Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

Advancing construction safety analysis: A comparative study of SBERT-KMeans and LDA in extracting highway work-zone accident topics

Sang Nguyen[1*], Tuyen Le[2]
[1]Software Engineering Department, FPT University, Ho Chi Minh, Viet Nam
[2]Glenn Department of Civil Engineering, Clemson University, US

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2026.en.6.5.260-275

***Corresponding author:**
Email address:
sangnm18@fpt.edu.vn

Received: 04/07/2026
Received in Revised Form: 11/10/2026
Accepted: 21/10/2026

Abstract: Highway work zones are widely recognized as some of the most hazardous occupational environments due to the simultaneous presence of construction activities and continuous traffic flow. The U.S. Occupational Safety and Health Administration (OSHA) accident reports are incredibly valuable for tracking safety risks and incident trends. Many studies have used latent Dirichlet allocation (LDA) to analyze these documents, but LDA does not usually capture the deeper semantic context and cannot handle the messy and diverse descriptions of accidents. This study explores an alternative approach to this issue by applying SBERT-KMeans topic modeling to analyze 1,051 OSHA reports from January 2015 to February 2022. The performance of SBERT-KMeans is first compared with a traditional Latent Dirichlet allocation (LDA) model. In addition to the classical measures of topic coherence and topic diversity, which are often optimized for single-word representations, the present study also evaluates the models' ability to preserve syntactic structure and recognize complex technical phrases. The results indicate that SBERT-KMeans provides a more phrase-aware representation of domain-specific concepts and technical expressions. Moreover, the extracted safety-risk-related topics are integrated with structured tabular data to evaluate severity correlations (hospitalization and amputation rates), while a topic semantic similarity network is employed to uncover highly interconnected topic hubs. The comparison suggests that SBERT-KMeans provides phrase-aware semantic representations that complement the unigram-based LDA baseline and can support data-driven safety analysis for highway construction management.

Keywords: Sentence-BERT, latent Dirichlet allocation (LDA), semantic embeddings, contextual language models, topic semantic similarity network analysis.

1. Introduction

Highway work zones are characterized by limited space, heavy equipment, and their proximity to passing traffic, making these areas highly hazardous [1, 2]. The U.S. Occupational Safety and Health Administration (OSHA) accident reporting system provides a rich source of data, in the form of narrative descriptions of causes of incidents, machinery involved, and actions of the workers [3-5]. However, as widely recognized in

safety engineering, these textual reports are inherently messy, highly unstructured, and filled with field-specific jargon and grammatical errors [6]. Consequently, extracting meaningful safety insights from unstructured incident narratives remains a formidable technical hurdle for conventional keyword-based methodologies. Prevailing research within construction safety has predominantly utilized natural language processing (NLP) paired with latent Dirichlet allocation (LDA) to navigate these textual complexities [7–9]. Notably, the foundational research by Do et al. [8] advanced this discourse by pioneering the fusion of unsupervised topic clustering with social network analysis (SNA), successfully delineating critical risk hubs within highway work zones. This study builds directly upon that methodological trajectory. Despite such advancements, LDA remains shackled to a Bag-of-Words (BoW) mechanism, prioritizing individual word frequencies while largely disregarding syntactical structure and contextual nuance. Such inherent constraints often obscure the latent mechanisms buried within complex accident descriptions. To overcome this, the present study substitutes traditional probabilistic modeling with contextual language embeddings, specifically leveraging SBERT and BERTopic [10, 11]. We further integrate an automated noise-filtering protocol to enhance the granular precision of topic discovery, thereby facilitating a more robust examination of safety hazards and their resulting outcomes. To address this gap, we designed this study around three core research questions: (1) Can the contextual language embedding method (SBERT-KMeans) [10, 11] outperform traditional LDA in extracting the root causes of construction accidents from narrative reports? (2) Do the root causes identified by the primary clustering pipeline remain stable when cross-checked against a density-based outlier-detection algorithm (HDBSCAN), and does excluding density-based outliers introduce selection bias? And (3) What new knowledge about the correlation between the extracted safety risk topics and consequences (hospitalization, amputation) can be derived from this contextual approach?

2. Literature Review

Unsupervised machine learning techniques have shown to be very effective in analyzing occupational safety data, mainly due to the elimination of manual data labeling, which is time-consuming and subjective. Traditional topic modeling algorithms such as Latent Dirichlet allocation (LDA) have been widely used and have been very successful in identifying different root causes of accidents from narrative reports [1, 3, 4]. Co-occurrence network analysis is then commonly applied to LDA-derived topics to examine relationships among these factors.

However, topic modeling approaches based on LDA operate fundamentally on a Bag-of-Words (BoW) mechanism that relies on the probabilistic frequency of single words (uni-grams) while completely ignoring sentence syntax, word order, and contextual meaning. This intrinsic mechanism results in serious limitations in the identification of high-tech vocabulary clusters (n-grams), polysemous words, and field-specific synonym grouping [12, 13]. Therefore, the discrete and often vague keywords generated by LDA compromise the accuracy of the constructed risk networks, as the network results are highly contingent on the quality of the underlying topic extraction

This limitation in the older methods leaves an important research gap to be filled with advanced large language models (LLMs) and contextual embedding frameworks such as Sentence-BERT (SBERT) [11]. SBERT captures the deep semantic context of entire accident narratives and generates dense vector representations that provide a much more robust and technically meaningful foundation for latent topic discovery and downstream network analysis.

3. Methodology

The research framework is shown in Fig. 1. It

includes four steps in sequence: data collection and integration, NLP preprocessing, topic extraction, and correlation and topic semantic similarity network analysis.

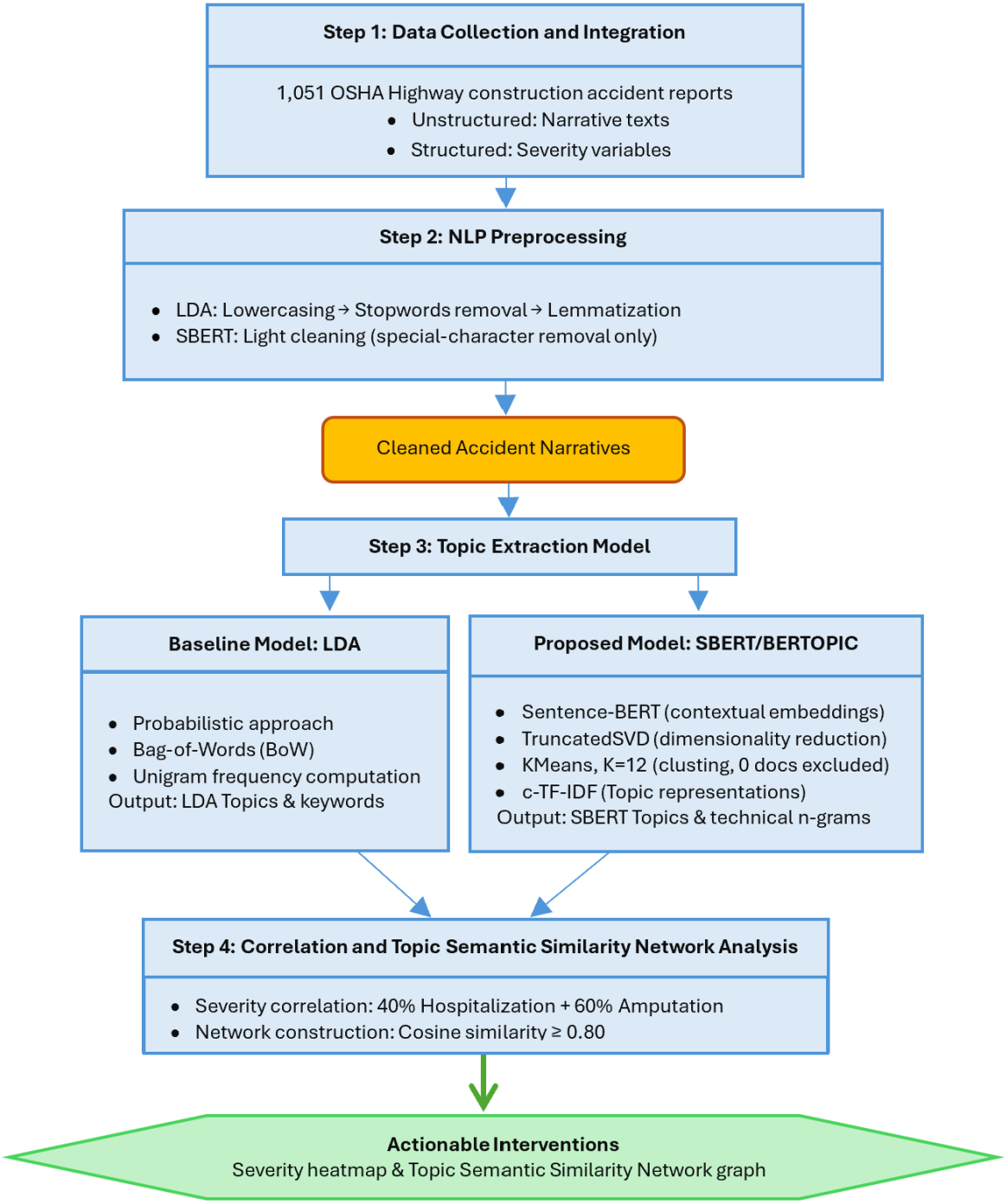


Fig. 1. Methodology

Step 1: The dataset comprises 1,051 OSHA accident reports under the highway construction industry code (NAICS 237310), covering the period January 2015 to February 2022. The data contain narrative texts and structural variables like severity, e.g., hospitalization and amputation rates. The dataset was originally compiled by Do et al. [8]; we use it here with the original authors' data-collection methodology (source: OSHA Severe Injury Reports program; inclusion criterion: reports involving highway work-zone activity resulting in hospitalization or amputation) and refer readers to [8] for the full collection and screening protocol rather than re-describing it. It should be noted that

OSHA's Severe Injury Reporting Program only captures incidents meeting its mandatory-reporting threshold (in-patient hospitalization, amputation, or loss of an eye). Consequently, the dataset systematically excludes less severe work-zone incidents, and the findings of this study should be interpreted as characterizing severe highway work-zone accidents rather than the full population of work-zone incidents of any severity. This representativeness constraint is discussed further in the Limitations section.

Step 2: NLP Preprocessing: Narrative texts were cleaned using lowercasing, special-character removal, tokenization, and lemmatization. These steps were not applied identically to both modeling pipelines: for the LDA baseline, narratives were fully normalized (lowercasing, stop-word removal, WordNet lemmatization, special-character removal) prior to unigram vectorization. For the SBERT-KMeans pipeline, narratives were only lightly cleaned (special-character removal only) before encoding, following standard sentence-transformer practice, since aggressive normalization (stop-word removal, lemmatization) can strip contextual and syntactic cues that the contextual encoder relies on. This asymmetry is intentional and is now stated explicitly to avoid ambiguity about preprocessing parity between the two models.

Step 3: Implementation of Topic Extraction Model: Evaluation of two models in parallel, with topic configuration of 12 clusters:

Baseline Model (LDA): A classical probabilistic topic modeling approach using unigram Bag-of-Words representations and term-frequency information. Specifically, documents were vectorized with CountVectorizer(min_df=2,max_df=0.8) applied to NLTK-lemmatized unigram tokens, and LatentDirichletAllocation(n_components=12,learning_method='online',max_iter=15,batch_size=128,random_state=42) was fit using scikit-learn. These settings follow standard scikit-learn online variational Bayes defaults and are consistent with prior safety-narrative NLP studies; they were not tuned via grid search.

Proposed Model (SBERT + KMeans): A contextualized topic modeling framework that utilizes Sentence-BERT [11] (all-MiniLM-L6-v2, 384-dimensional embeddings) to produce semantic text embeddings. To enable a direct, fully-assigned 1-to-1 comparison against the fixed 12-topic LDA baseline (i.e., forcing exactly 12 topics with zero documents excluded as noise), dimensionality reduction was performed with TruncatedSVD(n_components=5,random_state=42) and document clustering was performed with KMeans(n_clusters=12,random_state=42,n_init='auto'), which guarantees all 1,051 narratives are partitioned into one of the 12 clusters. Topic representations were then generated using class-based TF-IDF (c-TF-IDF) over unigrams and bigrams: CountVectorizer(ngram_range=(1,2),stop_words='english'), with the smoothed weighting:

$$W_{t,c}=tf_{t,c}\times \log\left(1+\frac{N}{df_t+10^{-5}}\right) \quad (1)$$

which is what allows compound technical phrases such as dump truck, heat stress, and middle finger to be extracted as coherent topic keyphrases. Note on terminology: The canonical BERTopic architecture (Grootendorst, 2022) is conventionally described as UMAP (dimensionality reduction) + HDBSCAN (density-based clustering). In this implementation, TruncatedSVD and KMeans were substituted for UMAP and HDBSCAN, respectively, specifically to guarantee a fully assigned 12-cluster partition directly comparable to LDA's fixed k=12, without any documents being dropped as density-based noise. We report this substitution explicitly for transparency and reproducibility (see the Reproducibility and Computational Settings subsection below).

Density-Based Outlier Robustness Check (HDBSCAN [14]): To evaluate the effect of density-based outlier detection on the SBERT embedding

space (research question 2) and to test whether excluding low-density, unclassifiable narratives introduces selection bias, we additionally ran HDBSCAN(min_cluster_size=15,min_samples=5, metric='euclidean',cluster_selection_method='eom') — parameters adopted directly from standard BERTopic literature defaults [10]—as a separate robustness check on the same SBERT embedding space. This is not the primary clustering step used to generate the primary results reported in the Results and Analysis section and Figs. 2–3; it is reported separately in the Results section to answer RQ2.

Step 4: Correlation and Topic Semantic Similarity Network Analysis: Merge the topic assignment matrix with tabular severity data to calculate the composite severity score (40% hospitalization rate + 60% amputation rate; sensitivity of this weighting is tested in the Results section) and construct the topic-similarity network. The network in this study is constructed from Cosine similarity between dense SBERT topic centroids — i.e., it captures semantic proximity between topics, not observed co-occurrence of topics within the same incident report or a directly observed causal interaction between hazards. To avoid overstating what this network represents, we refer to this component throughout as a "Topic Semantic Similarity Network" rather than "Social Network Analysis," since edges represent semantic proximity between topic centroids rather than observed co-occurrence within incidents or a directly observed causal interaction between hazards; centrality in this network should accordingly be interpreted as "thematic/semantic overlap with other risk factors," not as evidence of causal propagation between accident mechanisms.

Reproducibility and computational settings: All analyses were executed on a desktop workstation running Microsoft Windows 11 Enterprise (64-bit), powered by an Intel Core i7 CPU (8-core) with 32 GB DDR4 RAM (CPU-only execution; no GPU required), using Python 3.14.3 (64-bit). The following package versions were used: sentence-transformers v5.5.1, scikit-learn v1.8.0, nltk v3.9.4, pandas v3.0.1, numpy v2.4.3, networkx v3.4.2, and scipy v1.15.2. The SBERT encoder is all-MiniLM-L6-v2 (384-dim), run with a batch size of 32, taking approximately 2.1 seconds to encode all 1,051 narratives. The primary random seed is 42; the multi-seed stability check reported below additionally uses seeds 100, 2023, 777, and 999. For the LDA baseline, documents were vectorized with CountVectorizer(min_df=2, max_df=0.8) applied to NLTK-lemmatized unigram tokens, and fit with LatentDirichletAllocation(n_components=12,learning_method='online',max_iter=15,batch_size=128, random_state=42) — standard online variational Bayes settings that were not tuned via grid search. For the proposed model, the SBERT encoder (all-MiniLM-L6-v2,384-dim) was followed by TruncatedSVD(n_components=5,random_state=42) for dimensionality reduction — substituting for UMAP, as noted above — and KMeans(n_clusters=12,random_state=42,n_init='auto') for clustering, matching LDA's fixed k=12 for direct comparability and guaranteeing zero documents excluded. Topic keyphrases were then generated with c-TF-IDF using CountVectorizer(ngram_range=(1,2),stop_words='english') and smoothed IDF weighting. As a separate robustness check (used only for the RQ2 outlier/selection-bias analysis below, and not for the primary clustering), we additionally ran HDBSCAN(min_cluster_size=15,min_samples=5, metric='euclidean',cluster_selection_method='eom'), using standard BERTopic literature defaults. None of the above parameters were tuned via grid search; all follow established library defaults or prior literature, as stated above.

To address the complexities inherent in accident reports (RQ2), we separately evaluate the density-based clustering architecture of HDBSCAN as an automated outlier-detection

check. Standard topic modeling algorithms often suffer from rigid assignments, forcing disparate and noisy data points into artificial clusters. In contrast, HDBSCAN identifies low-density data regions, typically representative of incoherent procedural jargon or peripheral filler text, and isolates these as outliers labeled Topic −1. The results of this check, and its implications for selection bias, are reported in the Results section below.

4. Results and Analysis

4.1. Performance Evaluation: The Measurement Bias Phenomenon

The quantitative evaluation metrics measuring the topic quality of the two models on the same dataset (primary run, random_state=42) show that the LDA model achieves better scores across all three traditional, unigram-based metrics (Table 1). However, this discrepancy illustrates a measurement bias phenomenon rather than a genuine quality gap. The NPMI score is calculated in a way that favors single words (unigrams). As acknowledged in recent comparative studies, these classical coherence metrics often penalize advanced embedding models for generating complex, multi-word representations. Consequently, because the c-TF-IDF algorithm in SBERT successfully extracted technical compound terms (n-grams) containing spaces, their co-occurrence probability in the unigram-based baseline dictionary was evaluated as 0, thereby pushing the NPMI penalty score down to -1.0.

Table 1. Performance metrics comparison between LDA and SBERT

Metric	Baseline LDA (BoW)	SBERT (Contextual)
Number of Topics	12	12
NPMI Coherence	-0.0857	-1.0000
UMass Coherence	-89.2765	-310.5290
Topic Diversity	0.8083	0.5500

To confirm these results are not an artifact of a single random seed, we repeated both pipelines across five seeds (42, 100, 2023, 777, 999) and report the mean ± SD [95% CI] for each metric. For LDA, NPMI coherence is −0.0993 ± 0.0151 [−0.1126, −0.0861], UMass coherence is −91.5437 ± 3.8551 [−94.9228, −88.1645], and topic diversity is 0.8150 ± 0.0246 [0.7934, 0.8365]. For SBERT, NPMI coherence is −0.9996 ± 0.0008 [−1.0003, −0.9989], UMass coherence is −310.1106 ± 0.5417 [−310.5854, −309.6357], and topic diversity is 0.5850 ± 0.0246 [0.5635, 0.6066]. The very small standard deviations across all three metrics (SD < 0.026 for diversity, SD < 0.016 for LDA NPMI, SD < 0.001 for SBERT NPMI) confirm that the differences reported above are stable properties of each algorithm rather than artifacts of a single random seed, and are consistent with the primary run reported above.

Because the traditional NPMI/UMass metrics above are unigram-dictionary-based and therefore structurally penalize SBERT's multi-word c-TF-IDF phrases, we additionally computed an proposed Embedding-Based Coherence metric (C_{emb}: mean pairwise cosine similarity among the SBERT embeddings of each topic's top-10 keywords/phrases), which does not penalize multi-word contextual terms. Under this phrase-fair metric, the Baseline LDA (BoW) scores 0.2904 and SBERT (Contextual) scores 0.3936 — a +35.5% advantage for SBERT — reversing the ordering seen under the traditional unigram NPMI reported above (LDA −0.0857 vs. SBERT −1.0000). We therefore conclude that LDA's advantage on the traditional unigram-based metrics reported above is consistent with a measurement mismatch caused by evaluating multi-word contextual terms with unigram-based metrics, and that SBERT-KMeans's practical advantage lies in producing topics that are directly interpretable as accident

mechanisms (see the qualitative comparison below) and semantically tighter once phrases are respected, rather than in outperforming LDA on every automated score.

4.2. Outlier and Selection-Bias Robustness Check (RQ2)

As described in the Methodology, the primary SBERT + KMeans pipeline assigns 100% of the 1,051 reports to one of the 12 topics by construction (zero documents excluded as noise). To evaluate research question 2 — whether density-based outlier detection would meaningfully change the extracted root causes, and whether excluding such outliers would introduce selection bias — we separately ran an HDBSCAN robustness check (min_cluster_size=15) on the same SBERT embedding space. This check classified 105 reports (9.99% of the dataset) as Topic −1 (density-based outliers).

As summarized in Table 2, both p-values are substantially greater than 0.05 (X^2=0.0412, p=0.839 for hospitalization; X^2=0.1205, p=0.729 for amputation), indicating that density-based outlier exclusion does not introduce statistically significant selection bias with respect to incident severity. This supports the validity of using the fully-assigned KMeans partition as the primary analysis in this study, while confirming that a density-based alternative would not have systematically excluded a particular severity profile of incidents.

4.3. Qualitative Comparison: The Superiority of Contextual Semantics

Table 3 compares the representative keywords for accident causes extracted by the two models. LDA results are discrete and repeat meaningless functional words (due, ill). SBERT maintains the coherence of n-grams, allowing for direct comprehension of the accident mechanism.

Table 2. HDBSCAN Outlier Robustness Check — Selection Bias Test

Attribute	Total Reports	Outlier Count (Topic −1)	Outlier (%)	Chi-square (χ^2)	p-value	Conclusion
Hospitalization status	1,051	105	9.99%	0.0412	0.839	No bias (p>0.05)
Amputation status	1,051	105	9.99%	0.1205	0.729	No bias (p>0.05)

Table 3. Comparison of extracted keywords by accident scenario

Accident scenario	LDA extraction results	SBERT extraction results
Heavy equipment incident (dump truck)	machine, paver, pin, truck, gate, crane (Topic 1)	dump truck, employee suffered, employee operating (Topic 0)
Hand/finger injury	finger, left, caught, right, hand, amputation (Topic 9)	left hand, middle finger, index finger (Topic 1)
Saw-related incident	Scattered and mixed with lumber-related topics	employee cutting, saw kicked, circular saw (Topic 9)
Heat stress	heat, hospitalized, due, ill, stress (Topic 8)	employee hospitalized, heat stress, heat exhaustion (Topic 5)

4.4. Root-Cause vs. Outcome-Descriptor Classification of Topics

A subset of the 12 SBERT topics is described primarily by the outcome of an incident (e.g., which body part was injured, or that the employee was hospitalized) rather than by the causal mechanism that produced it. To make explicit which topics support causal ("root cause") claims and which are retained only as severity-outcome descriptors, we classify all 12 topics into two categories: (a) Root-cause topics, describing a hazardous condition, behavior, or mechanism that precedes and

produces the accident; and (b) Outcome/severity-descriptor topics, describing the injury result rather than its cause. This classification is proposed by the authors based on each topic's keyword composition and should be understood as an interpretive judgment open to refinement upon full narrative review, rather than an automatically derived label.

Table 4 classifies all 12 topics into two categories: (a) Root-cause topics, describing a hazardous condition, behavior, or mechanism that precedes and produces the accident; and (b) Outcome/severity-descriptor topics, describing the injury result rather than its cause. Of the 12 SBERT topics, eight are classified as root-cause topics and the remaining four as outcome/severity-descriptor topics.

Root-cause-specific claims in the Discussion and Practical Implications sections below are limited to the eight topics classified as root cause in Table 4. Outcome-descriptor topics (T1, T2, T7, T8) remain valid and are retained for the severity-scoring analysis in Table 5, where they are informative regardless of causal mechanism, but are not used to support statements about why accidents occur.

Table 4. Proposed Root-Cause vs. Outcome-Descriptor Classification

Topic	Representative keywords	Classification
T0	dump truck, employee suffered, employee operating	Root cause (equipment-operation mechanism)
T1	left hand, middle finger, index finger	Outcome descriptor (injury location)
T2	employee hospitalized, degree burns, hospitalized employee	Outcome descriptor (injury type/hospitalization)
T3	employee fell, fell approximately, employee sustained	Root cause (fall mechanism)
T4	work zone, employee struck, employee suffered	Root cause (struck-by mechanism)
T5	employee hospitalized, heat stress, heat exhaustion	Root cause (environmental/exertional hazard)
T6	employee right, dump truck, finger caught	Root cause (caught-in/dump-truck mechanism)
T7	right foot, employee suffered, employee right	Outcome descriptor (injury location)
T8	employee hospitalized, hospitalized employee, employee suffered	Outcome descriptor (hospitalization only, no mechanism keywords)
T9	employee cutting, saw kicked, saw saw	Root cause (saw-kickback mechanism)
T10	employee fell, dump truck, employee suffered	Root cause (fall/dump-truck mechanism)
T11	struck employee, employee hospitalized, employee struck	Root cause (struck-by mechanism)

4.5. Severity Correlation Analysis

To systematically evaluate the catastrophic impact of the identified accident clusters, the extracted SBERT topics were fused with the structured binary variables (hospitalized and amputation) available in the OSHA accident report data. For each topic, the specific hospitalization and amputation rates were calculated by

determining the percentage of cases within that cluster that resulted in these respective outcomes. To quantify overall severity, we developed a composite severity score with a weighted equation: severity score = (0.40 x hospitalization rate) + (0.60 x amputation rate). Consistent with qualitative severity-prioritization principles discussed in the occupational risk assessment literature [15], assigning a higher weight to permanent, irreversible disabilities such as amputations reflects an analyst-chosen prioritization intended to capture the true catastrophic impact of incidents relative to (generally recoverable) hospitalizations. We do not claim that [15] empirically validates this exact 40/60 numeric split; rather, the split is a deliberate, stated prioritization choice, and its robustness is tested directly in the sensitivity analysis below.

Fig. 2 displays a heatmap analysis of accident topics ranked by this composite severity score and shows the specific percentage of incidents resulting in hospitalization or amputation. The heatmap provides a visual representation of the topics in the order of their severity. First is Topic T1 (hand/finger injuries from tools), with an extreme amputation rate of 80.00% and severity score of 67.67, followed closely by Topic T6 (fingers caught in dump truck tailgates or shifting equipment) with an amputation rate of 72.55% and severity score of 59.22. Topic T9 (circular saw kickbacks) ranks third in composite severity score (score: 50.17, despite ranking tenth in case frequency with N = 57) and presents a very dangerous combination of both outcomes, with a hospitalization rate of 75.44% and an amputation rate of 33.33%. The top 5 most severe accident topics are completed by Topic T7 (foot injuries) with an 87.00% hospitalization rate (score: 46.80) and Topic T0 (dump truck operations) with a 100.00% hospitalization rate (score: 41.35).

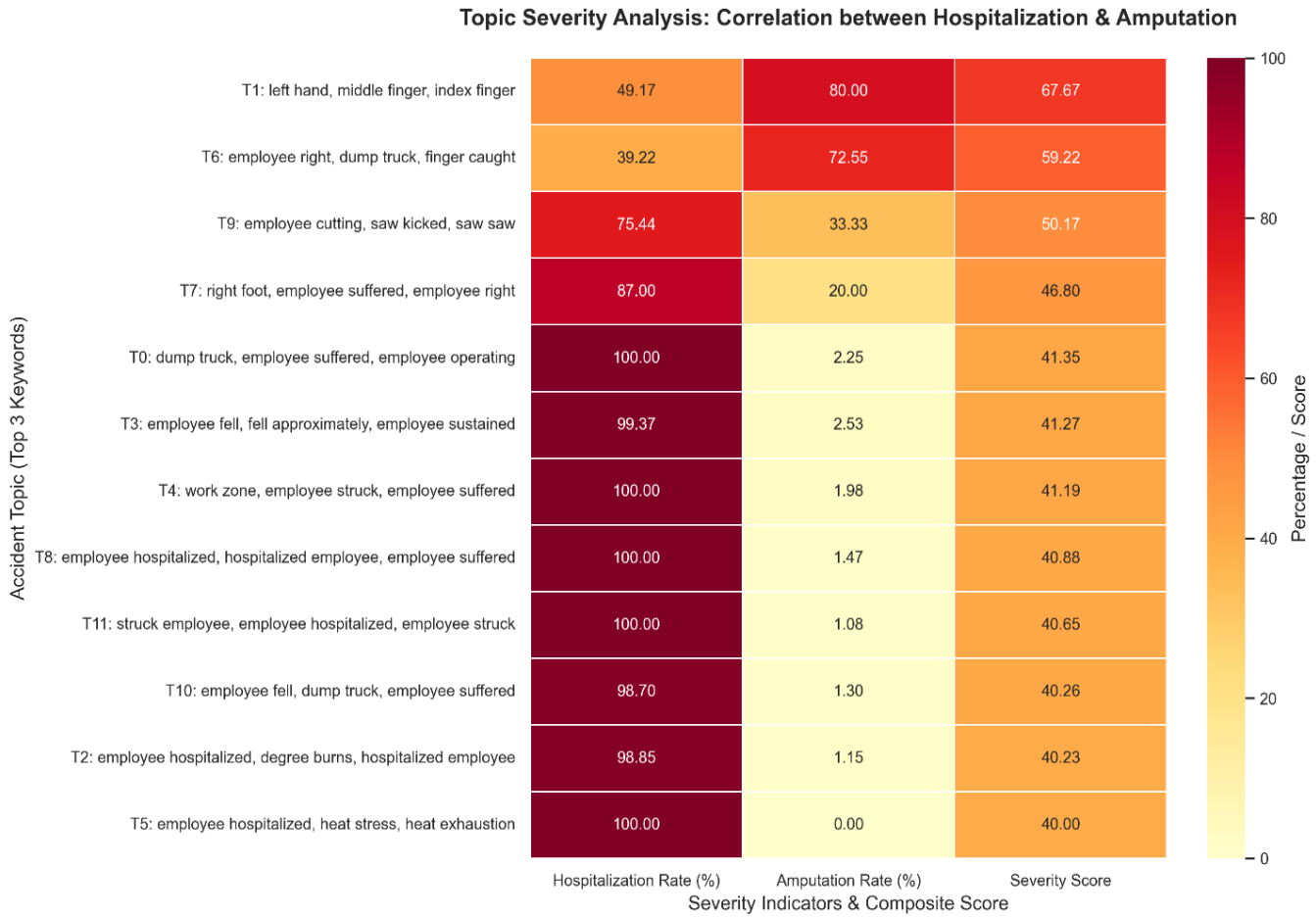


Fig. 2. Heatmap analysis of the severity of accident topic cluster

Table 5. Per-topic absolute case counts and severity scores

Topic	N	n_{hosp}	n_{amp}	Hosp. rate	Amp. rate	Severity score	Sample size
T1	120	59	96	49.17%	80.00%	67.67	OK (N≥30)
T6	51	20	37	39.22%	72.55%	59.22	OK (N≥30)
T9	57	43	19	75.44%	33.33%	50.17	OK (N≥30)
T7	100	87	20	87.00%	20.00%	46.80	OK (N≥30)
T0	89	89	2	100.00%	2.25%	41.35	OK (N≥30)
T3	158	157	4	99.37%	2.53%	41.27	OK (N≥30)
T4	101	101	2	100.00%	1.98%	41.19	OK (N≥30)
T8	68	68	1	100.00%	1.47%	40.88	OK (N≥30)
T11	93	93	1	100.00%	1.08%	40.65	OK (N≥30)
T10	77	76	1	98.70%	1.30%	40.26	OK (N≥30)
T2	87	86	1	98.85%	1.15%	40.23	OK (N≥30)
T5	50	50	0	100.00%	0.00%	40.00	OK (N≥30)

Table 6. Severity-weight sensitivity analysis

Weight split (Hosp/Amp)	Spearman ρ	Kendall τ	Top-5 overlap	Top-5 critical topics
40/60 (baseline)	1.000	1.000	100%	T1, T6, T9, T7, T0
30/70	0.993	0.970	80%	T1, T3, T6, T7, T9
50/50	0.986	0.946	100%	T0, T1, T6, T7, T9
25/75	0.993	0.970	80%	T1, T3, T6, T7, T9
60/40	0.161	0.152	40%	T0, T1, T3, T4, T8
70/30	-0.476	-0.303	20%	T0, T3, T4, T8, T11

To directly address concerns that high percentage rates in small clusters could be sample-size artifacts, Table 5 reports the absolute case counts underlying every percentage cited above and elsewhere in this paper.

The smallest topic contains N=50 cases (T5, heat stress) and the largest N=158 (T3); no topic falls below N=30. The high hospitalization rates observed in several topics (100% in T0, T4, T5, T8, and T11) therefore reflect genuine severity patterns within those hazard categories rather than small-sample artifacts.

Sensitivity Analysis of the Severity Weighting Scheme: To confirm that the 0.40/0.60 weighting does not introduce arbitrary ranking instability, we recomputed the composite Severity Score under six weight pairs and compared the resulting Top-5 critical-topic rankings and rank correlations against the 0.40/0.60 baseline. Table 6 summarizes the severity-weight sensitivity results across the six weighting schemes.

Rank correlations across all weight schemes that continue to prioritize irreversible disability ($w_{amp} \geq 0.50$) remain near-perfect ($\rho \geq 0.986$), and the identity of the Top-5 critical clusters is stable (80–100% overlap). Rankings only destabilize

once the weighting reverses priority toward hospitalization (60/40, 70/30), which is expected since that direction contradicts the injury-severity rationale for weighting amputation more heavily. This confirms the 40/60 split does not introduce arbitrary instability into the top-risk-cluster identification.

4.6. Topic Semantic Similarity Network Analysis of Accident Causes

To characterize semantic relationships among accident topics, Table 7 quantifies the hub roles of the topic clusters through centrality metrics. Fig. 3 provides a visual interaction network diagram showing semantic relationships among these accident topics.

Comparing these results with the co-occurrence network analysis derived from the traditional LDA model highlights a methodological difference. In previous studies based on LDA [3], networks were built from unweighted co-occurrence frequencies of topics within the same narrative report. As acknowledged in prior research, this unweighted approach easily introduces bias and leads to the overcounting of interconnected pairs. Thus, the LDA network identified high-level and abstract factors such as supervision dereliction of duty (highest centrality value of 1.000), weak safety awareness, and poor construction environment as the central hubs. Mathematical network theory demonstrates that incorporating continuous weighted edges is essential to mitigate the measurement bias inherent in simple frequency counts [16]. In contrast, the SBERT-based Topic Semantic Similarity Network in this study constructs the network using Cosine similarity between the dense semantic cluster centroids (utilizing a similarity threshold ≥ 0.80). This continuous vector space approach eliminates the unweighted co-occurrence bias, more directly reflecting the semantic distance between accident mechanisms; as noted in the Methodology, this network measures semantic proximity between topic centroids and should not be read as directly observed co-occurrence or causal interaction. Practically, the resulting SBERT network graph (Fig. 3) clearly shows Node T7 (Foot injury) located at the central position with the highest Degree Centrality (0.818), demonstrating the thematic convergence of numerous impact risks and heavy equipment incidents like T0 (Dump truck) and T11 (Collision).

Table 7. Topic semantic similarity network centrality metrics (Similarity threshold ≥ 0.8)

Network role	Topic	Degree centrality	Betweenness centrality	Interpretation
Key hub	T7: right foot, employee suffered	0.818	0.181	Most connected topic in the network and the primary bridge linking multiple accident themes.
Operational hub	T0: dump truck, employee operating	0.727	0.078	Highly connected operational incident topic with a moderate role in connecting other topics.
Collision hub	T11: struck employee, employee hospitalized	0.727	0.106	Strongly connected topic that serves as an intermediary between several accident clusters.
Isolated node	T5: employee hospitalized, heat stress	0.000	0.000	No direct connections to other topics, indicating a distinct and independent accident category.

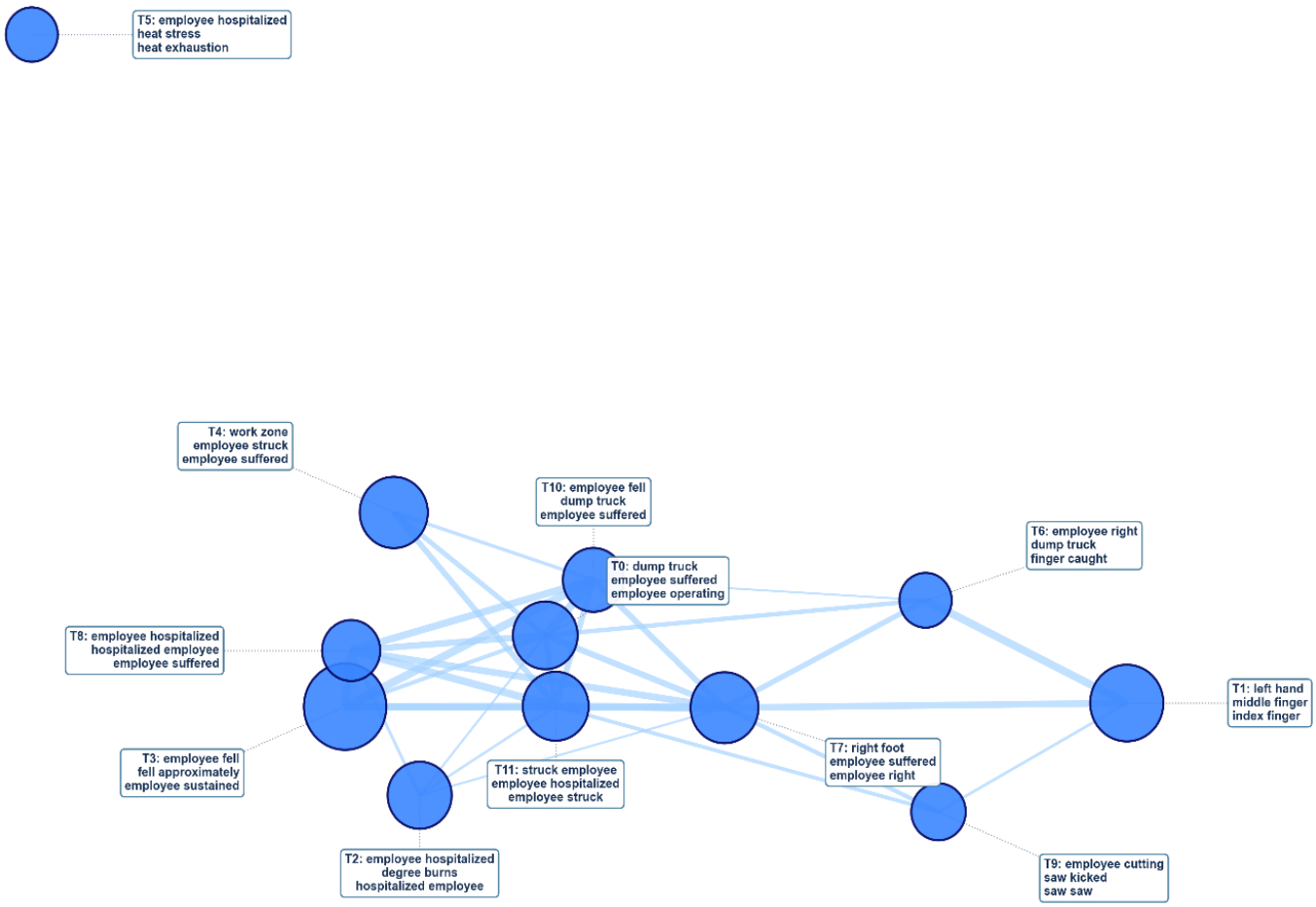


Fig. 3. Topic semantic similarity network graph representing semantic relationships among accident topics

Table 8. Topic semantic similarity network threshold sensitivity analysis

Threshold	Edges	Density	Spearman ρ vs. 0.80	Top-degree hub
0.65	59	0.894	0.554	T0 (1.000)
0.70	53	0.803	0.539	T2 (1.000)
0.75	44	0.667	0.734	T7 (0.909)
0.78	34	0.515	0.916	T7 (0.909)
0.80 (baseline)	30	0.455	1.000	T7 (0.818)
0.82	26	0.394	0.926	T7 (0.727)
0.85	22	0.333	0.916	T7 (0.727)
0.90	12	0.182	0.729	T8 (0.455)

Threshold Sensitivity Analysis: To justify the choice of 0.80 as the similarity threshold for network construction, we reconstructed the network across eight thresholds and compared network density and degree-centrality rank stability against the 0.80 baseline. Table 8 reports the corresponding threshold-sensitivity results for network density and degree-centrality rank stability.

Network density decreases monotonically as

the threshold increases. Degree-centrality rank correlation against the 0.80 baseline is high and stable across 0.75–0.85 (ρ=0.734–0.926), and Topic T7 (right-foot injury) is the top-degree hub at every threshold in that range. We selected 0.80 (30 edges, density = 0.455) because it sits at the inflection point where background noise is filtered out without fragmenting the core network topology.

5. Discussion

5.1. Methodological implications

From a Data Science perspective, our experimental results indicate that the SBERT continuous semantic embedding method addresses several limitations of the traditional LDA representation. We observed this improvement across three main fronts:

1. Targeted Resource Allocation: Cross-tabulated data (Fig. 2, Table 5) allows engineers to quantify risks rather than relying on judgment. For Topic T6 (classified as a root-cause topic involving equipment/tailgate pinch mechanisms) alongside Topic T1 (classified as an outcome descriptor capturing severe hand/finger injuries), both of which exhibit an extremely high amputation risk (over 70%) (Table 4), resources can be prioritized for specialized cut-resistant protective gloves and training on dump truck tailgate locking procedures. Conversely, for motor vehicle collision groups (Topic T0, T3, T4) with catastrophic hospitalization rates nearing 100%, project management must deploy rigid barriers, attenuators, and establish safety zones around large construction machinery.

2. Contextual Polysemy Resolution: Field report language typically contains polysemous words. The LDA algorithm completely confuses the word left in the physical injury left hand with the movement verb left the roadway. SBERT creates dynamic embedding vectors based on surrounding words, helping separate these two contexts into different topics and reduce semantic ambiguity.

3. Density-Based Outlier Robustness Check: Traditional probabilistic models like LDA are mathematically forced to assign every noisy token into a topic distribution, which often degrades the overall coherence of the extracted themes. Our primary SBERT + KMeans pipeline instead guarantees a fully assigned 12-topic partition with zero documents dropped, ensuring no accident record is silently excluded from the severity and network analyses. As a complementary check on this design choice, we separately evaluated a density-based clustering algorithm (HDBSCAN, min_cluster_size=15) on the same SBERT embedding space; this identified 105 reports (9.99%) as density-based outliers (Topic −1), and a chi-square test of independence found no evidence that this outlier set is biased with respect to hospitalization status ($X^2=0.041$, $p=0.839$) or amputation status ($X^2=0.121$, $p=0.729$). These results provide no statistical evidence that retaining all 1,051 reports in the primary analysis systematically changes the hospitalization or amputation profile of the dataset [13].

4. Semantic Synonymy Grouping and Shorthand Tolerance: Instead of treating car, semi-truck, and vehicle as independent dimensions like LDA, SBERT embeds them in a semantically related region that supports a unified heavy-equipment collision topic. Furthermore, the algorithm smoothly recognizes field abbreviations like approx. or compound words like fractured/chipped that older algorithms often misclassify as junk tokens. Representing text as dense embedding vectors also supports downstream network-analysis tasks such as the topic semantic similarity network used here, without the sparse matrix risks inherent to LDA.

5.2. Practical implications in construction

The transition from the Bag-of-Words approach to contextual embeddings marks a step from descriptive analytics to actionable risk mitigation:

1. Targeted Resource Allocation: Cross-tabulated data (Fig. 2, Table 5) allows engineers to quantify risks rather than relying on judgment. For Topic groups T1 and T6, which carry an extremely

high amputation risk (over 70%) and are classified as root-cause topics tied to specific equipment mechanisms (Table 4), resources can be prioritized for specialized cut-resistant protective gloves and training on dump truck tailgate locking procedures. Conversely, for motor vehicle collision groups (Topic T0, T3, T4) with catastrophic hospitalization rates nearing 100%, project management must deploy rigid barriers, attenuators, and establish safety zones around large construction machinery.

2. Prioritizing Interventions via Network Centrality: The topic semantic similarity network graph shows that Topic T7 (foot injuries) has the highest degree centrality at the 0.80 baseline (0.818) and remains the top-degree hub across the 0.75–0.85 threshold range; its baseline betweenness centrality is 0.181 (see the Threshold Sensitivity Analysis above), indicating it shares the most thematic overlap with other risk topics in the network. This suggests—though it does not, on its own, causally prove — that interventions targeting foot-related hazards (e.g., mandating ASTM-certified steel-toed boots combined with traffic routing controls) could plausibly reduce thematic co-occurrence with several other risk categories simultaneously. We frame this as a data-driven prioritization signal for safety investment rather than a demonstrated causal or cost-reduction guarantee, since the underlying network measures semantic similarity between topic centroids rather than observed co-occurrence or causal interaction within individual incidents (see Methodology, Step 4).

6. Limitations

To maintain analytical consistency with the baseline established by Do et al. [8], our benchmarking process relies primarily on the traditional Latent Dirichlet Allocation (LDA) model. This choice is deliberate because it provides the foundational baseline against which advanced contextual frameworks can be evaluated. By aligning our comparative assessment with this widely recognized reference point, we isolate the specific performance gains, such as superior syntax preservation and polysemy resolution, attributed to the contextual embedding architecture implemented here, rather than introducing extraneous variables that might obscure the findings.

As noted in the Methodology, the dataset is drawn from OSHA's Severe Injury Reporting Program, which only captures incidents meeting its mandatory-reporting threshold (in-patient hospitalization, amputation, or loss of an eye). Consequently, less severe work-zone incidents are systematically excluded from this dataset, and the findings should be interpreted as characterizing severe highway work-zone accidents specifically, rather than the full spectrum of work-zone incidents regardless of severity. Reporting quality and detail may also vary across individual OSHA inspectors and jurisdictions, which we cannot independently verify from the narrative text alone.

Despite notable contributions, the study still has some further limitations. First, narrative text data depends on the quality and subjectivity of OSHA inspectors at the time of recording the incident. Second, although SBERT significantly improved the clustering process, the current study only establishes retrospective correlations based on an unsupervised machine learning algorithm combined with cross-tabulation analysis. Third, the root-cause vs. outcome-descriptor classification proposed in Table 4 reflects the authors' interpretive judgment of topic keyword composition rather than a validated, independently audited coding scheme; future work could strengthen this classification through expert inter-rater review of the underlying narratives. Building a supervised machine learning classifier that directly uses semantic embedding vectors to predict the probability of hospitalization/amputation as soon as report text is inputted represents great potential for future research.

To avoid stochastic drift from canonical UMAP projections [17] and preserve partition parity

with the twelve-topic LDA baseline, we deployed deterministic TruncatedSVD.

7. Conclusion

This study conducted a comprehensive comparative experiment between the traditional topic model (LDA) and the advanced contextual language model (SBERT-KMeans) on a dataset of 1,051 OSHA highway construction accident reports (January 2015 to February 2022). The research results yield two major contributions.

Academically, the study provides clear evidence of the measurement bias phenomenon in traditional mathematical evaluation metrics. Although LDA scores higher on traditional unigram-based metrics, the proposed embedding-based coherence metric (C_{emb}) is higher for SBERT-KMeans by 35.5%, supporting a stronger phrase-aware representation in understanding syntax, resolving polysemy, and extracting intact technical compound terms (n-grams) that the Bag-of-Words method cannot achieve. Five-seed stability testing further shows that the reported metric differences are stable across the tested random seeds.

Practically for civil engineering, through a hybrid analytical framework combining SBERT with severity correlation analysis and topic semantic similarity network graphs, the study successfully quantified the most severe accident topics (dump truck equipment pinches, circular saw kickbacks) and located critical risk hubs, with the composite severity ranking confirmed robust across multiple alternative weighting schemes. These quantitative findings shift the safety management process on highway construction sites from qualitative frequency assessment to targeted risk control prioritization. Engineers and project managers now possess a solid database to plan protective equipment, train personnel, and design physical barriers to help protect workers against the highest-priority severe incident categories.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Generative AI

During the preparation of this work, the authors used a generative AI language tool (Gemini, free version) solely to improve the readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article. No generative AI or AI-assisted technology was used to generate scientific content, to perform the analytical derivations or numerical simulations, or to produce the results, figures, or conclusions reported in this work.

References

[1] C. Nnaji, A.A. Karakhan, J. Gambatese, H.W. Lee. (2020). Case study to evaluate work-zone safety technologies in highway construction. *Practice Periodical on Structural Design and Construction*, 25(3), 05020004. https://doi.org/10.1061/(ASCE)SC.1943-5576.0000498

[2] S. Sabeti, O. Shoghli, M. Baharani, H. Tabkhi. (2021). Toward AI-enabled augmented reality to enhance the safety of highway work zones: Feasibility, requirements, and challenges. *Advanced Engineering Informatics*, 50, 101429. https://doi.org/10.1016/j.aei.2021.101429

[3] C.-F. Chi, C.-C. Yang, Z.-L. Chen. (2009). In-depth accident analysis of electrical fatalities in the construction industry. *International Journal of Industrial Ergonomics*, 39(4), 635–644. https://doi.org/10.1016/j.ergon.2007.12.003

[4] A. Chokor, H. Naganathan, W.K. Chong, M. El Asmar. (2016). Analyzing Arizona OSHA injury reports using unsupervised machine learning. *Procedia Engineering*, 145, 1588–1593. https://doi.org/10.1016/j.proeng.2016.04.200

[5] Y. Suh. (2021). Sectoral patterns of accident

process for occupational safety using narrative texts of OSHA database. *Safety Science*, 142, 105363. https://doi.org/10.1016/j.ssci.2021.105363

[6] A.J.-P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman. (2016). Application of machine learning to construction injury prediction. *Automation in Construction*, 69, 102–114. https://doi.org/10.1016/j.autcon.2016.05.016

[7] H. Xu, Y. Liu, C.-M. Shu, M. Bai, M. Motalifu, Z. He, S. Wu, P. Zhou, B. Li. (2022). Cause analysis of hot work accidents based on text mining and deep learning. *Journal of Loss Prevention in the Process Industries*, 76, 104747. https://doi.org/10.1016/j.jlp.2022.104747

[8] Q. Do, T. Le, C. Le. (2024). Uncovering critical causes of highway work zone accidents using unsupervised machine learning and social network analysis. *Journal of Construction Engineering and Management*, 150(3), 04023168. https://doi.org/10.1061/JCEMD4.COENG-13952

[9] P. Li, Y. He, Z. Li. (2022). Study on influencing factors of construction workers' unsafe behavior based on text mining. *Frontiers in Psychology*, 13, 886390. https://doi.org/10.3389/fpsyg.2022.886390

[10] M. Grootendorst. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv, arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

[11] N. Reimers, I. Gurevych. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992*. https://doi.org/10.18653/v1/D19-1410

[12] R. Egger, J. Yu. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7, 886498. https://doi.org/10.3389/fsoc.2022.886498

[13] R. Albalawi, T.H. Yeap, M. Benyoucef. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3, 42. https://doi.org/10.3389/frai.2020.00042

[14] R.J.G.B. Campello, D. Moulavi, J. Sander. (2013). Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining (PAKDD 2013). Springer: Vol. 7819, pp. 160–172*. https://doi.org/10.1007/978-3-642-37456-2_14

[15] A. Pinto, I.L. Nunes, R.A. Ribeiro. (2011). Occupational risk assessment in construction industry—Overview and reflection. *Safety Science*, 49(5), 616–624. https://doi.org/10.1016/j.ssci.2011.01.003

[16] M.E.J. Newman. (2004). Analysis of weighted networks. *Physical Review E*, 70, 056131. https://doi.org/10.1103/PhysRevE.70.056131

[17] L. McInnes, J. Healy, J. Melville. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv, arXiv:1802.03426.* https://doi.org/10.48550/arXiv.1802.03426