratives, sbert_model)
```

```
  // Step 5: Dimensionality Reduction using UMAP
```

```
  umap_model ← INITIALIZE_UMAP (umap_params)
```

```
  umap_embeddings ← umap_model.FIT_TRANSFORM (embeddings)
```

```
  // Step 6: Optimal k Selection for Clustering
```

```
  optimal_k ← SELECT_K_USING_SILHOUETTE (umap_embeddings)
```

```
  // Step 7: Kmeans Clustering
```

```
  kmeans_model ← INITIALIZE_KMEANS (optimal_k)
```

```
  initial_clusters ← kmeans_model.FIT_PREDICT (umap_embeddings)
```

```
  // Step 8: Sub-clustering for Clusters without Clear Dominant Patterns
```

```
  final_clusters ← COPY (initial_clusters)
```

```
  for each cluster c in initial_clusters do
```

```
    if NOT HAS_CLEAR_DOMINANT_PATTERNS (c, grouped_labels) then
```

```
      sub_clusters ← HIERARCHICAL_CLUSTERING(c)
```

```
      final_clusters ← UPDATE_LABELS (final_clusters, sub_clusters)
```

```
    end if
```

```
  end for
```

```
  // Step 9: Cluster Characterization
```

```
  cluster_patterns ← []
```

```
  for each cluster fc in final_clusters do
```

```
    pattern ← EXTRACT_DOMINANT_LABELS (fc, grouped_labels)
```

```
    cluster_patterns.append(pattern)
```

```
  end for
```

```
  return final_clusters, umap_embeddings, cluster_patterns
```

```
end procedure
```

```

procedure CONCAT_TEXT_FIELDS (variables)
  text ← ""
  for each var in variables do
    text ← text + var + ". "
  end for
  return text
end procedure

procedure EXTRACT_DOMINANT_LABELS (cluster, grouped_labels)
  dominant_labels ← {}
  total ← SIZE (cluster)

  for each variable in grouped_labels do
    label_counts ← COUNT_LABELS (cluster, variable)
    max_label ← null
    max_prop ← 0

    for each label in label_counts do
      prop ← label_counts[label] / total
      if prop > max_prop then
        max_label ← label
        max_prop ← prop
      end if
    end for

    // A cluster is considered to have a dominant label for a variable
    // if one label appears in >50% of instances (threshold = 0.5)
    if max_prop > 0.5 then // dominance threshold
      dominant_labels[variable] ← (max_label, max_prop)
    end if
  end for

  return dominant_labels
end procedure

```

. (continued).

### 3.1. MSHA dataset

MSHA is responsible for the inspection and enforcement of health and safety regulations for mining operations in the U.S. to prevent fatal and non-fatal injuries and occupational illnesses. Under MSHA rules and regulations, designated MSHA representatives record accidents and notify MSHA regarding injuries and illnesses occurring at the mine. This reporting is primarily done through Form 7000-1, which is completed by the principal officer in charge of health and safety or the supervisor after the accident. Fatal accidents and accidents defined in 30 CFE 50.10 (including injuries with reasonable potential to cause death, entrapment for more than 30 min, unplanned inundation, unplanned ignition or

explosion, unplanned fires, and damage to hoisting equipment) are reported immediately, while all other accidents, injuries, and illnesses are to be reported within 10 days. These accident reports are maintained in the MSHA database, which includes important information related to all accidents, injuries, and illnesses reported by mine operators and contractors. While MSHA has records from prior years available through the NIOSH site, the publicly accessible database includes data from January 1, 2000, onwards. The MSHA injury dataset used for this study provides information on 247514 accidents from January 1, 2000, to December 31, 2022. Each instance of data has text-based narratives along with other text-based variables that provide information associated with accidents, such as subunit, mining equipment, classification, accident type, injury source, nature of the injury, injured body parts, and degree of injury, based on MSHA's classification. These variables detail key characteristics of each accident, such as degree of injury including

**Table 1**

Example of systematic text variables merging with accident narrative from the MSHA.

| Subunit                 | Mining Equipment                                                    | Classification           | Accident Type                | Injury Source           | Nature of Injury | Injured Body Parts     | Narrative                                                                                                      | Merged Narrative                                                                                                                                                                                                                                                                                                                                                |
|-------------------------|---------------------------------------------------------------------|--------------------------|------------------------------|-------------------------|------------------|------------------------|----------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Strip, Quarry, open pit | Carriage-Mounted drills, rail, Rubber-Tired, jumbo, Air-Track drill | Slip or fall of a person | Fall onto or against objects | Surface mining machines | Fracture, chip   | Lower Leg/Tibia/Fibula | Employee was in the process of getting down off the drill rig and slipped and fell hitting leg on the mud pan. | Strip, Quarry, open pit. Carriage-Mounted drills, rail, Rubber-Tired, jumbo, Air-Track drill. Slip or fall of person. Fall onto or against objects. Surface mining machines. Fracture, chip. Lower Leg/Tibia/Fibula. Employee was in the process of getting down off the drill rig and slipped and fell hitting leg on the mud pan. Hitting leg on the mud pan. |

**Table 2**

Grouping of labels for each variable.

| Injured Body Parts              | Grouping                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|---------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Head                            | 'Mouth/Lip/Teeth/Tongue/Throat/Taste', 'Eye(S) Optic Nerve/Vision', 'Ear(S) Internal & External', 'Ear (S) External', 'Jaw Include Chin', 'Nose/Nasal Passages/Sinus/Smell', 'Head, NEC', 'Neck', 'Face, Multiple Parts', 'Face, NEC', 'Ear(s) internal & Hearing', 'Brain', 'Head, Multiple Parts', 'Scalp', 'Skull'                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Upper extremities               | 'Finger(S)/Thumb', 'Arm, NEC', 'Arm, Multiple Parts', 'Hand (Not Wrist or Fingers)', 'Forearm/Ulnar/Radius', 'Shoulders', 'Collarbone/Clavicle/Scapula', 'Elbow', 'Wrist', 'Upper Arm/Humerus', 'Upper Extremities', 'Upper Extremities, Multiple'                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Torso                           | 'Back (Muscles/Spine/S-Cord/Tailbone)', 'Abdomen/Internal Organs', 'Trunk, NEC', 'Chest (Ribs/Breastbone/Chest Organs)', 'Trunk, Multiple Parts', 'Hips (Pelvis/Organs/Kidneys/Buttocks)'                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Lower extremities               | 'Leg, Multiple Parts', 'Toe(S)/Phalanges', 'Lower Extremities', 'Ankle', 'Leg, NEC', 'Thigh/Femur', 'Foot (Not Ankle/Toe)/Tarsus/Metatarsus', 'Knee/Patella', 'Lower Leg/Tibia/Fibula', 'Lower Extremities, Multiple Parts'                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Multiple body parts             | 'Multiple Parts (More Than One Major)', 'Body Systems', 'Body Parts, NEC'                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <b>Mining Equipment</b>         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Loaders                         | "Load-haul-dump, Scoop tram, Transloader, Unitrac, S&S Battery", "Mucking machine, Overshot loader, Cryderman", "Shovel, Power shovel, Backhoe, Track hoe, Dragline - Big Muskie", "Pneumatic blasting agent loader, Driller loader, Prill loader", "Front-end loader, Tractor-shovel, Payloader, High lift, Skip loader", "Gathering arm loader, Conway loader, Coal loading machine", "Conveyor, Belt feeder, Stage loader, Hopper shaker, Belt structure"                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| Transportation vehicles/ Trucks | "Basket, Basket scaler, Self-propelled hydraulic boom", "Trucks, Service truck, Utility truck, Pickup, Water truck, Fuel truck", "Crane, Cherry picker, Lift basket, Scissor truck, Boom truck", "Ore haulage trucks - off highway trucks", "Ore haulage trucks - on highway trucks", "Boats, Barges, Other water transportation", "Mine car, Ore or coal car, Boxcar, Hopper car", "Air transportation, Planes, Helicopter", "Scraper loader, Tractor scraper, Pan scraper, Elevating scraper", "Locomotive, (motor) - rail-mounted (Battery, Steam, Electric, Air)", "Mine car, Timber truck, Nipper truck", "Shuttle car, Buggy, Ram car, Young buggy, Teletram car", "Elevator, Skip, Cage, Buckets, Mancage, Slope car", "Slusher, Scraper hoist, Slusher hoist", "Pump, Slurry pump, Sump pump", "Gunite machine, Shotcrete", "Bulldozer, Dozer, Crawler tractor, Push cat", "Tractor, Supply car", "Chute", "Raise climber", "Forklift, Plant cat, Telehandler, Hyster, Lift truck, Skid steer", "Tugger, Air winch, Air hoist, Air jack, Electric winch", "Track equipment, Re-railer, 335 track shifter, Track jack", "Mancar, Mantrip, Personnel carrier, Porta bus, Jeep, Jitney, ATV" |
| Supporting/Bolting equipment    | "Longwall subparts, Duke, Dowdy jack, Ramjack, Longwall shield", "Rock or roof bolting machine, Pinning machine, Truss bolter"                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Drilling equipment              | "Rock drill, Jackleg, Airleg, Drifter, Stoper, Buzzy, Jackhammer", "Hand tools (powered)", "Hand tools (not powered)", "Carriage-mounted drills, Rail, Rubber-tired, Jumbo, Air-track drill", "Machine, NEC - Wheelbarrow, Well drilling Rig, Post hole auger", "Electric drills, Coal drills", "Auger machine, Highwall miner, Auger highwall drill, Auger drill"                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Other machines and equipment's  | "Packaging machine, Bagger, sewing machine, Palletizer", "Welding machine, Torch, Cutting torch, Arc welder, Air arc", "Hoist, Car dropper, Hydraulic jack, Mine car retriever, Condor", "Blow pipe, Blow gun, Air lance", "Longwall machine", "Road grader, Motor grader, Motor patrol, Grader, Road scraper", "Continuous miner, Tunnel borer, Road header", "Cutting machine, Chain cutter", "Fan", "Air compressor", "Dimension stone machinery, Gangwire saw, Guillotine,                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |

**Table 2 (continued)**

|                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Processing equipment              | Hydrosplit", "Rock dusting machine, Trickle duster", "Raw coal storage, Tipple, Dump bins", "Dredge", "Impactor", "Hydraulic jets, Monitor", "Milling machinery, Block press, Ballast machine", "Washers", "Classifier, Screw classifier, Spiral classifier, Cyclone classifier", "Flotation and filters", "Mill, grinding (Rod, Ball, Autogenous, Pug, Hammer)", "Bench grinder, Drill press, Band/Table saw, Sandblaster", "Grizzlies for coarse screening, Scalping or skimming of bulk material", "Crusher, Breaker, Mills (ball and rod), Feeder breaker", "Screen, Vibrator, Shaker, Rock screener, Scalper" |
| <b>Accident type</b>              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Without injuries                  | "Accident type, without injuries"                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| Fall                              | "Fall from headframe, derrick or tower", "Fall to lower level, NEC", "Fall from ladders", "Fall from piled material", "Fall onto or against objects", "Fall from machine", "Fall to the walkway or working surface", "Fall from scaffolds, walkways, platforms", "Fall down raise, shaft, or manway", "Fall downstairs"                                                                                                                                                                                                                                                                                            |
| Caught in                         | "Caught in, under or between a moving and a stationary object", "Caught in, under or between running or meshing objects", "Caught in, under or between collapsing material or buildings", "Caught in, under or between two or more moving objects", "Caught in, under or between NEC"                                                                                                                                                                                                                                                                                                                              |
| Contact with                      | "Contact with cold objects or substances", "Contact with heat", "Contact with cold", "Contact with electrical current", "Contact with hot objects or substances")                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| Struck by                         | "Struck by ... NEC", "Struck by flying object", "Struck against a moving object", "Struck by falling object", "Struck against stationary object", "Struck by rolling or sliding object", "Struck by concussion", "Struck by powered moving object"                                                                                                                                                                                                                                                                                                                                                                 |
| Over-exertion and bodily reaction | "Over-Exertion NEC", "Over-exertion in lifting objects", "Over-exertion in welding or throwing objects", "Over-exertion in pulling or pushing objects", "Bodily reaction, NEC"                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Burns                             | "Flash burns (electric)", "Flash burns (welding)", "Inhalation of radiations, caustics, toxic and noxious substances", "Ingestion of radiations, caustics, toxic and noxious substances"                                                                                                                                                                                                                                                                                                                                                                                                                           |
| Drowning                          | "Drowning"                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Rubbed                            | "Rubbed or abraded"                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| NEC                               | "NEC"                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |

accident only, fatality, permanent total or permanent partial disability, days away from work only, days away from work and restricted activity, days restricted activity only, no days away from work with no restrictions, occupational illness, injuries due to natural causes, injuries involving non-employees, and all other cases including first aid, and are extracted and evaluated for cluster-based analysis of accident patterns in this study.

### 3.2. Pre-processing

To conduct this study, the MSHA accidents and injuries dataset is used, which is extracted from the MSHA database and pre-processed using the Python programming language. Pandas (pandas, 2024) and NumPy (2004) packages are used in the Python environment to process the txt file provided by the online MSHA database (Mine Safety and Health Administration). The data extraction and pre-processing are

**Table 3**

SBERT model used for the extraction of embeddings of narratives.

| SBERT model (All-MiniLM-L12-v2) |              |
|---------------------------------|--------------|
| Maximum Sequential Length       | 256 words    |
| Output Dimensions               | 384          |
| Normalized Embeddings           | True         |
| Pooling                         | Mean Pooling |

conducted in two steps: 1) merging of relevant text-based features with accident narratives for text enhancement and their pre-processing, and 2) extraction of important variables from metadata and label grouping. The first step in the data extraction and pre-processing is to process and clean the text-based event narratives that represent all the available accident records; however, each event has text-based variables apart from narratives, which contain valuable information related to each event. Merging the text from these variables with narratives will enhance their semantic richness by incorporating additional relevant information. Therefore, these text-based variables, i.e., subunit, mining equipment, classification, accident type, injury source, nature of the injury, and injured body parts, are concatenated with the accident narrative for each instance before processing the text for spelling correction and lower casing. Table 1 provides an example of an instance of combining the text-based event narrative with other text-based variables. The selection of these specific text-based variables for merging was guided by domain expertise and their recognized importance in occupational safety research, as they collectively capture critical dimensions of accident characterization. Specifically, they represent contextual aspects such as the location of the incident (subunit), the equipment involved (mining equipment), the regulatory classification of the event (classification), the mechanism through which the accident occurred (accident type), and the proximate cause and consequences of the event (injury source, nature of injury, and injured body parts). Once selected, the actual merging process follows a systematic, automated concatenation approach where all selected variables are sequentially appended to the accident narrative consistently across all data instances. The Python libraries NLTK (NLTK) and re (Regular expressions operations) are used for spelling correction and lower casing of merged text-based narrative. Moreover, the word 'Employee' was referred to as 'E.E.' in many instances, which were corrected using regular expressions.

The second step involves the selection of key variables from the MSHA dataset that are deemed essential for accidents pattern analysis. The MSHA data includes variables that directly capture essential aspects of mining accidents. These variables - subunit, mining equipment, classification, accident type, injury source, nature of the injury, injured body parts, and degree of injury - inherently describe where accidents occurred, what equipment was involved, how they happened, and their impact on workers. Therefore, these variables were extracted for clustering analysis. To facilitate meaningful accident pattern analysis across clusters, the labels within each of these variables are grouped according to similarity in functionality, type, or operational context. For example, the 'Mining equipment' variable contains information about the equipment involved in accidents. To gain more meaningful insight, the equipment is grouped based on equipment types, such as transportation vehicles, loaders, trucks, etc. This grouping helps to understand the broader categories of equipment involved in accidents rather than specific individual equipment. Therefore, mining equipment, accident type, and injured body type variables are grouped based on domain expert knowledge. The mining equipment, accident type, and injured body part have originally 70, 45, and 47 labels, respectively, based on MSHA's classification. These labels are grouped into 6, 10, and 5 for clustering analysis, respectively. Table 2 provides the grouping of these variables in detail, where the injured body part is grouped as head, upper extremities, torso, lower extremities, and multiple body parts. Similarly, mining equipment is grouped as loaders, transportation vehicles, support/bolting equipment, drilling equipment, processing equipment, and other machines and equipment. Also, accident type labels are grouped as without injuries, fall, caught in, contact with, struck by, over-exertion & bodily reaction, burns, drowning, rubbed, and Not Elsewhere Classified (NEC). Moreover, the original labels are used for the subunit, nature of injury, injury source, classification, and degree of injury variables labels because the labels of these variables didn't require any further grouping, as MSHA's grouping is significant for final analysis.

### 3.3. Sentence embedding

To analyze text-based accident data, text from accident narratives needs to be converted into a form that a computer can understand and process. Natural Language Processing (NLP) techniques provide several methods to convert text into a vector matrix for the computer to understand and process. Traditional methods like one-hot vector representation have shortcomings such as loss of word order, word independence, and high dimensionality. To solve these shortcomings, the distributive representation of words is considered as a more suitable and powerful representation of words to capture the semantic meaning of words. The transformer-based state-of-the-art NLP framework like BERT can be used to extract the richer meaning of the text in numerical form. SBERT, a modification of the pre-trained BERT framework, uses Siamese and triplet network structures to extract semantically meaningful and dense vector representations for sentences. In this study, pre-trained SBERT model *all-MiniLM-L12-v2* is used to extract the sentence embeddings for accident text data such that semantically similar narratives have the same representation and are close to each other in vector space. The sentence-transformers library in Python is utilized to implement this embedding process. The *all-MiniLM-L12-v2* model processes input sequences up to 256 words and produces 384-dimensional normalized output vectors. It employs mean pooling to combine individual word embedding into a single sentence-level representation. Table 3 provides a detailed overview of the model's specifications.

### 3.4. Dimensionality reduction

The dimensionality reduction is the central component in high-dimensional data visualization and clustering analysis. The sentence transformer used to extract the sentence embeddings for MSHA accident narratives results in a high-dimensional vector representation of these narratives. Specifically, each narrative is transformed into a 384-dimensional vector, capturing its semantic content in a numerically rich but complex format. While these high-dimensional embeddings contain valuable information, they pose challenges for direct analysis. It is impractical to perform clustering analysis or visual examination on such high-dimensional data due to the curse of dimensionality causing data sparsity, computational inefficiency from poor algorithm scaling, and human limitations in perceiving relationships beyond three dimensions. However, these high-dimensional sentence embeddings can be reduced to low-dimensional embeddings for robust clustering and visual examination using dimensionality reduction techniques such as Principal component analysis (PCA), t-SNE, or UMAP. PCA is a linear technique that has been most commonly used; however, nonlinear techniques like t-SNE and UMAP have been recognized as having more importance. t-SNE is very efficient in preserving the local structure of the data. Still, it has a high computational cost, loses large-scale information (inter-cluster relationship), and is unable to represent large datasets effectively. UMAP, on the other hand, is a graph-based manifold learning algorithm that accurately represents the local and global structure with less computational cost (Becht et al., 2019).

UMAP addresses the challenges mentioned earlier by efficiently reducing the dimensionality of our 384-dimensional text data embeddings. It constructs a high-dimensional graph representation of the data and optimizes a low-dimensional layout to minimize the cross-entropy between the two. This process relies on three main assumptions: uniform data distribution on a Riemannian manifold, locally constant Riemannian metric, and local connectedness of the underlying manifold.

The algorithm begins by constructing a weighted k-neighbor graph, where the weight of an edge between points  $x_i$  and  $x_j$  is defined as:

$$\sum_{j=1}^k \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = w(x_i, x_j) \quad (1)$$

Here,  $d$  is a distance metric,  $\rho_i$  is a local connectivity parameter, and  $\sigma_i$  is



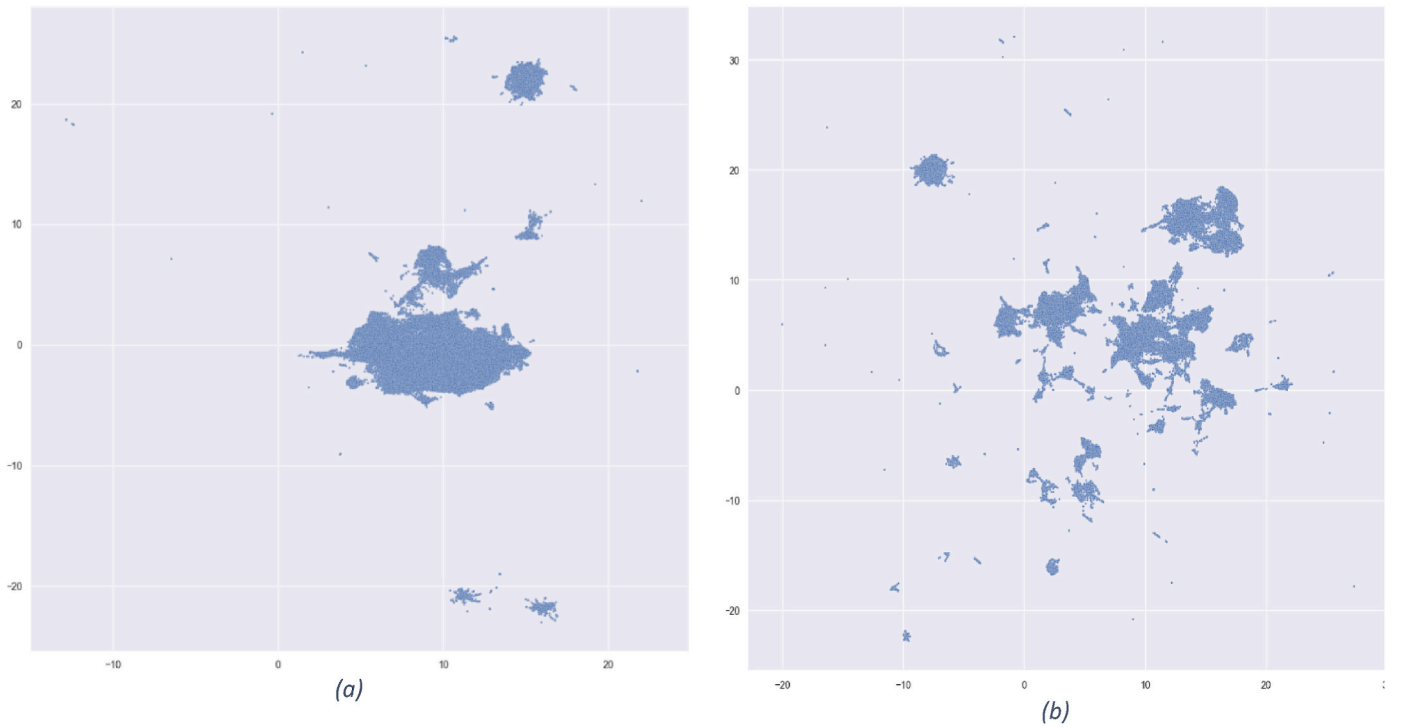


Fig. 2. 2-D Visualization of low-dimensional UMAP embeddings of (a) narratives only; (b) narratives merged with other text-based variables.

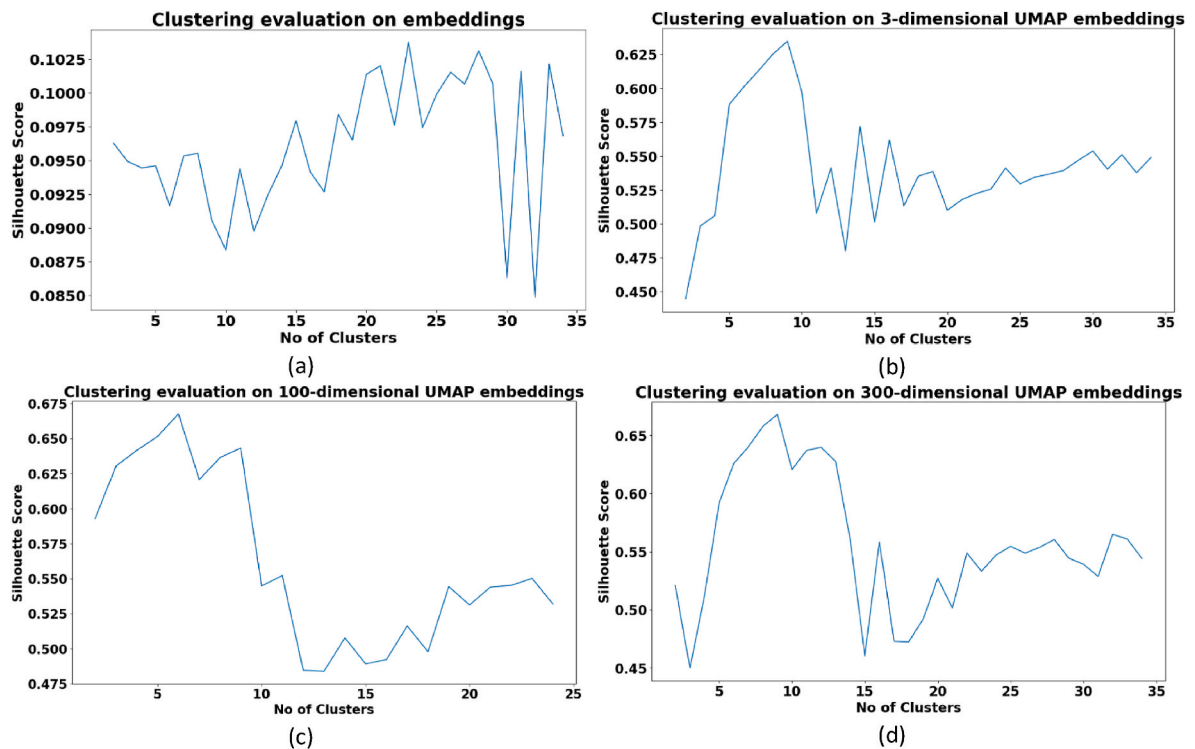


Fig. 3. K-means clustering evaluation using silhouette scores across different numbers of clusters: a) sentence embeddings without dimensionality reduction, showing consistently low silhouette scores (0.085–0.102) indicating poor cluster separation, b) 3-dimensional UMAP embeddings, demonstrating improved cluster quality with peak silhouette score (0.625) at 9 clusters, c) 100-dimensional UMAP embeddings, showing optimal clustering at 6 clusters (silhouette score 0.67), and d) 300-dimensional UMAP embeddings, also indicating effective clustering with peak performance at 9 clusters (silhouette score 0.65).

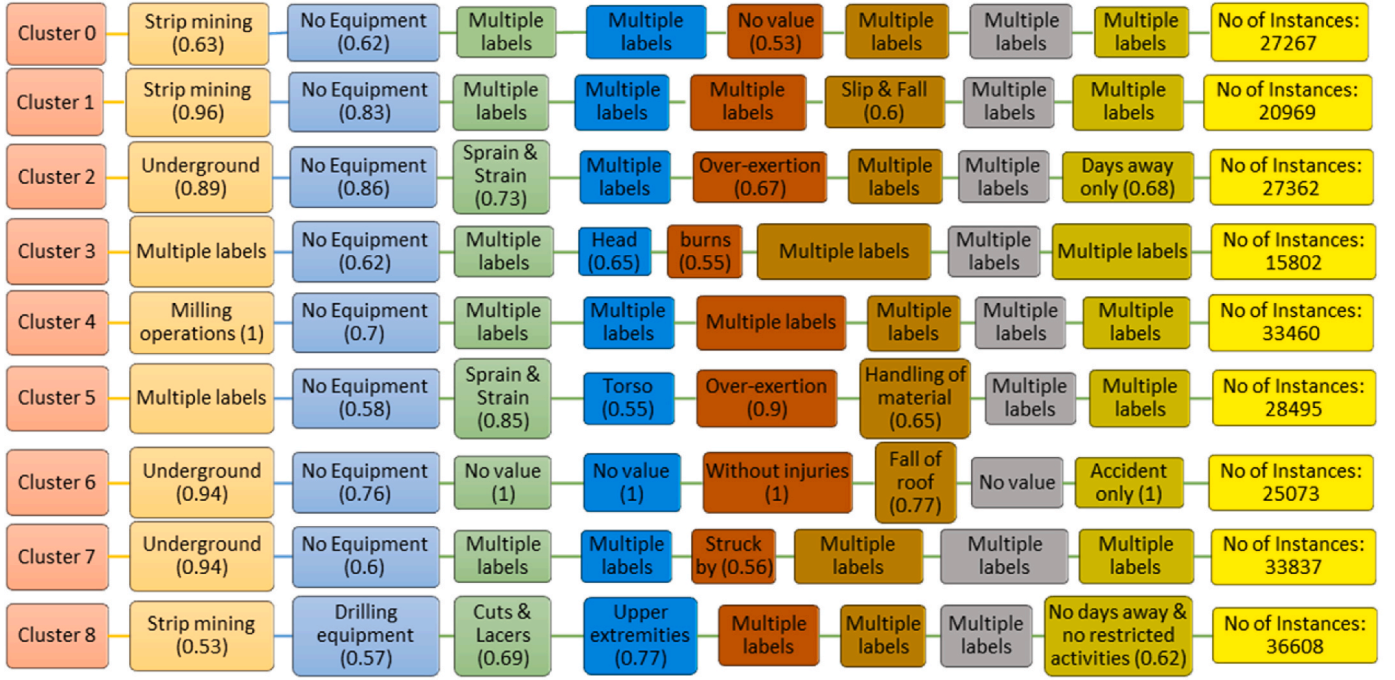
chosen to normalize local neighbor weights. UMAP then uses attractive and repulsive forces to evolve an equivalent weighted graph in low-dimensional space, with the attractive force given by:

$$\frac{-2ab \|y_i - y_j\|_2^{2(b-1)}}{1 + \|y_i - y_j\|_2^2} w(x_i, x_j) (y_i, y_j), \text{ and} \quad (2)$$

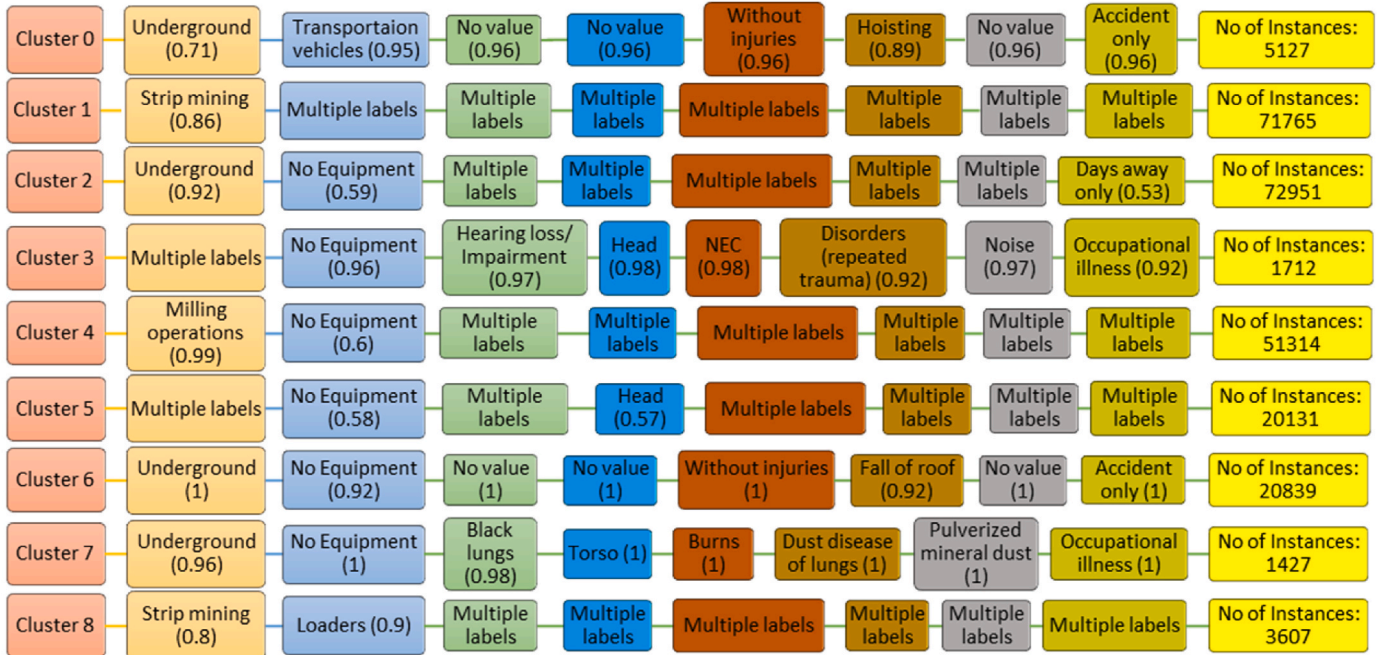
$$\frac{2b}{(\epsilon + \|y_i - y_j\|_2^2)(1 + \alpha y_i - y_j^{2b})} w(x_i, x_j) (y_i, y_j) \quad (3)$$

Where  $y_i$  and  $y_j$  are low-dimensional coordinates representing the projected positions of accident narratives in the reduced dimensional space, corresponding to the original high-dimensional SBERT embeddings  $x_i$  and  $x_j$  of two different accident narratives, and  $a$  and  $b$  are hyperparameters.

The performance of UMAP is significantly influenced by three key hyperparameters: output dimensionality ( $n\_components$ ), local neighborhood size ( $n\_neighbors$ ), and minimum distance between embedded points ( $min\_dist$ ). The  $n\_components$  parameter determines the dimensionality of the low-dimensional embeddings, balancing information

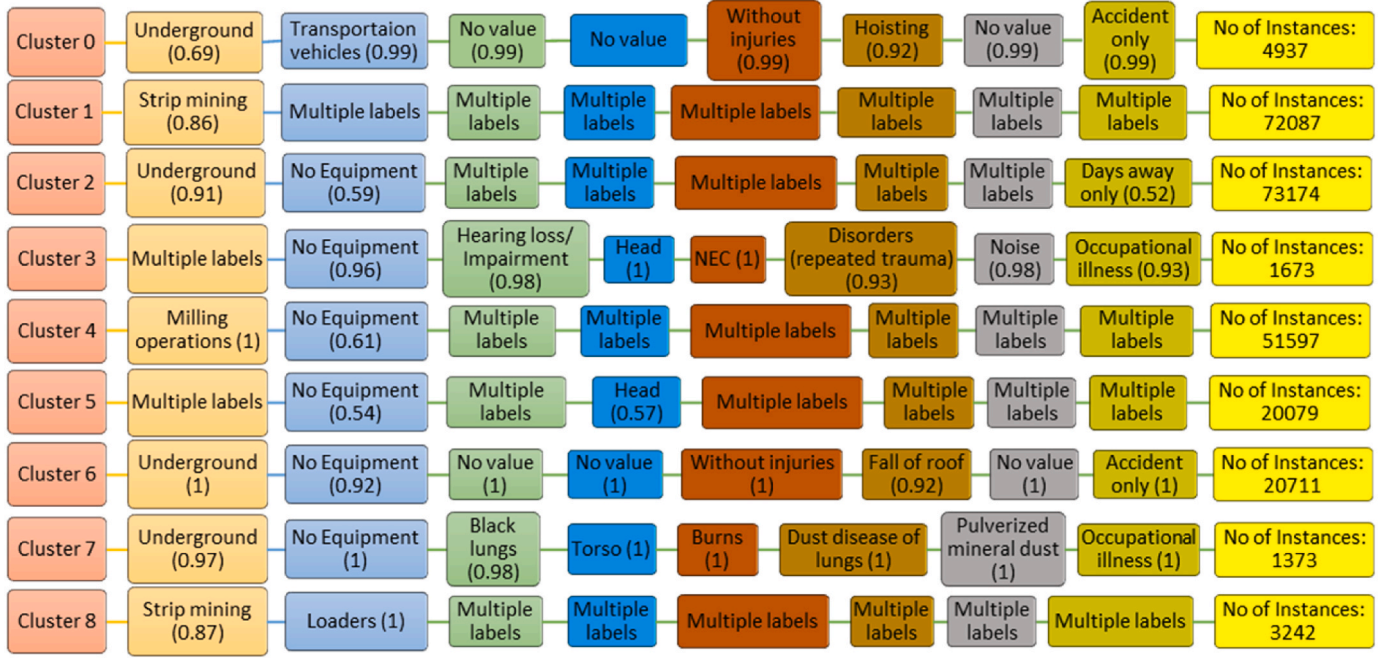


(a)

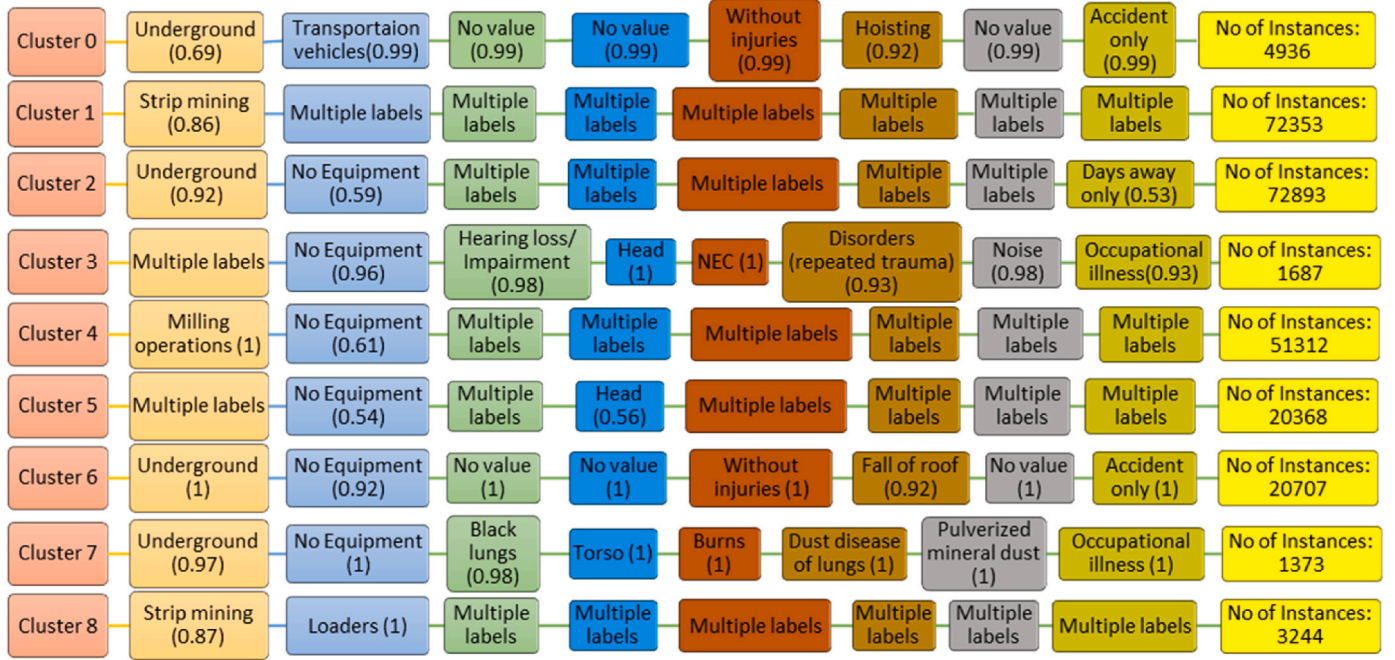


(b)

Fig. 4. k-means clustering analysis on merged narratives embeddings a) without dimensionality reduction, b) 3-dimensional UMAP embeddings, c) 100-dimensional UMAP embeddings, and d) 300-dimensional UMAP embeddings



(c)



(d)

### Variables

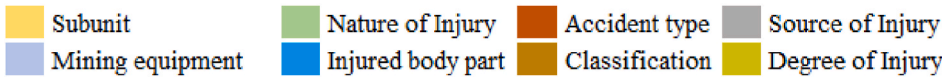


Fig. 4. (continued).

preservation with interpretability. The  $n\_neighbors$  parameter affects the balance between local and global structure preservation, with larger values capturing more global patterns at the cost of computational efficiency. The  $min\_dist$  parameter influences the spread of points in the embeddings, with smaller values promoting tighter clusters and larger values preserving global topology. To understand the detailed theory and mathematics behind the algorithm, please see (McInnes et al., 2018).

In the context of MSHA accident text data, these parameters play crucial roles. We reduce the 384-dimensional embeddings of text data, originally generated by the SBERT model, to 3 dimensions ( $n\_components = 3$ ) to enable effective visualization and clustering while preserving key semantic relationships. The  $n\_neighbors$  parameter is tuned to balance the capture of similarities between closely related incidents and broader patterns across different accident types. The  $min\_dist$  parameter is adjusted to achieve a clear separation between distinct



accident categories while maintaining the overall structure of relationships between different types of incidents. For this study, we optimize these hyperparameters using a trial-and-error method, experimenting with various values to achieve the most insightful clustering and visualization of the MSHA accident text data.

The main objective of applying UMAP in this study is to reduce the dimensionality of the SBERT-generated 384-dimensional accident narrative embeddings while preserving semantic relationships between accidents. This dimensionality reduction serves two primary goals: plotting low-dimensional representations for visual examination of data and facilitating clustering of similar accident reports. UMAP's ability to maintain both local and global structures in the data is crucial, as it helps reveal patterns and similarities across different types of accidents while preserving nuanced relationships within closely related incidents. By reducing the high-dimensional embeddings to a lower-dimensional space, we can visualize MSHA data variables for each cluster corresponding to text narratives, enabling comprehensive visual cluster analysis. These lower-dimensional representations, combined with other MSHA data variables, allow for detailed characterization of accident patterns within each identified cluster.

In this study, UMAP implementation is achieved using Python's UMAP library (McInnes et al., 2018). It is important to note that the visual examination is based on 3-D plots; however, for the presentation of results for readers, 2-D plots are presented in this article. To allow readers to explore the low-dimensional UMAP representations of the clustered data, a Jupyter notebook that generates these interactive 3D visualizations, and the underlying dataset are provided and can be accessed via the online repository (Danish).

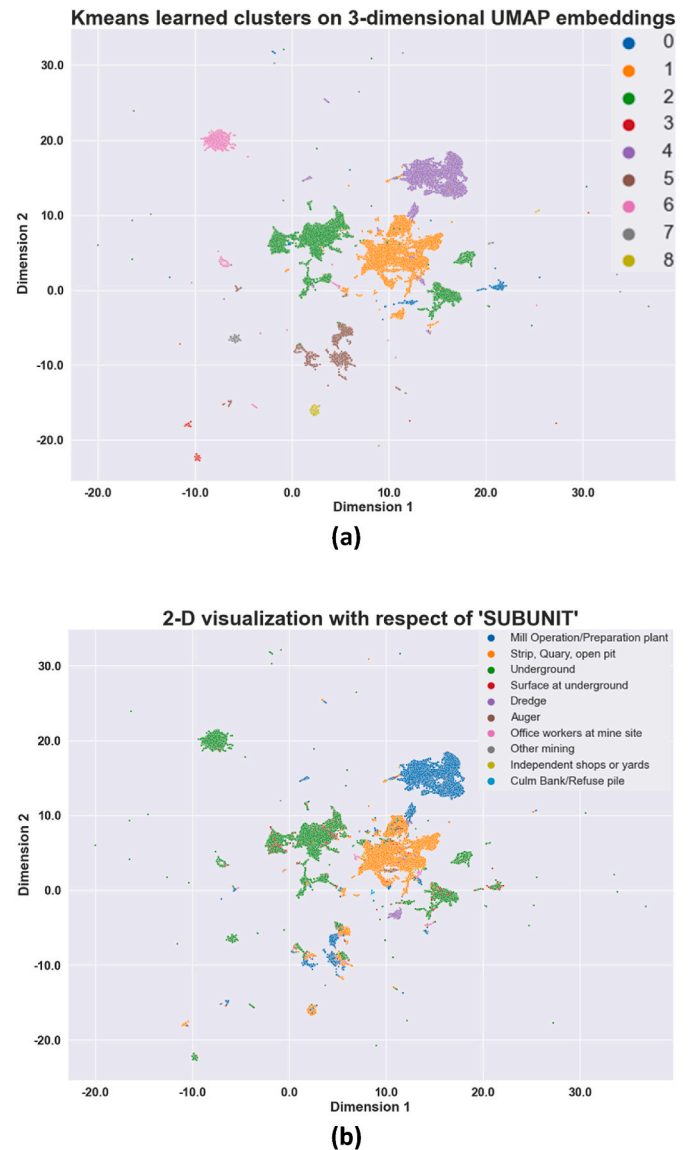
### 3.5. Clustering

Clustering is an unsupervised machine learning method used to identify the groups of similar objects within data, without having prior knowledge about the categories and their characteristics. The main goal of clustering algorithms is to attain higher intra-cluster homogeneity and inter-cluster heterogeneity, i.e., clustering data such that points within each cluster share similar properties compared to points of neighboring clusters. Hierarchical and non-hierarchical clustering are two commonly used techniques - where in hierarchical clustering, clusters are created in a predefined order, and similar clusters are grouped hierarchically. In contrast, in non-hierarchical clustering, clusters are formed by grouping or splitting the data based on evaluation criteria.

The main idea behind the cluster analysis conducted in this study is to group similar accident reports in such a way that each cluster represents accidents sharing common characteristics and patterns. The algorithm adopted was k-means clustering which is one of the most efficient non-hierarchical clustering algorithms for data segmentation due to its simplicity and computational efficiency. k-means clustering splits the given data into a predefined k number of clusters such that data points in every cluster have similar traits compared to points within other clusters. It starts by choosing random k-centroids assigning each data point to the closest centroid. The mean value of each cluster is calculated, and new clusters are formed considering the clusters' mean value as the new centroids. Similar steps are repeated until centroids no longer change their positions (Wang et al., 2019). The distance between each point and the centroid is measured using Cosine or Euclidean distance. The number of clusters, i.e., k, is predefined based on different metrics. For this study, the silhouette score is used to evaluate the learned cluster. One can refer (Baarsch and Celebi, 2012) for a detailed description of this method.

## 4. Results

This study analyzes 247,873 accident events from the MSHA dataset using the proposed methodology. Fig. 2 illustrates the effect of adding

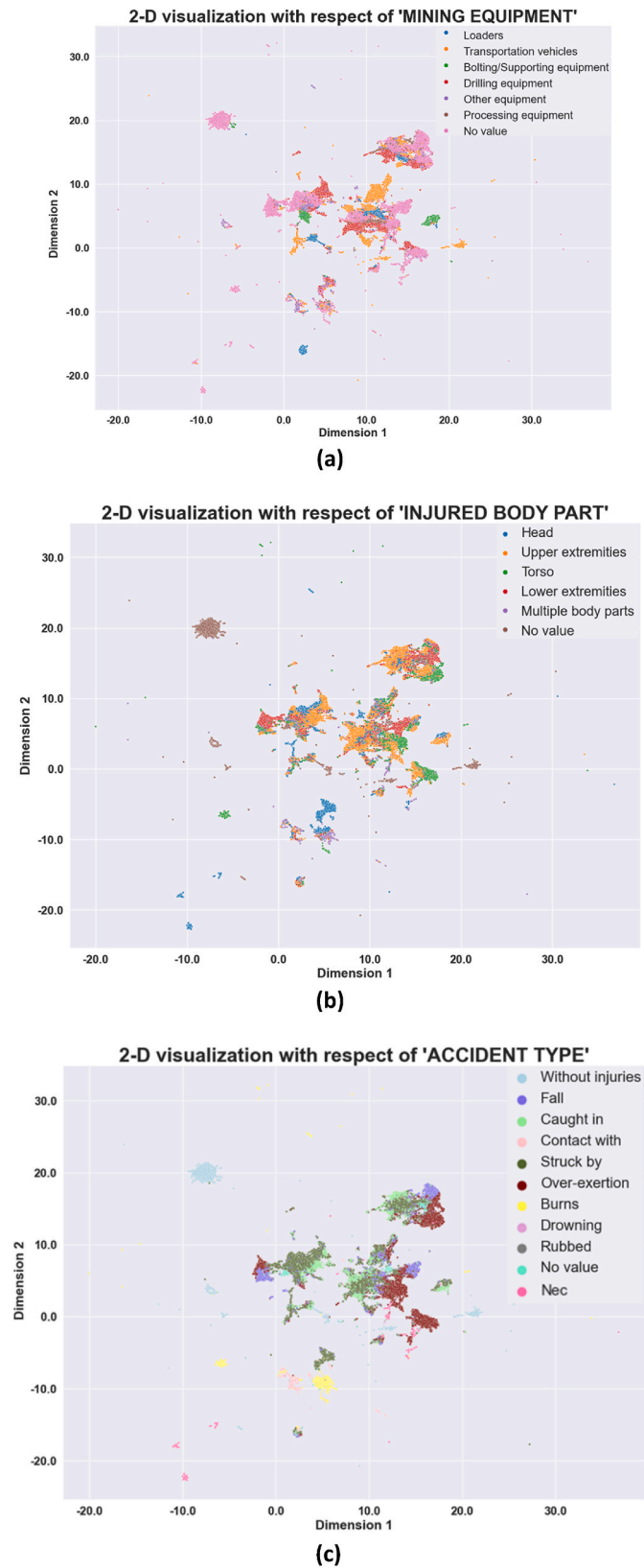


**Fig. 5.** 2-D visualization of 3-dimensional UMAP embeddings showing: a) nine distinct k-means learned clusters, and b) their correspondence with mining operation types (subunits), demonstrating that the clustering algorithm naturally grouped accidents based on operational contexts (e.g., underground, strip/quarry, mill operations).

text information from variables to the narratives on UMAP embeddings. The 3-dimensional UMAP representation of narratives-only embeddings, visualized in 2D in Fig. 2a, reveals six distinct clusters. However, most of the data points are concentrated in one large cluster, with the remaining clusters being relatively small, indicating limited differentiation between accidents when using narrative text alone. This clustering pattern could suggest that the narratives alone may not contain sufficient distinguishing information to effectively separate different types of accidents into distinct groups.

In contrast, the UMAP representation of merged narratives, visualized in 2D in Fig. 2b, shows more distinct and well-separated clusters compared to narratives-only embeddings. This improved structure indicates that incorporating additional text-based variables provides valuable context, enhancing the differentiation between accident types. The clearer cluster separation suggests that merged narratives capture more nuanced accident information, facilitating more effective grouping of similar incidents.

These low-dimensional UMAP embeddings of merged narratives are



**Fig. 6.** 2-D visualization of 3-dimensional UMAP embeddings demonstrating clustering of accidents by: a) mining equipment involvement, b) injured body parts, and c) accident types, showing how different accident characteristics are distributed in the embedding space.

used for clustering analysis to find the hidden patterns for accident narratives. However, clustering on low-dimensional UMAP representations may lead to the loss of essential data patterns or biased clustering

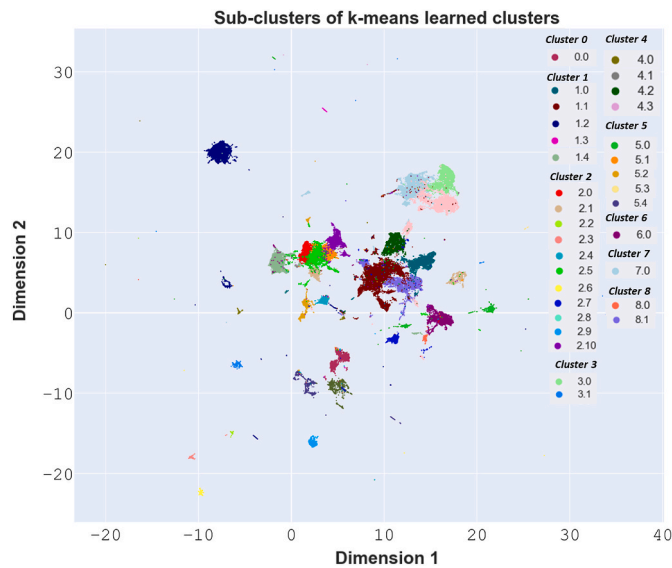
results if the low-dimensional UMAP components are not the accurate intrinsic dimensionality of the high-dimensional data. The intrinsic dimensionality of high-dimensional data depends on the characteristics



**Table 4**

Cluster 0, 6, and 7 dominating labels for each variable.

| Cluster # | Subunit     | Mining equipment      | Classification        | Injured body part     | Accident type    | Degree of Injury     | Nature of Injury      | Injury Source           |
|-----------|-------------|-----------------------|-----------------------|-----------------------|------------------|----------------------|-----------------------|-------------------------|
| 0         | Underground | Transportation        | Hoisting              | <sup>a</sup> No Value | Without injuries | Accidents only       | <sup>a</sup> No Value | <sup>a</sup> No Value   |
| 6         | Underground | <sup>a</sup> No Value | Fall of roof          | <sup>a</sup> No Value | Without injuries | Accidents only       | <sup>a</sup> No Value | <sup>a</sup> No Value   |
| 7         | Underground | <sup>a</sup> No Value | Dust disease of lungs | Torso                 | Burns            | Occupational illness | Black lungs           | Pulverized mineral dust |

<sup>a</sup> Entries marked as 'No value' indicate missing (Nan) label information for the corresponding variable.**Fig. 7.** Sub-clusters of each k-means learned cluster.

of the dataset. Therefore, it is important to find a sufficient number of low-dimensional UMAP representations of each data point for robust cluster analysis. In order to find the accurate low-dimensional representation of merged narrative embeddings, several numbers of low-dimensional UMAP components, i.e., 3, 100, and 300, are used for k-means clustering. Fig. 3a, b, and 3c illustrate the performance of k-means clustering on sentence embeddings without dimensionality reduction, and after applying UMAP to reduce the embeddings to 3, 100, and 300 dimensions, respectively. The silhouette score metric is used to find the optimal number of clusters for each aforementioned case. It can be seen in Fig. 3a that the k-means clustering on embeddings without dimensionality reduction does not provide a significant silhouette score for any number of clusters in high-dimensional space which is due to the high-dimensionality of data. In the case of 3-dimensional and 300-dimensional UMAP embeddings, the high silhouette score is achieved for 9 clusters. While the silhouette score suggested an optimal cluster number of 6 for the 100-dimensional UMAP embeddings, the difference in scores between 6 and 9 clusters was marginal. Consequently, 9 clusters were selected to maintain consistency for comparison across 3-dimensional, 100-dimensional, and 300-dimensional UMAP embeddings. This decision allows for a more comprehensive analysis of hazard categories across all three UMAP representations, potentially uncovering nuanced patterns in the accident data that might be obscured with fewer clusters.

Fig. 4a provides the cluster analysis on narrative embeddings without dimensionality reduction, whereas Fig. 4b, c, and 4d give the details of each k-means learned cluster on 3, 100, and 300 low-dimensional UMAP embeddings, respectively. In each figure, colored boxes represent different variables, with the text inside indicating the dominant label for that variable within the cluster, and accompanying

values in parentheses denote the proportion of accidents associated with the given label, ranging from 0 (0 %) to 1 (100 %). A label is considered dominant for a variable in a cluster when it accounts for more than 50 percent of the accidents within that cluster.

It is evident from Fig. 4b, c, and 4d that the k-means learned clusters on 3, 100, and 300 UMAP low-dimensional embeddings represent similar information in all three cases. This shows that the information is not lost with dimensionality reduction, as the k-means learned clusters on 300-dimensional UMAP embeddings represent similar information compared to the k-means learned clusters on 3 and 100-dimensional UMAP embeddings. However, the k-means learned clusters on embeddings without dimensionality reduction differ from all the above-mentioned cases, as seen in Fig. 4a. This is due to the curse of high dimensionality of data, and the effect can be seen in clustering, as most clusters don't have dominating labels associated with them in this case. Therefore, for this particular dataset, clustering on sentence embeddings without dimensionality reduction doesn't significantly extract the hidden patterns and clusters within the data, which would be helpful for the hidden pattern analysis. Moreover, it is evident that the number of low-dimensional UMAP embeddings has minimal effect on the clustering of data, so one can choose low-dimensional UMAP embeddings for clustering to reduce the computational complexity and mitigate the effect of the curse of dimensionality. Therefore, in this study, the further results and discussion are based on k-means clustering of 3-dimensional UMAP embeddings for accident pattern characterization.

#### 4.1. K-means cluster analysis on 3-dimensional UMAP embeddings

As indicated in the preceding section, the assessment of k-means clustering on 3-dimensional UMAP embeddings yielded 9 optimal clusters, and the 2-D graphical representation of these low-dimensional UMAP embeddings, color-coded by cluster numbers, is depicted in Fig. 5a. The preliminary cluster analysis of these k-means learned clusters on 3-dimensional UMAP embeddings shows that the clusters are spatially arranged based on the Subunit variable, which describes the mining location or type, such as strip/open-pit, underground, or mill, among others. This grouping of clusters is also evident in the 2-D visualization of low-dimensional UMAP embeddings of merged narratives, color-coded by the 'subunit' variable, shown in Fig. 5b. It is evident that accidents in underground mining operations are represented by clusters 0, 2, 6, and 7, whereas accidents in strip mining operations are represented by clusters 1 and 8. Cluster 4 is indicative of accidents in milling operations, while clusters 3 and 5 are associated with multiple mining activities, indicating that accidents in these clusters may involve complex or diverse factors.

However, a more in-depth examination of each cluster with respect to other variables in the MSHA dataset, including nature of injury, accident type, source of injury, mining equipment, injured body part, classification, and degree of injury, reveals that only clusters 0, 6, and 7 possess distinct and dominant labels related to all of these variables, as illustrated in Fig. 6 and Table 4.

Cluster 0 is predominantly associated with accidents related to underground mining operations, primarily involving transportation vehicles. The majority of incidents in this cluster are reported as "no-injury"



Fig. 8. K-means learned clusters along with their sub-clusters and dominating labels for each variable



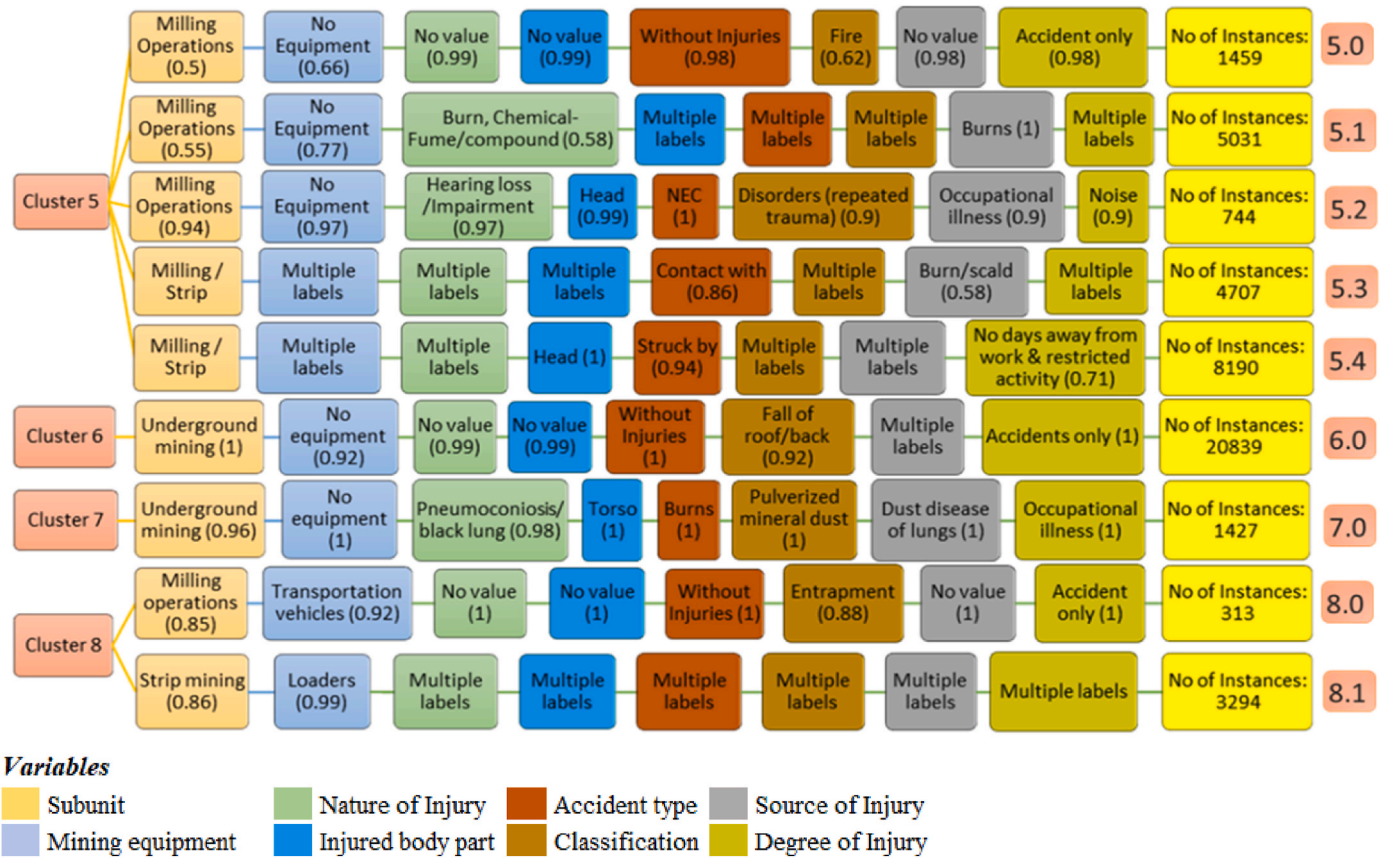


Fig. 8. (continued).

events and do not result in disruption to mining activities. Similarly, cluster 6 accidents also pertain to underground mining operations and are predominantly characterized by “no-injury” incidents and have no disruption to mining activities. However, these accidents are primarily associated with fall of the roof or back and are not typically linked to equipment use. Cluster 7 accidents are also associated with underground mining operations but are largely categorized as occupational health incidents rather than immediate safety hazards. The dominant incident type in this cluster is dust disease of the lungs, specifically black lung disease, caused by exposure to respirable coal dust, which falls under the category of occupational illness.

In contrast, clusters 1, 2, 3, 4, 5, and 8 exhibit a more diverse distribution of labels across key variables, necessitating a more nuanced analysis for effective safety hazard assessment. Identifying the predominant labels associated with critical variables such as mining equipment, accident type, and injured body part within these clusters is essential for addressing vital safety-related questions, such as a) what type of injuries are associated with specific mining equipment, b) what type of equipment is involved in accidents with respect to specific mining type, and c) the patterns of specific body part injuries in relation to particular accident types or mining equipment.

For instance, while cluster 1 encompasses accidents associated with strip/open-pit mining, it lacks a clear dominant pattern in other crucial variables, limiting its immediate utility for identifying specific safety hazards associated with strip/open-pit mining. Visual examination of these clusters, as illustrated in Fig. 6a, b, and 6c suggest that information related to mining equipment, injured body part, and accident type forms distinct sub-groups within each k-means derived cluster. Therefore, to extract more detailed insights from these complex clusters, it is necessary to identify and analyze these sub-clusters within these k-means learned clusters that lack dominant labels for these crucial variables. Accordingly, for this study, the criterion for sub-clustering is based on

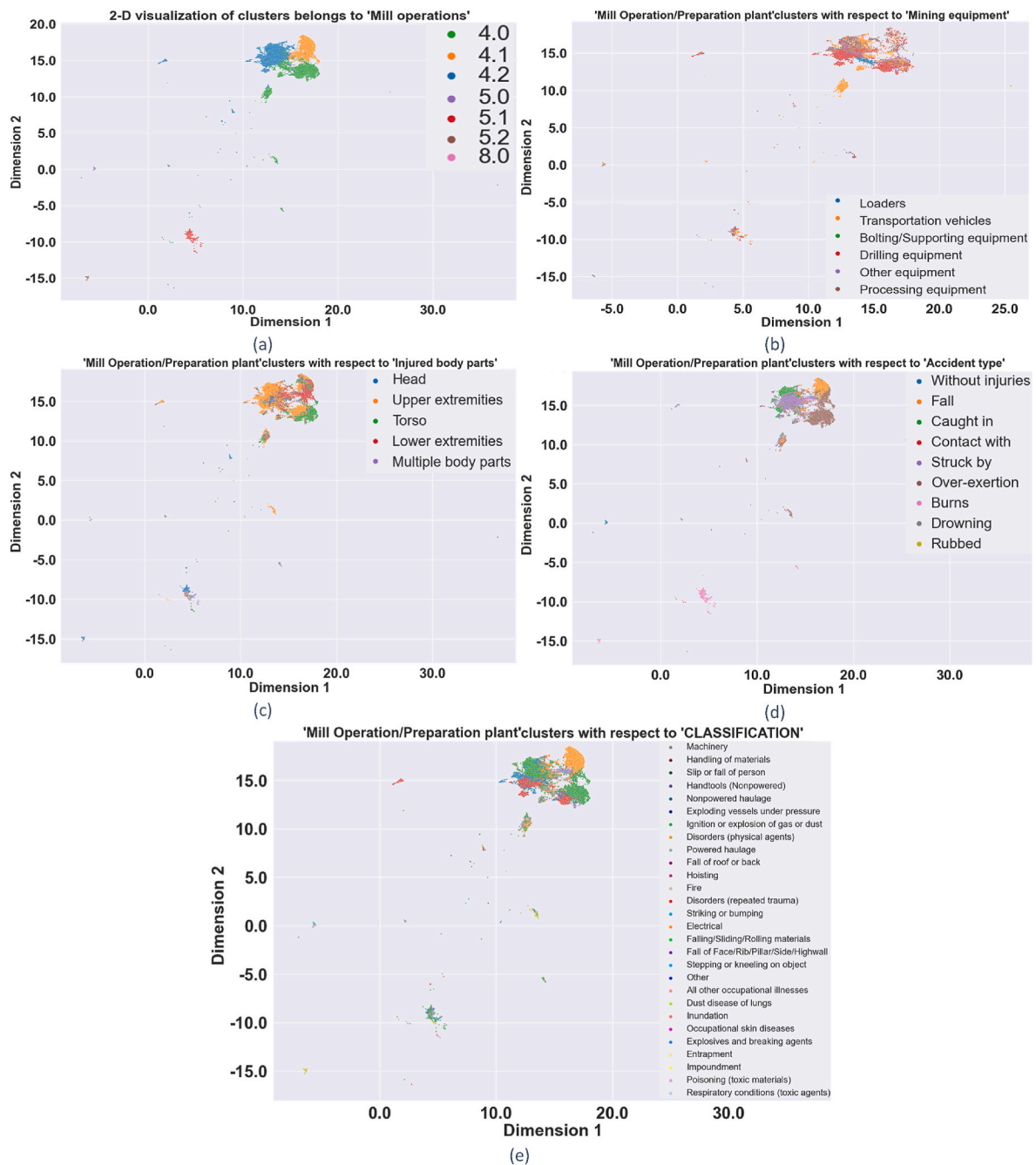
the missing dominant labels associated with these variables, including subunit, mining equipment, injured body part, and accident type. Hence, sub-clustering analysis using dendrogram and hierarchical agglomerative clustering techniques is conducted for clusters 1, 2, 3, 4, 5, and 8, aiming to reveal more granular patterns and associations within the accident data.

#### 4.2. Sub-clustering of k-means learned clusters

To further analyze the k-means learned clusters 1, 2, 3, 4, 5, and 8, post-processing is conducted with the application of dendrogram analysis and hierarchical agglomerative clustering. The dendrogram is a tree diagram illustrating the hierarchical relationship between data points and can be used to identify the number of clusters for given data. Thus, the agglomerative hierarchical method with ward linkage is used on k-means learned clusters to find sub-groups representing unique information related to the mining equipment, accident type, and injured body part variables for detailed safety insight. This approach aims to provide insights regarding injured body parts and the degree of injury associated with specific mining equipment, accident types, and mining methods. The heuristic approach was taken to find the number of sub-clusters for each k-means learned cluster by analyzing its relevant dendrograms manually. The 2-D visualization of the resulting sub-cluster and their respective clusters are presented in Fig. 7. Fig. 8 details each k-means learned cluster and its sub-clusters, each represented by its dominating label for each variable.

### 5. Discussion

As previously mentioned, the initial k-means learned clusters are separated in space based on the “SUBUNIT” variable. These clusters are further subdivided into sub-clusters through the application of



**Fig. 9.** 2-D visualization of clusters belonging to 'Milling/Preparation plant' operations, visualized with respect to (a) clusters number and variables; (b) Mining Equipment, (c) Injured body part, (c) Accident type, and (d) Classification.

agglomerative clustering techniques using dendrogram analysis. Through analysis of predominant variable labels in each of these sub-clusters, distinct accident patterns were characterized. The following section provides an in-depth discussion on all the clusters after sub-clustering.

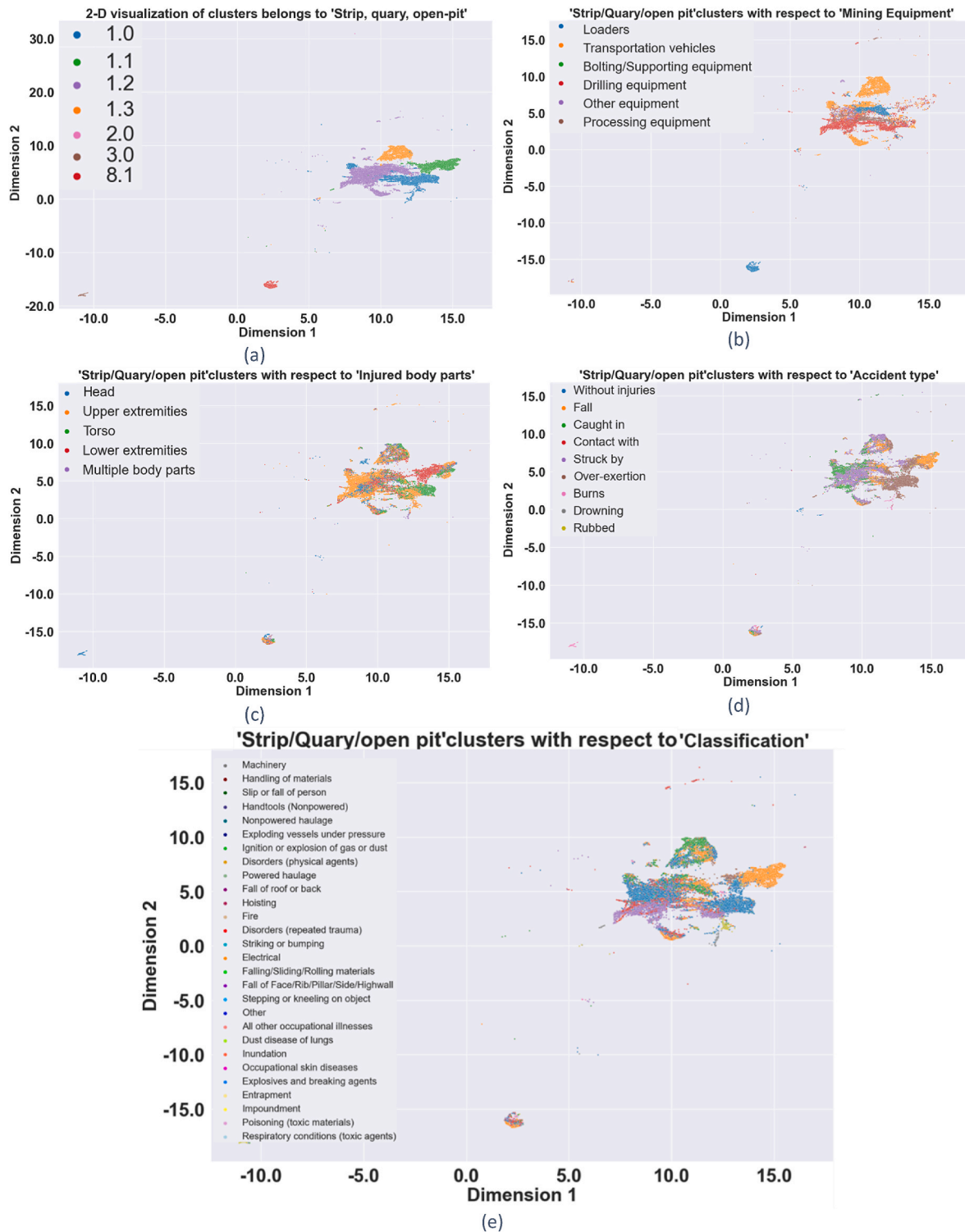
### 5.1. Clusters representing milling operations/preparation plants

The present study aimed to conduct clustering and sub-clustering in order to determine the group of instances that carry unique information related to accidents. Seven clusters (4.0, 4.1, 4.2, 5.0, 5.1, 5.2, and 8.0) were identified as belonging to milling operations. The 2-D visualizations of these clusters with respect to all variables are presented in Fig. 9. Each of the seven clusters presents unique information about accidents

in milling operations. A high proportion of instances within clusters 4.0, 4.1, 5.0, 5.1, and 5.2 have the 'No equipment' label dominant for the 'Mining Equipment' variable. The proportions of accidents with this dominant label in these clusters are 0.52, 0.95, 0.66, 0.77, and 0.97, respectively.

Clusters 4.0 and 4.1 indicate a predominance of sprain & strain injuries (0.65 and 0.55 respectively), with cluster 4.0 showing a higher incidence of torso injuries (0.5) due to over-exertion (0.75) and material handling (0.54). This likely reflects repetitive manual lifting or poor ergonomics in material transfer stations within milling operations (Horberry et al., 2013). These patterns suggest the need for ergonomic training, lifting aids, and task rotation to reduce physical strain (Apud, 2012). In contrast, cluster 4.1 exhibits lower extremity injuries (0.51) from slip and fall incidents (0.8). The high rate of "No equipment"

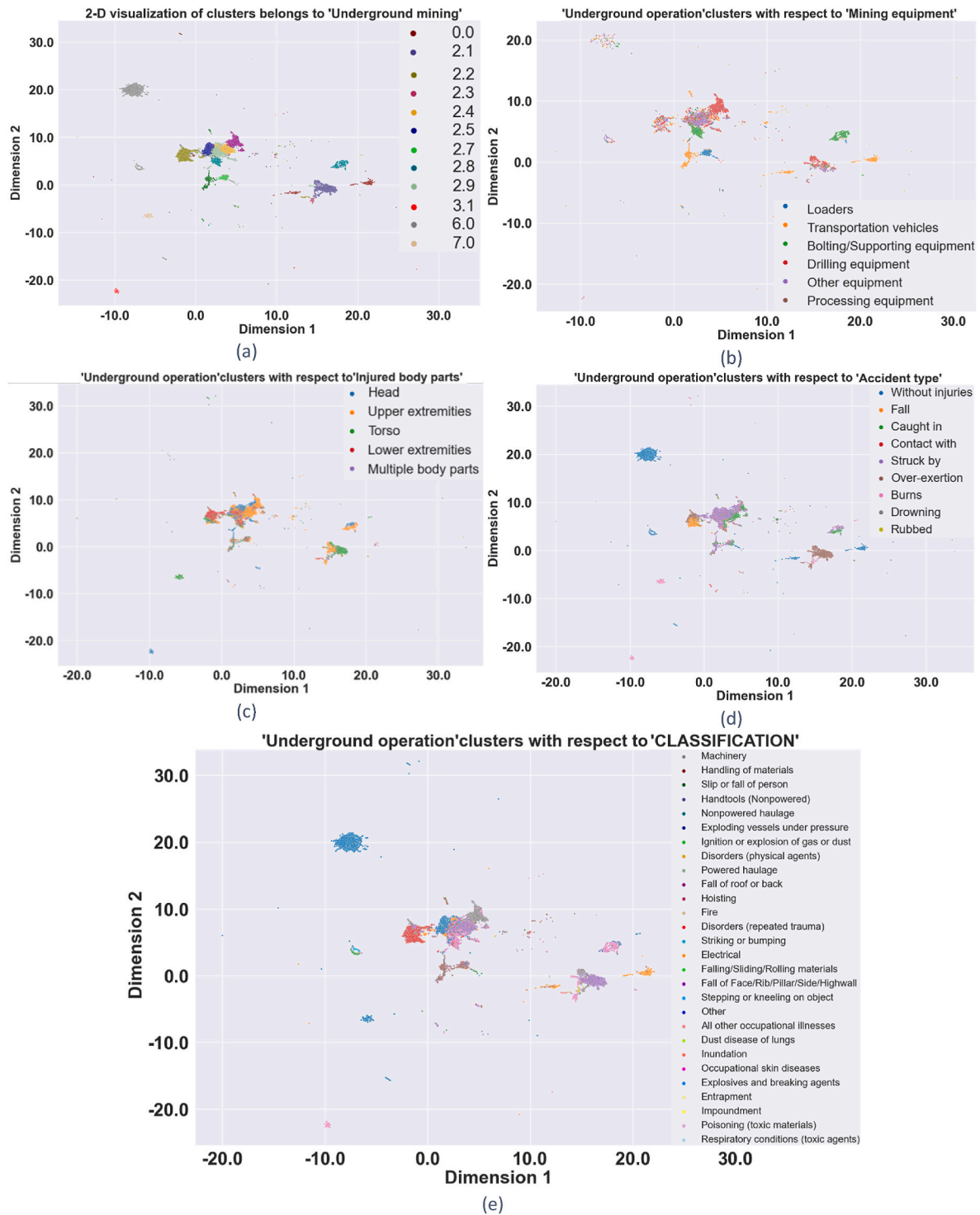




**Fig. 10.** 2-D visualization of clusters belonging to 'Strip/Quarry/Open pit' mining operations, visualized with respect to (a) clusters numbers and variables; (b) Mining Equipment, (c) Injured body part, (d) Accident type, and (e) Classification.

implies these are human-factors incidents, possibly linked to wet or uneven flooring, poor lighting, or inadequate footwear (Komljenovic et al., 2008). This highlights a potential gap in housekeeping practices or slip prevention programs. Cluster 4.2 is characterized by upper extremity injuries (0.69) from being struck by objects or vehicles (0.5). This cluster could represent workers operating in proximity to mobile equipment or handling heavy components without proper guards or spacing. These insights emphasize the importance of work zone

management, clear visual cues, and better equipment guarding in shared workspaces (Stemn, 2019). Notably, clusters 5.0 and 8.0 predominantly involve no-injury accidents (0.98 and 1.0), yet they point to high-risk events. Cluster 5.0 is primarily associated with fire-related incidents (62 %), suggesting operational conditions where combustible materials or heat sources are present, but the events often result in near misses (Hansen, 2018). Fire hazards remain a critical cause of severe accidents in mining operations; therefore, implementing robust prevention and



**Fig. 11.** 2-D visualization of clusters belonging to 'Underground' mining operations, visualized with respect to (a) clusters numbers and variables; (b) Mining Equipment, (c) Injured body part, (d) Accident type, and (e) Classification.

mitigation strategies is essential to minimize the occurrence and impact of such incidents. Cluster 8.0 was characterized by transportation vehicle interactions (0.92), which may reflect incidents involving collision risks or loss of vehicle control, but again without reported injuries (Hrica et al., 2023). These findings reinforce the need for incident follow-up and proactive hazard mitigation, even for no-injury events. Cluster 5.1 is exclusively burns-related (1.0), while 5.2 focuses on hearing loss and impairment (0.97) due to noise exposure (0.9). These clusters likely reflect chronic hazard exposure in specialized areas, such

as welding stations or high decibel grinding operations. These insights point to a need for routine monitoring, PPE compliance audits, and possibly engineering controls to mitigate long-term health impacts (Onder et al., 2012).

## 5.2. Clusters representing strip/quarry/open pit mining operations

Further analysis revealed clusters 1.0, 1.1, 1.2, 1.3, 2.0, 3.0, and 8.1 belong to strip/quarry/open pit mining. The 2-D visualization of these

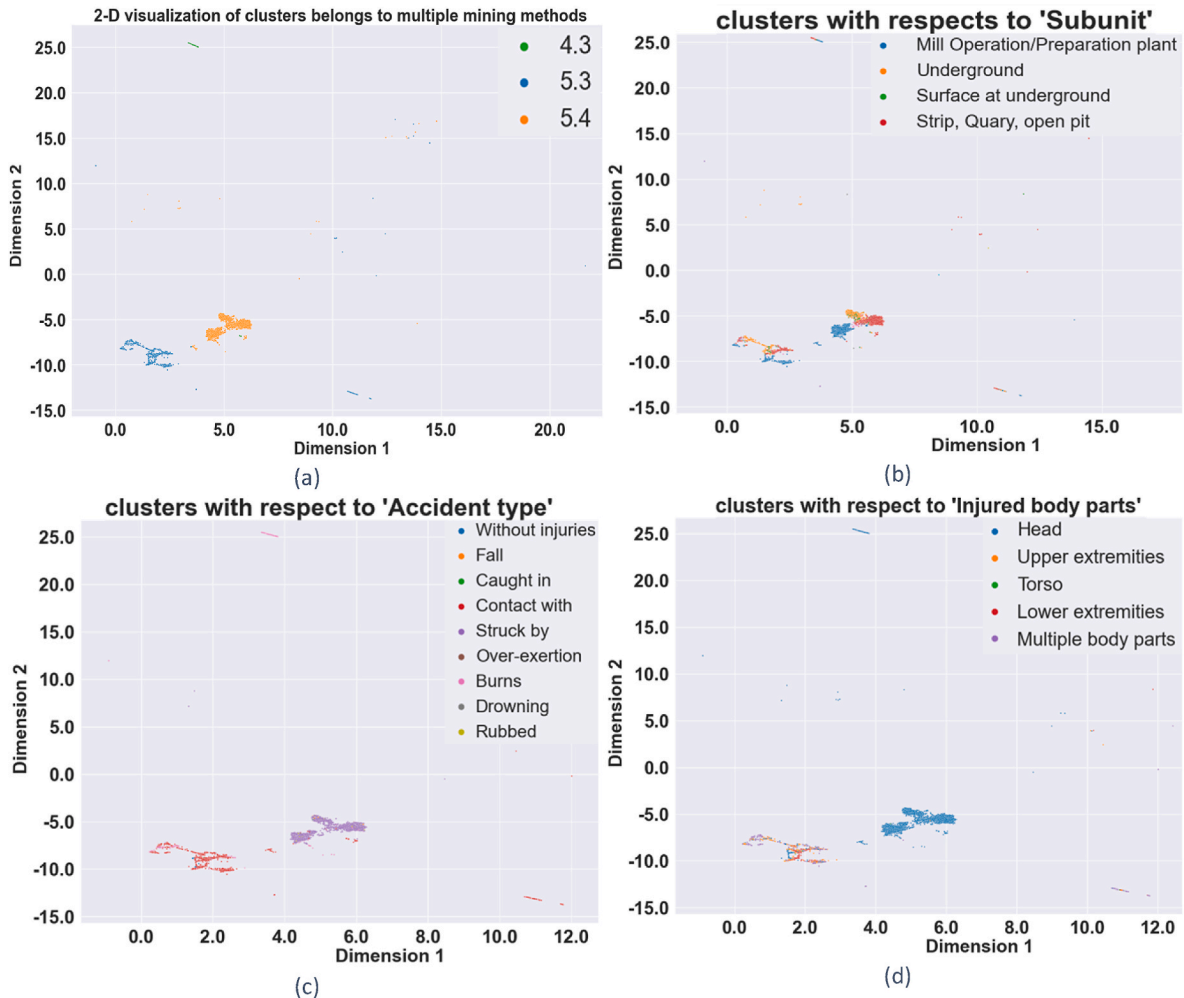
**Table 5**

Clusters share the same information and differentiate based on the mining operation.

| Clusters No | Sub-Cluster No | Differentiating variables and labels | Common and dominating labels                                                                                  |
|-------------|----------------|--------------------------------------|---------------------------------------------------------------------------------------------------------------|
| 1           | 1.0            | Strip/open pit mining                | No equipment, Handling of Material, Torso, over-exertion, Strain & Sprain                                     |
| 2           | 2.1            | Underground mining                   |                                                                                                               |
| 4           | 4.0            | Milling operation                    |                                                                                                               |
| 1           | 1.1            | Strip/open pit mining                | No equipment, Slip & Fall, Lower extremities, Fall, Strain & Sprain                                           |
| 2           | 2.2            | Underground mining                   |                                                                                                               |
| 4           | 4.1            | Milling operations                   |                                                                                                               |
| 0           | 0.0            | Strip mining                         | No equipment, Torso, Dust disease of lungs, Occupational illness, Pneumoconiosis/black lungs, Pulverized dust |
| 7           | 7.0            | Milling operations                   |                                                                                                               |
| 3           | 3.0            | Strip/open pit mining                | No equipment, disorder (repeated trauma), Head, Burns-NEC, Occupational illness, Hearing Loss, Noise          |
| 3           | 3.1            | Underground mining                   |                                                                                                               |
| 5           | 5.2            | Milling operations                   |                                                                                                               |

clusters in relation to all variables is depicted in Fig. 10. It was observed that accidents belonging to Clusters 1.0 and 1.1 primarily involve no equipment (0.98 and 0.96 respectively) and share similarities with

clusters 4.0 and 4.1 in terms of the predominant variables, i.e., Classification, Accident type, Injured body parts, and Nature of injury, except for the Subunit, which differentiates these clusters. This pattern suggests that injuries related to manual labor and human error, rather than equipment failure, are prevalent across both milling and open-pit mining contexts (Simpson and Horberry, 2009). These findings highlight the importance of behavioral safety interventions and ergonomic practices in physically demanding tasks (Kowalski et al., 2011). Cluster 1.3 primarily involves transportation vehicles (0.98) and shows a higher proportion of injuries to multiple body parts. This may indicate more severe or complex accidents, such as vehicle rollovers or collisions, where multiple impact points are involved (Digges and Eigen, 2004). These findings revealed the need for enhanced training in vehicle operation, and stricter traffic management on mine sites. Cluster 8.1 has loaders as the dominant 'Equipment Type' (0.99), suggesting a higher frequency of loader-related accidents affecting multiple body parts. This could be due to operational or visibility challenges during loading and dumping operations (Kecojevic and Radomsky, 2004). This finding demonstrated the need for better machine operator training, improved line-of-sight, and advanced collision avoidance systems to prevent multi-point injuries (Bazaluk et al., 2023). Cluster 3.0, similar to cluster 5.2, shows a higher proportion of hearing loss and impairment (0.96) due to noise, where no equipment (0.98) is directly involved in causing injuries. This likely reflects long-term occupational exposure to ambient industrial noise. The results emphasize the importance of consistent hearing conservation programs, noise monitoring, and use of PPE, especially in tasks



**Fig. 12.** 2-D visualization of clusters representing multiple mining operations with respect to; (a) clusters numbers, (b) subunit, (c) accident type, and (d) injured body part.

that may not involve active equipment (Onder et al., 2012). Cluster 1.2, like cluster 4.2 shows a higher frequency of injuries to the upper extremities (0.52) due to being struck by objects or vehicles (0.5). This suggests frequent exposure to moving machinery or falling material in open-pit mining environments (Souza et al., 2024). Interventions such as better guarding, proper material handling protocols, and hazard zone demarcation could potentially mitigate these risks. Cluster 2.0 uniquely represents dust disease of the lungs, i.e., black lungs caused by coal dust (1.0). This cluster captures chronic occupational health risks rather than acute injury events, indicating the importance of long-term respiratory health monitoring, airborne dust control technologies, and stricter exposure limits in open-pit coal mining operations (Colinet, 2020).

### 5.3. Clusters representing underground mining operations

The analysis further revealed that clusters 0.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 2.10, 3.1, 6.0, and 7.0 belong to underground mining operations. The 2-D visualizations of these clusters are visualized in relation to all variables for the purpose of visual representation, as depicted in Fig. 11. It is observed that clusters 2.1, 2.2, 2.3, 2.4, 2.5, 3.1, 6.0, and 7.0 primarily involve accidents that do not involve equipment, and each cluster provides distinct information pertaining to accidents in underground mining operations. The proportions of accidents with a 'No Equipment' label in these clusters are 0.84, 0.96, 0.97, 0.94, 0.96, 0.98, 0.92, and 0.96 respectively. This suggests that a significant portion of underground mining accidents arise from manual tasks or environmental conditions, underscoring the importance of human-factor interventions and proactive workplace design in mitigating risks (Bahn, 2013).

Clusters 2.1 and 2.2 are characterized by sprain and strain injuries (0.5 and 0.6, respectively), resulting in days away from work (0.71 and 0.66). Cluster 2.1 emphasizes torso injuries (0.67) primarily from over-exertion (0.96), while cluster 2.2 highlights lower extremity injuries (0.58) predominantly from slips and falls (0.74). These clusters share similarities with clusters 1.0, 4.0, and 1.1, 4.1, respectively, differentiated primarily by mining location, i.e., subunit. The confined and uneven terrain in underground mines likely intensifies these risks, highlighting a need for enhanced ergonomic practices, physical task rotation, and improved flooring or traction measures (Jones and James, 2018).

Cluster 2.3 represents joint/tendon/muscle inflammation injuries (0.98) to the upper extremities (0.95) caused by over-exertion (0.98). This may point to task-specific ergonomic deficiencies such as improper tool design, repetitive manual motions, awkward posture, and issues often under-monitored in underground settings. Targeted interventions like redesigned tools, task redesign, and muscle fatigue monitoring could reduce these injuries (Butlewski et al., 2015). Cluster 2.4 is associated with material handling (0.89), causing cut and laceration injuries (0.65) to upper extremities (0.85), typically resulting in no days away from work and restricted activity (0.63). The low severity, yet high frequency, accidents suggest that these are routine tasks with recurring minor risks. Safety gloves, safe materials handling protocols, and task automation could reduce the volume of such injuries (Stewart et al., 2001). Cluster 2.5 predominantly involves employees being struck by (0.97) falling rock due to roof or back collapse (0.72), often resulting in

days away from work (0.56). This cluster reflects one of the most hazardous underground scenarios and highlights the urgent need for consistent ground support systems, real-time monitoring for rock movement, and better training on recognizing ground control hazards (Maiti and V Khanzode, 2009).

Clusters 2.6, 2.7, and 2.8 accidents are equipment related. Cluster 2.6 accidents are related to drilling equipment (0.92) involving hand tools (non-powered) (0.89), causing cut and laceration wounds (0.5) to upper extremities (0.68) due to being struck by objects or equipment (0.57), leading to no day away from work and no restricted activity (0.68). The result suggests lower-risk but high-frequency exposures, needing better training on manual tool use and protective gear usage (Arciniega-Rocha et al., 2023). Cluster 2.7 accidents involve loaders (0.98) in powered haulage (0.67), often causing days away from work (0.58) due to being struck by equipment (0.53). The result indicates operational challenges in narrow haulage pathways and visibility issues in underground mining. A better traffic control, collision avoidance technologies, and clear operator-pedestrian separation could potentially minimize these accidents (Moniri-Morad et al., 2024). Cluster 2.8 involves supporting and bolting equipment (0.96), which causes injuries to the upper extremities (0.56) and results in days away from work (0.5), mostly associated with machinery (0.72). The result revealed that high-risk maintenance and reinforcement tasks require specialized operator training, automated bolting systems, and better ergonomics (Papic and Kovacevic, 2016).

Cluster 3.1 accidents are related to disorder (repeated trauma) (0.95) caused by repeated exposure to high levels of noise leading to hearing loss and impairment (0.97), further affirming the widespread challenge of occupational noise in mining. A sustainable hearing conservation program and regular audiometric testing are key to minimizing this occupational injury and illness (Sun et al., 2019). Clusters 0.0 and 2.10 accidents are associated with transportation vehicles (0.95 and 0.99). Cluster 0.0 accidents, categorized under hoisting (0.88), predominantly result in no injuries (0.95). Cluster 2.10 accidents, categorized under power haulage (0.5), involve being struck by transportation equipment (0.66), often leading to days away from work (0.63). These highlight the contrasting aspects of underground transportation, low frequency but high-severity outcomes, and emphasize the continued need for rigorous traffic management, safety interlocks, and regular vehicle inspections (Burgess-Limerick, 2011).

### 5.4. Clusters representing dredge mining operations

Cluster 1.4 represents accidents that are associated with dredge mining (0.9). The analysis revealed that there is no dominant label for any variable in this cluster except the mining equipment, indicating that there are multiple causes of accidents, including over-exertion, struck by, and falls, mostly involving no equipment (0.58). A significant proportion of accidents are associated with handling material, slip & fall incidents, and hand tools (non-powered). This diversity in accident types suggests that dredge mining activities present a wide range of potential hazards, with the potential for injuries to various body parts (Mohapatra et al., 2017). Additionally, a higher proportion of accidents in dredge mines lead to no days away from work and restricted activity, implying that while the incidents are frequent, their severity is generally low. This may reflect better hazard awareness or prompt response mechanisms, but it could also indicate underreporting of minor injuries in this work setting.

### 5.5. Clusters representing similar information but distinguishable based on the mining method

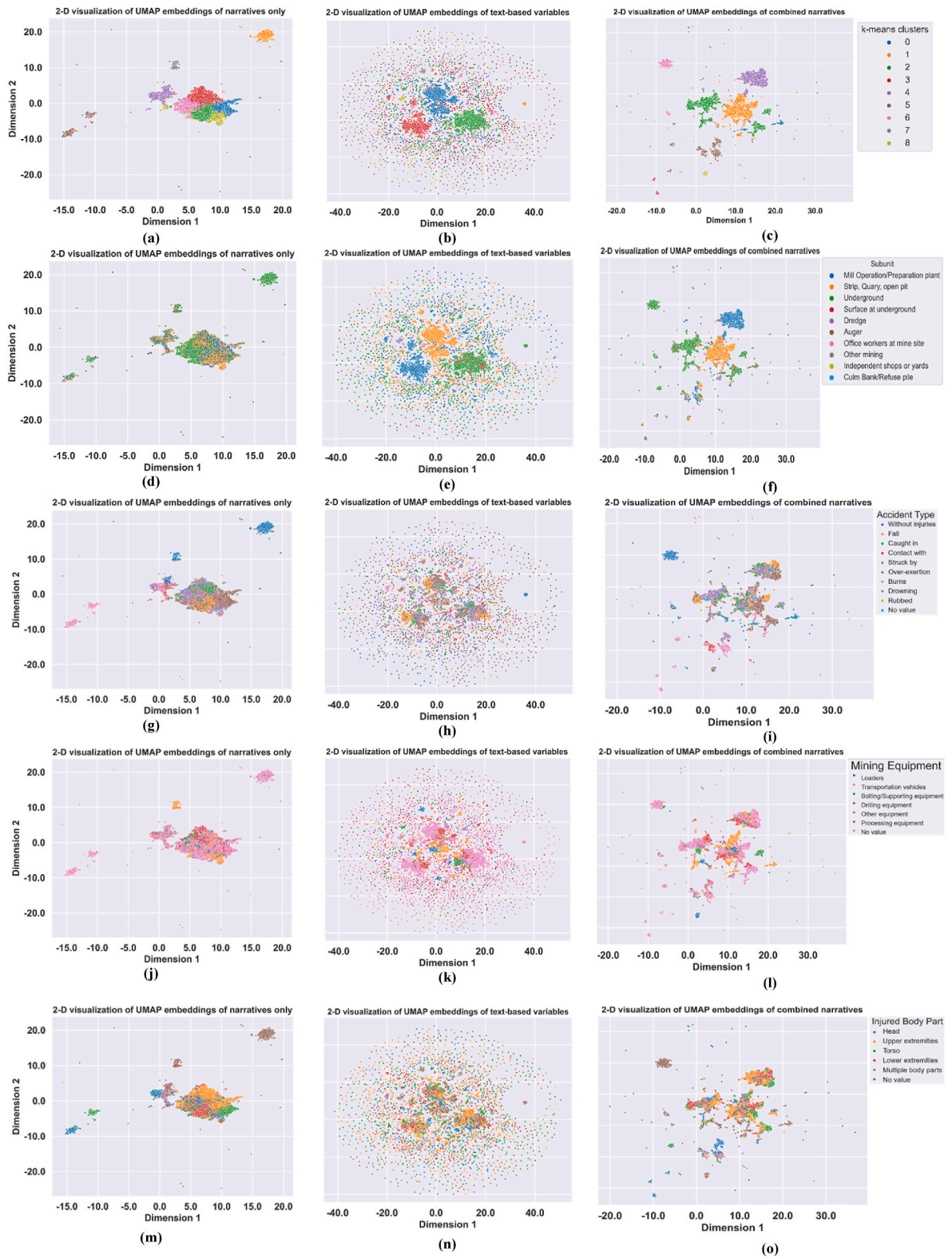
Several clusters exhibit similar accident characteristics but are differentiated by mining method. Clusters 1.0, 2.1, and 4.0 present comparable information regarding strain and sprain-related injuries (0.56, 0.5, and 0.65, respectively) to the torso (0.51, 0.67, and 0.5) due

**Table 6**

Performance comparison of traditional methods with proposed methodology for accident narrative clustering.

| Method                | Maximum Sillhouette Score | Optimum Number of Clusters |
|-----------------------|---------------------------|----------------------------|
| TF-IDF + PCA + kmeans | 0.476                     | 2                          |
| GloVe + PCA + kmeans  | 0.309                     | 4                          |
| BERT + PCA + kmeans   | 0.394                     | 2                          |
| SBERT + UMAP + kmeans | 0.625                     | 9                          |





**Fig. 13.** 2-D visualization of 3-dimensional UMAP embeddings of three distinct cases: narratives only (left column), text-based variables only (middle column), and combined narratives with text-based variables (right column), color-coded with respect to k-means clusters (a, b, c), Subunit (d, e, f), Accident type (g, h, i), Mining Equipment (j, k, l), and Injured body part (m, n, o). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

to over-exertion (0.68, 0.96, and 0.75) while handling materials (0.5, 0.82, and 0.54). These clusters are separated into different groups as they belong to separate mining subunit, i.e., strip mining (0.89),

underground mining (0.9), and milling operations (0.99), respectively. This differentiation highlights how workplace ergonomics and task demands differ by mining method, even when the resulting injuries are

physiologically similar. For instance, over-exertion in milling may result from repetitive handling in controlled indoor environments, whereas in strip mining it could contribute from large-scale material movement in uneven terrains. Table 5 provides the details of all the identical clusters based on the dominating variables, i.e., mining equipment, classification, injured body parts, accident type, and nature of injury - differentiated by the type of mining subunit.

Clusters 1.1, 2.2, and 4.1 illustrate strain and sprain-related injuries (0.51, 0.6, and 0.55) to lower extremities (0.5, 0.58, and 0.51) due to slip and fall incidents (0.79, 0.74, and 0.8) in strip mining (0.93), underground mining (0.87), and milling operations (1), respectively. While the injury types and mechanisms are comparable across these clusters, the underlying environmental conditions, wet surfaces in milling, confined underground passages, or unstable quarry slopes, likely contribute to subtle differences in incident dynamics and prevention strategies (Nowrouzi-Kia et al., 2018).

Clusters 0.0 and 7.0 both represent occupational hazards related to dust disease of lungs (1.0 and 1.0) caused by exposure to pulverize mineral dust (1.0 and 0.98). These clusters span across different subunits but are united by exposure to fine particulate matter, pointing to a systemic occupational health risk across mining contexts where dust mitigation remains inadequate (Singh et al., 2025). Similarly, clusters 3.0, 3.1, and 5.2 present occupational hazards associated with noise-induced hearing loss and impairment (0.97 each) in strip mining (0.96), underground mining (0.97), and milling operations (0.97), respectively. The persistence of such occupational illnesses across subunits highlights the pervasive nature of long-term exposure risks, often overlooked in favor of more immediate physical injuries (Ntlhakana et al., 2020).

#### 5.6. Clusters with multiple mining methods clustered together

Clusters 4.3, 5.3, and 5.4 exhibit distinct behavior, as they are not separated based on mining subunit but rather on other dominant features, primarily Accident type, as can be seen in Fig. 12. These clusters have no dominating label for mining subunit; however, most instances belong to strip mining and milling operations in all three clusters. While most instances within these clusters are drawn from strip mining and milling operations, the clustering is driven by commonalities in injury profiles and sources.

Cluster 4.3 involves burn-related injuries (0.5), specifically electric arc burns categorized as 'Flash burns,' primarily affecting the eyes (0.97) and often associated with welding machines (0.95). The characteristics of this cluster highlight very specific but serious hazard in mining environments where maintenance, fabrication, or repair work involving welding is common. The concentration of eye injuries due to arc exposure signals the need for targeted safety interventions, such as enhanced protective eyewear protocols and improved arc flash shielding (Ezisi, 2019). Cluster 5.3 is characterized by contact burns or scalds (0.58) affecting various body parts, mainly caused by contact with hot objects and heat (0.86). This pattern suggests generalized thermal hazards, possibly stemming from equipment surfaces, pipelines, or processing units, especially in milling environments (Xiaojie et al., 2011). Cluster 5.3 differentiates from cluster 4.3 in terms of the type of burns and injured body parts. Cluster 5.3 injuries are mainly causing burns or scalds injuries to all body parts. In contrast, cluster 5.3 injuries are primarily associated with flash burns to the eyes and sources mostly from welding activities, reflecting more diffuse heat-related risks, implying a broader need for hazard identification and preventive equipment maintenance (Zaitsev et al., 2024). Cluster 5.4 is grouped based on eye injuries (1.0) similar to cluster 4.3; however, these injuries result from being struck by equipment (0.94) and do not involve burns. This separation from 4.3 suggests an alternate risk scenario, potentially involving flying debris, tool recoil, or operational mishaps during equipment handling. Although the affected body part (eye) overlaps with Cluster 4.3, the injury mechanism is different, illustrating how

similar injury outcomes can arise from distinct operational conditions. This reinforces the importance of multifactorial hazard assessments when designing safety interventions (Wu et al., 2022).

#### 5.7. Comparative analysis with traditional methods

To evaluate the effectiveness of the proposed SBERT-UMAP framework, we conducted a comparative analysis against traditional text representation and dimensionality reduction methods using consistent evaluation criteria, applying all methods to the same merged accident narratives to ensure fair comparison. The comparative baselines included TF-IDF + PCA, GloVe + PCA, and BERT + PCA. The quantitative evidence for this enhanced performance is presented through silhouette scores, a critical metric for evaluating both cluster cohesion and the degree of separation between clusters in a given embedding space.

The proposed SBERT-UMAP framework demonstrated superior performance, particularly in its ability to delineate clearer semantic trends and achieve more distinct cluster separation within the high-dimensional representational space. This efficacy stands in marked contrast to the comparative methods. TF-IDF + PCA achieved a maximum silhouette score of 0.476 (at 2 clusters), GloVe + PCA yielded 0.309 (at 4 clusters), and BERT + PCA achieved 0.394 (at 2 clusters). In contrast, the SBERT-UMAP framework achieved significantly higher scores of 0.625 for 3-dimensional embeddings (at 9 clusters), as presented in Table 6. For these comparative methods, the limitations in capturing nuanced semantic relationships (as with TFIDF and GloVe), or the constraints imposed by PCA's linear projection on richer embeddings like BERT, often resulted in a less interpretable and more conflated semantic space from the unstructured narrative data. The consistently lower silhouette scores indicate clusters that are less cohesive internally and poorly separated from each other, often leading to a scenario where a significant portion of the data points are clustered together without meaningful differentiation, thereby hindering the clear identification and distinction of accident categories within their respective data representations. This compelling performance underscores the SBERT-UMAP framework's enhanced capacity to produce clusters that are both internally robust and distinctly separated, leveraging SBERT's rich sentence-level semantic embeddings and UMAP's non-linear dimensionality reduction to create a more discriminative semantic space for accident narrative analysis.

#### 5.8. Impact of merging text-based variables with accident narratives

In this study, the effectiveness of merging text-based variables with accident narratives was evaluated through a comparative analysis of clustering results. This process involved applying the same methodology to three distinct cases: narratives only, text-based variables only, and combined narratives with text-based variables. The resulting 2-D visualizations of UMAP embeddings for each case are presented in Fig. 13 (a), (b), and (c), respectively.

Fig. 13 (a) shows the clustering of narratives only, revealing several distinct clusters with a significant portion of the data concentrated in one large cluster. Fig. 13 (b) shows the clustering of text-based variables, which exhibits a more dispersed distribution of data points across the feature space. The combination of narratives and text-based variables, as shown in Fig. 13 (c), results in a markedly improved clustering structure. The combined data yields more distinct and well-separated clusters, suggesting enhanced differentiation between accident types. This improvement is particularly evident when compared to both the narrative-only and text-only variable clustering, demonstrating that the merger contributes valuable contextual information to the analysis. Similarly, the subsequent visualizations offer a comparison of 2D representations with respect to various variables across all three cases: subunit (Fig. 13 d, e, f), accident types (Fig. 13 g, h, i), mining equipment (Fig. 13 j, k, l), and injured body parts (Fig. 13 m, n, o).

To quantitatively assess the impact of merging text-based variables with narratives, the study employed the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) metrics. These measures evaluate the similarity between clustering results, with values ranging from  $-1$  to  $1$ , where  $1$  indicates perfect agreement. The comparison between narratives only and merged narratives yielded an ARI of  $0.1796$  and an AMI of  $0.3183$ , while the comparison between text-only variables and merged narratives resulted in an ARI of  $0.3771$  and an AMI of  $0.3565$ . These scores reveal a notable lack of strong agreement between either the narrative-only or text-only variable cases when compared to the case of the merged narratives. The relatively low ARI and AMI values in both comparisons indicate substantial differences in cluster assignments and information content. While the text-only variables case shows slightly higher similarity to the merged case than the narratives-only case, both comparisons demonstrate significant dissimilarity. This indicates that the merged case produces a unique clustering structure that is not fully captured by either individual component.

This analysis suggests that the combination of narratives and text-based variables produces a distinct clustering that leverages information from both narratives and text-based variables. This underscores the value of merging text-based data for a more comprehensive analysis of accident patterns.

### 5.9. Practical implications and framework applications

The identified clusters across various mining operations through the SBERT-UMAP-kmeans framework yielded distinct and actionable accident patterns across various mining operations, providing a granular understanding of hazard causation that can directly inform targeted safety interventions. For instance, in underground mining, the high proportion of "no-equipment" incidents, frequently associated with specific injury patterns such as torso injuries from over-exertion ( $67\%$ ), lower extremity injuries from slips/falls ( $58\%$ ), and upper extremity injuries from over-exertion ( $95\%$ ) and material handling ( $85\%$ ), highlights a critical need for enhanced ergonomic training, proper lifting techniques, and improved housekeeping protocols. These insights suggest that safety programs in such environments should extend beyond equipment-centric hazards to emphasize manual task safety, aligning with research that demonstrates the prevalence of work-related musculoskeletal disorders (WMSDs) in mining, often affecting the back, shoulders, and knees due to overexertion and falls (Weston et al., 2016). Similarly, the strong association of specific clusters with equipment like drilling tools, loaders, and bolting equipment, as well as the high frequency of vehicle-related accidents ( $98\%$  transportation,  $99\%$  loaders) in strip/quarry/open pit operations often leading to multiple body part injuries, indicates the necessity for machinery-specific safety controls, rigorous maintenance, and comprehensive operator training. This aligns with systematic reviews highlighting that a significant portion of heavy machinery accidents are often linked to ineffective or inappropriate training methods (Duarte et al., 2021). Furthermore, the consistent observation of noise-induced hearing loss ( $96\text{--}97\%$ ) across diverse mining types (strip, underground, milling) points to a pervasive occupational health hazard. This finding resonates with studies reporting high prevalence rates of hearing loss among noise-exposed workers in the mining sector, reaching as high as  $24\%$  (Lawson et al., 2019). Additionally, analysis of MSHA citations reveals that  $25.6\%$  of coal mines and  $14.7\%$  of non-coal mines were cited for noise violations, with larger mines showing higher citation rates, reinforcing the widespread nature of this hazard across the industry. This reinforces the need for robust hearing conservation programs, including regular noise monitoring, provision of hearing protection, and audiometric testing, which are crucial for mitigating hearing loss risks in the industry (Sun and Azman, 2018). These specific, data-driven patterns enable safety professionals to implement interventions precisely aligned with the most impactful hazards identified within each operational context.

This proposed SBERT-UMAP framework also possesses significant

potential for diverse real-world applications in mining safety. It is highly applicable for the instantaneous analysis of accident risk at mining sites, providing a robust mechanism for real-time, site-level decision-making. New accident narratives or near-miss reports can be rapidly processed through the pre-trained SBERT model to generate embeddings, which are then projected into the established UMAP-derived semantic space. By assessing their proximity to existing clusters, mine operators can quickly infer potential hazard patterns and understand associated risks, enabling immediate activation of appropriate preventative measures. This capability fosters proactive safety management, facilitating swift responses to emerging or recurring hazards even in data-sparse settings. Moreover, the framework demonstrates strong potential for integration with other predictive solutions, such as injury severity prediction models. The semantically rich SBERT embeddings, capturing high-dimensional accident characteristics, can serve as powerful input features for separate machine learning models trained to predict injury severity (e.g., fatal, days away from work, restricted activity, or specific injury classifications). The clustering results themselves could also provide valuable contextual features, enhancing the predictive accuracy of such downstream models and allowing for a more holistic accident analysis system that combines pattern identification with predictive capabilities.

Although our study focuses on the MSHA dataset, the proposed method is generalizable and can be applied to other safety narrative datasets. These include safety report datasets from organizations such as WorkSafeBC (Canada), Workers' Compensation reports from various U. S. states, the UK Health and Safety Executive (HSE), or internal incident reports maintained by mining companies. However, unlike the MSHA dataset, many of these sources are not publicly accessible. Furthermore, our framework leverages pre-trained language models and unsupervised learning, it does not rely on specific field names or data formats. With minimal preprocessing, the approach can be readily adapted to other mining incident reports or occupational safety datasets.

### 5.10. Limitation and future work

While this methodology demonstrates effectiveness, several limitations emerge from the results. The clustering analysis reveals overlapping patterns across different mining operations, making it challenging to establish clear distinctions between certain clusters. This is particularly evident in clusters sharing similar accident characteristics but differentiated only by mining methods, such as clusters 1.0, 2.1, and 4.0, which show comparable patterns of strain and sprain-related injuries. Additionally, some clusters like 4.3, 5.3, and 5.4 exhibit mixed operational contexts, with no dominant mining subunit, indicating limitations in clearly categorizing certain types of accidents within specific operational settings. Furthermore, the results from narratives-only clustering demonstrate limited pattern distinction, suggesting potential limitations in SBERT's ability to capture complex relationships in technical mining-specific narratives without additional structured information. This limitation arises because general-purpose language models like SBERT, while powerful, may not adequately grasp the implicit context, industry-specific jargon, and subtle causal chains unique to mining incidents when relying solely on raw text.

Future work could focus on addressing these limitations to further enhance the framework's capabilities. Specifically, the framework could be significantly improved by developing safety-specific language models to better capture technical terminology and contextual relationships in accident narratives. This could involve fine-tuning existing transformer-based models (e.g., a BERT variant like SciBERT, or a model specifically pre-trained on large mining safety corpora) or exploring domain-adaptive pre-training strategies on extensive datasets of mining-related texts. Additionally, it would be valuable to investigate alternative or advanced embedding techniques beyond standard SBERT. For instance, exploring models that incorporate external knowledge graphs of mining terminology, or integrating multi-modal fusion algorithms if



other forms of data (e.g., images, sensor data) become available alongside narratives, could provide richer and more contextually aware sentence embeddings. Validating this enhanced framework across diverse datasets would further help establish more robust pattern identification techniques for accident analysis.

## 6. Conclusion

This study introduces and implements a novel NLP-based framework for analyzing and clustering accident narratives in the U.S. mining industry. By integrating a transformer-based model (Sentence-BERT), nonlinear dimensionality reduction (UMAP), and unsupervised machine learning (k-means clustering), this framework effectively addressed critical limitations of traditional accident analysis methods, particularly their inability to capture nuanced semantic relationships and integrate unstructured data with structured metadata. A primary contribution of this work is demonstrating the enhanced effectiveness of merging text-based variables with accident narratives. This pre-embedding data fusion significantly improved the differentiation between accident types, leading to more distinct and semantically rich clusters compared to analyzing narratives or metadata in isolation. This integrated approach facilitates a more comprehensive understanding of accident causation patterns. The framework's reliance on UMAP dimensionality reduction also proved crucial for computational efficiency and pattern extraction, with the clustering approach exhibiting robustness across varying low-dimensional embeddings (3, 100, and 300 dimensions).

The clustering analysis on UMAP low-dimensional embeddings effectively characterized distinct accident patterns across various mining operations. The combination of k-means clustering and hierarchical agglomerative sub-clustering revealed nuanced relationships between injured body part, nature of injuries, accident type, and mining equipment across different operational contexts (e.g., underground, strip/quarry/open pit, milling, and dredge mining). These insights uncovered several semantically coherent clusters that highlight specific hazard characteristics unique to each operational environment, offering valuable insights for targeted hazard detection and safety interventions. For instance, the findings underscore prevalent issues such as non-equipment related sprain/strain injuries, specific equipment-related incidents, and occupational health concerns like noise-induced hearing loss and dust diseases, often differentiated by mining subunit. These cluster-level insights reveal concentrated patterns that can support more targeted, data-informed safety interventions.

The framework introduced in this study demonstrates an effective approach for mining accident pattern analysis by leveraging the rich information contained in narratives and associated text-based data. Its ability to uncover relationships between various accident characteristics provides a systematic framework for understanding accident occurrence patterns. From a practical standpoint, this clustering framework offers utility not only for analyzing historical trends but also for supporting real-time, site-level decision-making. Individual mines with limited data can project their accident narratives into the shared semantic space established by this study. By assessing their proximity to existing clusters, mine operators can infer similarities to broader trends, enabling them to identify potential hazard patterns and align safety planning with national-level insights - even in data-sparse settings.

## CRedit authorship contribution statement

**Abid Ali Khan Danish:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Snehamoy Chatterjee:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition. **Kumar Vaibhav Raj:** Writing – review & editing, Validation. **Rennie Kaunda:** Writing – review & editing. **Hugh Miller:** Writing – review & editing, Validation. **Aref Majdara:** Writing – review & editing, Validation.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Abid Ali Khan Danish reports that this study is partially funded by the National Institute for Occupational Safety and Health. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We gratefully acknowledge the National Institute for Occupational Safety and Health, United States (NIOSH) for their partial support of this research under Contract Number 75D30121C12375. The second author also acknowledges the support provided by the Witte Family Faculty Fellow funds.

The authors also acknowledge the use of AI language models in improving the clarity and readability of this manuscript's technical writing, while maintaining the complete scientific integrity of the content.

## Data availability

Publicly available data and source is cited in manuscript.

## References

- Ahmadi, E., Muley, S., Wang, C., 2024. Automatic construction accident report analysis using large language models (LLMs). *J. Intell. Constr.* <https://doi.org/10.26599/jic.2024.9180039>.
- Amoako, R., Buaba, J., Brickey, A., 2021. Identifying risk factors from MSHA accidents and injury data using logistic regression. *Mining, Metall. Explor.* 38 (1), 509–527. <https://doi.org/10.1007/s42461-020-00347-x>.
- Apud, E., 2012. Ergonomics in mining: the Chilean experience. *Hum. Factors* 54 (6), 901–907. <https://doi.org/10.1177/0018720812456701>.
- Arciniega-Rocha, R.P., Erazo-Chamorro, V.C., Szabo, G., 2023. The prevention of industrial manual tool accidents considering occupational health and safety. *Saf. Now* 9 (3). <https://doi.org/10.3390/safety9030051>.
- Ayesha, S., Hanif, M.K., Talib, R., 2020. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Inf. Fusion* 59, 44–58. <https://doi.org/10.1016/j.inffus.2020.01.005>.
- Baarsch, J., Celebi, M.E., 2012. Investigation of internal validity measures for K-means clustering. In: *Proceedings of the International Multiconference of Engineers and Computer Scientists*. sn, pp. 14–16.
- Babić, K., Martinčić-Ipsić, S., Meštrović, A., 2020a. Survey of neural text representation models. *Inf.* 11 (11), 1–32. <https://doi.org/10.3390/info11110511>.
- Babić, K., Martinčić-Ipsić, S., Meštrović, A., 2020b. Survey of neural text representation models. *Information* 11 (11). <https://doi.org/10.3390/info11110511>.
- Bahn, S., 2013. Workplace hazard identification and management: the case of an underground mining operation. *Saf. Sci.* 57, 129–137. <https://doi.org/10.1016/j.ssci.2013.01.010>.
- Bazaluk, O., Anisimov, O., Saik, P., Lozynskyi, V., Akimov, O., Hrytsenko, L., 2023. Determining the safe distance for mining equipment operation when forming an internal dump in a deep open pit. *Sustainability* 15 (7). <https://doi.org/10.3390/su15075912>.
- Becht, E., et al., 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37 (1), 38–44.
- BLS, “Number and rate of nonfatal work injuries and illnesses in private industries,” US BUREAU OF LABOR STATISTICS. [Online]. Available: <https://www.bls.gov/charts/injuries-and-illnesses/number-and-rate-of-nonfatal-work-injuries-and-illnesses-by-industry.htm>.
- Bridgell, R., Tolliver, D.D., 2024. Railroad accident analysis by machine learning and natural language processing. *J. Rail Transp. Plan. Manag.* 29 (October 2022), 100429. <https://doi.org/10.1016/j.jrtpm.2023.100429>.
- Burgess-Limerick, R., 2011. Injuries associated with underground coal mining equipment in Australia. *Ergon. Open J.* 4 (1), 62–73. <https://doi.org/10.2174/1875934301104010062>.
- Butlewski, M., Dahlke, G., Drzewiecka, M., Pacholski, L., 2015. Fatigue of miners as a key factor in the work safety system. *Procedia Manuf.* 3, 4732–4739. <https://doi.org/10.1016/j.promfg.2015.07.570>.
- Colinet, J.F., 2020. The impact of Black lung and a methodology for controlling respirable dust. *Mining, Metall. Explor.* 37 (6), 1847–1856. <https://doi.org/10.1007/s42461-020-00278-7>.
- A. A. K. Danish, “UMAP 3D visualization - MSHA data. Github.” [Online]. Available: [https://github.com/adanish91/UMAP\\_3D\\_visualization\\_MSHA\\_data](https://github.com/adanish91/UMAP_3D_visualization_MSHA_data).



- Darda'u Rafindadi, A., et al., 2025. Caught-In/Between accidents in the construction industry: a systematic review. *Saf. Now* 11 (1), 1–23. <https://doi.org/10.3390/safety11010012>.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *NAACL HLT 2019 - 2019 Conf. vol. 1. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., Mlm*, pp. 4171–4186.
- Digges, K.H., Eigen, A.M., 2004. "CRASH attributes that influence the severity of rollover crashes. In: *OCCUPANT AND VEHICLE RESPONSES IN ROLLOVERS*," Publ. Soc. Automot. Eng. [Online]. Available: <https://api.semanticscholar.org/CorpusID:106622021>.
- Duarte, J., Torres Marques, A., Santos Baptista, J., 2021. Occupational accidents related to heavy machinery: a systematic review. *Saf. Now* 7 (1). <https://doi.org/10.3390/safety7010021>.
- Ezisi, C.N., 2019. Risk factors for work-related eye injuries among stone quarry workers: a field report. *Niger. J. Ophthalmol.* 27 (1) [Online]. Available: [https://journals.lww.com/njoo/fulltext/2019/27010/risk\\_factors\\_for\\_work\\_related\\_eye\\_injuries\\_among\\_7.aspx](https://journals.lww.com/njoo/fulltext/2019/27010/risk_factors_for_work_related_eye_injuries_among_7.aspx).
- Figueres-Esteban, M., Hughes, P., van Gulijk, C., 2016. Visual analytics for text-based railway incident reports. *Saf. Sci.* 89, 72–76. <https://doi.org/10.1016/j.ssci.2016.05.009>.
- Ganguli, R., Miller, P., Pothina, R., 2021. Effectiveness of natural language processing based machine learning in analyzing incident narratives at a mine. *Minerals* 11 (7), 1–13. <https://doi.org/10.3390/min11070776>.
- Hansen, R., 2018. Fire statistics from the mining industry in New South Wales, Queensland and Western Australia. [https://www.researchgate.net/publication/323773014\\_Fire\\_statistics\\_from\\_the\\_mining\\_industry\\_in\\_New\\_South\\_Wales\\_Queensland\\_and\\_Western\\_Australia](https://www.researchgate.net/publication/323773014_Fire_statistics_from_the_mining_industry_in_New_South_Wales_Queensland_and_Western_Australia).
- Horberry, T., Burgess-Limerick, R., Fuller, R., 2013. The contributions of human factors and ergonomics to a sustainable minerals industry. *Ergonomics* 56 (3), 556–564. <https://doi.org/10.1080/00140139.2012.718800>.
- Hozumi, Y., Wang, R., Yin, C., Wei, G.-W., 2021. UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Comput. Biol. Med.* 131, 104264. <https://doi.org/10.1016/j.cmbiomed.2021.104264>.
- Hrica, J.K., Bellanca, J.L., Benbourenane, I., Carr, J.L., Stabryla, K.M., 2023. A rapid review of collision avoidance and warning technologies for mining haul trucks, 39 (4), 1357–1389. <https://doi.org/10.1007/s42461-022-00633-w>.
- Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J., 2023. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* 622, 178–210. <https://doi.org/10.1016/j.ins.2022.11.139>.
- Jiang, B., Wang, K., 2024. Railway accident causation prediction with improved transformer model based on lexical information and contextual relationships. *Knowledge-Based Syst.* 296 (April), 111897. <https://doi.org/10.1016/j.knsys.2024.111897>.
- Jing, S., et al., 2022. Correlation analysis and text classification of chemical accident cases based on word embedding. *Process Saf. Environ. Prot.* 158, 698–710. <https://doi.org/10.1016/j.psep.2021.12.038>.
- Jones, O.F., James, C.L., 2018. Task rotation in an underground coal mine: a pilot study. *Work* 59 (2), 285–294. <https://doi.org/10.3233/WOR-172669>.
- Kecejevic, V., Radomsky, M., 2004. The causes and control of loader- and truck-related fatalities in surface mining operations. *Inj. Control Saf. Promot.* 11 (4), 239–251. <https://doi.org/10.1080/156609704.233/289779>.
- Kemajl, Z., Stojance, M., Gzim, I., Ledi, M.L., 2024. Comprehensive analysis of the mining accident forecasting and risk assessment methodologies: case study – Stanterg Mine. *Min. Miner. Depos.* 18 (2), 11–17. <https://doi.org/10.33271/mining18.02.011>.
- Komljenovic, D., Groves, W.A., Kecejevic, V.J., 2008. Injuries in U.S. mining operations – a preliminary risk analysis. *Saf. Sci.* 46 (5), 792–801. <https://doi.org/10.1016/j.ssci.2007.01.012>.
- Kowalski, T., rakoffler K.M., 2011. *Psychological and Behavioral Aspects of Occupational Safety and Health in the US Coal Mining Industry*, first ed. Routledge [Online]. Available: <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315598697-13/psychological-behavioral-aspects-occupational-safety-health-us-coal-mining-industry-kathleen-kowalski-trakoffler-charles-vaught-linda-jansen-mcwilliams-dori-reissman>.
- Lawson, S.M., Masterson, E.A., Azman, A.S., 2019. Prevalence of hearing loss among noise-exposed workers within the Mining and Oil and Gas extraction sectors, 2006–2015. *Am. J. Ind. Med.* 62 (10), 826–837. <https://doi.org/10.1002/ajim.23031>.
- Li, J., Wang, J., Xu, N., Hu, Y., Cui, C., 2018. Importance degree research of safety risk management processes of urban rail transit based on text mining method. *Information* 9 (2), 26.
- Ma, Z., Chen, Z.S., 2024. Mining construction accident reports via unsupervised NLP and Accimap for systemic risk analysis. *Autom. Constr.* 161 (February), 105343. <https://doi.org/10.1016/j.autcon.2024.105343>.
- Maiti, J., V Khanzode, V., 2009. Development of a relative risk model for roof and side fall fatal accidents in underground coal mines in India. *Saf. Sci.* 47 (8), 1068–1076. <https://doi.org/10.1016/j.ssci.2008.12.003>.
- May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R., 2019. On measuring social biases in sentence encoders. *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. 1*, 622–628. <https://doi.org/10.18653/v1/n19-1063>.
- McInnes, L., Healy, J., Melville, J., 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.* 3 (29), 861.
- Mine Safety and Health Administration. Mine data retrieval System. <https://www.msha.gov/data-and-reports/mine-data-retrieval-system>.
- Mohapatra, S., Shaikh, A., Nayak, P., Navada, R., 2017. Hazards and health risks encountered by manual sand dredgers from udupi, India: a cross-sectional Study. *J. Clin. Diagn. Res.* 11 (7), LC11–LC17. <https://doi.org/10.7860/JCDR/2017/28960.10280>.
- Moniri-Morad, A., Shishvan, M.S., Aguilar, M., Goli, M., Sattarvand, J., 2024. Powered haulage safety, challenges, analysis, and solutions in the mining industry; a comprehensive review. *Results Eng.* 21, 101684. <https://doi.org/10.1016/j.rineng.2023.101684>.
- National Institute for Occupational Safety and Health (NIOSH). Number and rate of occupational mining fatalities by year, 2000 - 2023. <https://www.cdc.gov/NIOSH-Mining/MMWC/Fatality/NumberAndRate?StartYear=2000&EndYear=2023&SelectedOperatorType=&SelectedMineType=>.
- National Institute for Occupational Safety and Health (NIOSH). Number and rate of nonfatal lost-time injuries, 2000 - 2023. <https://www.cdc.gov/NIOSH-Mining/MMWC/Injuries/NumberAndRate?StartYear=2000&EndYear=2023&SelectedOperatorType=&SelectedMineType=>.
- NLTK. NLTK documentation. <https://www.nltk.org/>.
- Nowrouzi-Kia, B., et al., 2018. A systematic review of lost-time injuries in the global mining industry. *Work* 60 (1), 49–61. <https://doi.org/10.3233/WOR-182715>.
- Nthakana, L., Nelson, G., Khoza-Shangase, K., 2020. Estimating miners at risk for occupational noise-induced hearing loss: a review of data from a South African platinum mine. *South African J. Commun. Disord.* 67 (2), 1–8. <https://doi.org/10.4102/sajcd.v67i2.677>.
- NumPy, 2004. NumPy documentation. <https://numpy.org/>.
- Onder, S., 2013. Evaluation of occupational injuries with lost days among opencast coal mine workers through logistic regression models. *Saf. Sci.* 59, 86–92. <https://doi.org/10.1016/j.ssci.2013.05.002>.
- Onder, M., Onder, S., Mutlu, A., 2012. Determination of noise induced hearing loss in mining: an application of hierarchical loglinear modelling. *Environ. Monit. Assess.* 184 (4), 2443–2451. <https://doi.org/10.1007/s10661-011-2129-0>.
- pandas, 2024. Pandas documentation. <https://pandas.pydata.org/>.
- Papic, L., Kovacevic, S., 2016. Human factor in mining machines maintenance operations. In: *2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO)*, pp. 456–465. <https://doi.org/10.1109/SMRLO.2016.80>.
- Qiao, Y., Xiong, C., Liu, Z., Liu, Z., 2019. Understanding the behaviors of BERT in ranking. <https://arxiv.org/abs/1904.07531>.
- Ramos, P.M.S., Macedo, J.B., Maior, C.B.S., Moura, M.C., Lins, I.D., 2022. Combining BERT with numerical variables to classify injury leave based on accident description. *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* <https://doi.org/10.1177/1748006X221140194>, 1748006X221140194.
- Reimers, N., Gurevych, I., 2019. Sentence-BERT: sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 3982–3992. <https://doi.org/10.18653/v1/d19-1410>.
- Rose, R.L., Puranik, T.G., Mavris, D.N., 2020. Natural language processing based method for clustering and analysis of aviation safety narratives. *Aerospace* 7 (10), 1–22. <https://doi.org/10.3390/aerospace7100143>.
- Sarkar, S., Vinay, S., Raj, R., Maiti, J., Mitra, P., 2019. Application of optimized machine learning techniques for prediction of occupational accidents. *Comput. Oper. Res.* 106, 210–224. <https://doi.org/10.1016/j.cor.2018.02.021>.
- Shao, B., Hu, Z., Liu, D., 2019. Using improved principal component analysis to explore construction accident situations from the multi-dimensional perspective: a Chinese Study. *Int. J. Environ. Res. Publ. Health* 16 (18). <https://doi.org/10.3390/ijerph16183476>.
- Simpson, G., Horberry, T., 2009. *Understanding Human Error in Mine Safety*, 1st Editio. CRC Press, London. <https://doi.org/10.1201/9781315549194>.
- Singh, V.V., Lakra, A.R., Gautam, S., 2025. Seasonal impact of particulate matter exposure on urinary health of mine workers in iron ore mining environment. *Geosystems and Geoenvironment* 4 (1), 100342. <https://doi.org/10.1016/j.geogeo.2024.100342>.
- Smetana, M., de Salles, L., Sukharev, I., Khazanovich, L., 2024. Highway construction safety analysis using large Language models. *Appl. Sci.* 14 (4). <https://doi.org/10.3390/app14041352>.
- Souza, J., Castelo Branco, J., Baptista, J.S., 2024. Work accidents related to heavy equipment in the open pit extractive industry: a systematic review. In: *Azeires, P.M., Melo, R.B., Carneiro, P., et al. (Eds.), Occupational and Environmental Safety and Health V. Springer Nature Switzerland, Cham*, pp. 783–791. [https://doi.org/10.1007/978-3-031-38277-2\\_62](https://doi.org/10.1007/978-3-031-38277-2_62).
- Srinivasan, P., Nagarajan, V., Mahadevan, S., 2019. Mining and Classifying Aviation Accident Reports. *AIAA Aviat. 2019 Forum*. <https://doi.org/10.2514/6.2019-2938>.
- Stemm, E., 2019. Analysis of injuries in the Ghanaian mining industry and priority areas for research. *Saf. Health Work* 10 (2), 151–165. <https://doi.org/10.1016/j.shaw.2018.09.001>.
- Stewart, B.M., Clark, C., Patton, P., 2001. *Materials handling accident reduction in underground mines - synopsis*. <https://stacks.cdc.gov/view/cdc/9091>.
- Sun, K., Azman, A.S., 2018. Evaluating hearing loss risks in themining industry through MSHA citations. *J. Occup. Environ. Hyg.* 15 (3), 246–262. <https://doi.org/10.1080/15459624.2017.1412584>.
- Sun, K., Azman, A.S., Camargo, H.E., Dempsey, P.G., 2019. Risk assessment of recordable occupational hearing loss in the mining industry. *Int. J. Audiol.* 58 (11), 761–768. <https://doi.org/10.1080/14992027.2019.1622041>.
- Tan, Z.L., Chen, X., Song, Q.Z., Chen, X.C., 2017. Analysis for the potential hazardous risks of the coal mines based on the so-called text mining. *J. Saf. Env.* 4, 1262–1266.

- Tixier, A.J.P., Hallowell, M.R., Rajagopalan, B., Bowman, D., 2016. Application of machine learning to construction injury prediction. *Autom. Constr.* 69 (1), 102–114. <https://doi.org/10.1016/j.autcon.2016.05.016>.
- V Raj, K., Tarshizi, E.K., 2020. Advanced application of text analytics in MSHA metal and nonmetal fatality reports. In: *SME Annual Meeting & Expo*. Phoenix, AZ, USA.
- Valcamonico, D., Baraldi, P., Amigoni, F., Zio, E., 2022. A framework based on Natural Language Processing and Machine Learning for the classification of the severity of road accidents from reports. *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* <https://doi.org/10.1177/1748006X221140196>, 1748006X221140196.
- Vaswani, A., et al., 2017. Attention is all you need. In: *Adv. Neural Inf. Process. Syst.*, 2017-Decem Nips, pp. 5999–6009.
- Wang, H., Zhou, C., Li, L., 2019. Design and application of a text clustering Algorithm based on parallelized K-Means clustering. *Rev. d'Intelligence Artif* 33 (6), 453–460.
- Wang, F., et al., 2024. A Comprehensive Survey of small Language models in the era of large Language models: techniques, enhancements, applications, Collaboration with LLMs, and trustworthiness. <http://arxiv.org/abs/2411.03350>.
- Weston, E., Nasarwanji, M.F., Pollard, J.P., 2016. Identification of work-related musculoskeletal disorders in mining. *J. safety, Heal. Environ. Res.* 12 (1), 274–283 [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/27294012%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4902268>.
- Williams, T., Betak, J., Findley, B., 2016. Text mining analysis of railroad accident investigation reports. In: *ASME/IEEE Joint Rail Conference*. American Society of Mechanical Engineers. V001T06A009.
- Wu, M., Hu, N., Ye, Y., Wang, Q., Wang, X., 2022. Multi-hazard risk characterization and collaborative control oriented to space in non-coal underground mines. *Sci. Rep.* 12 (1), 16452. <https://doi.org/10.1038/s41598-022-20437-8>.
- Xiaojie, Y., Qiaoyun, H., Jiewen, P., Xiaowei, S., Dinggui, H., Chao, L., 2011. Progress of heat-hazard treatment in deep mines. *Min. Sci. Technol.* 21 (2), 295–299. <https://doi.org/10.1016/j.mstc.2011.02.015>.
- Xu, N., Ma, L., Liu, Q., Wang, L., Deng, Y., 2021. An improved text mining approach to extract safety risk factors from construction accident reports. *Saf. Sci.* 138 (Jun). <https://doi.org/10.1016/j.ssci.2021.105216>.
- Zaitsev, A., Parshakov, O., Semin, M., 2024. Heat emissions from mining machinery: implications for microclimatic conditions in underground workings. *Mining* 4 (4), 1075–1092. <https://doi.org/10.3390/mining4040059>.
- Zhang, F., 2022. A hybrid structured deep neural network with Word2Vec for construction accident causes classification. *Int. J. Constr. Manag.* 22 (6), 1120–1140. <https://doi.org/10.1080/15623599.2019.1683692>.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2019. BERTScore: evaluating text generation with BERT, 1–43. <http://arxiv.org/abs/1904.09675>.
- Zhang, Q., Liu, Z., Pan, S., 2025. The rise of small Language models. *IEEE Intell. Syst.* 40 (1), 30–37. <https://doi.org/10.1109/MIS.2024.3517792>.
- Zhong, B., Pan, X., Love, P.E.D., Ding, L., Fang, W., 2020a. Deep learning and network analysis: classifying and visualizing accident narratives in construction. *Autom. Constr.* 113. <https://doi.org/10.1016/j.autcon.2020.103089>.
- Zhong, B., Pan, X., Love, P.E.D., Sun, J., Tao, C., 2020b. Hazard analysis: a deep learning and text mining framework for accident prevention. *Adv. Eng. Informatics* 46. <https://doi.org/10.1016/j.aei.2020.101152>.