CCC 2025 **Proceedings of the Creative Construction Conference** (2025)
Edited by: Miroslaw J. Skibniewski & Miklós Hajdu & Žiga Turk
https://doi.org/10.22260/CCC2025/0049

AUTOMATED CLASSIFICATION OF CONSTRUCTION WORKERS' WALKING-RELATED ACCIDENTS IN ACCIDENT CASE DATABASE

Ho-Young Lee, Jongwoo Cho, Tae Wan Kim

Incheon National University, Incheon, Korea

Abstract

Construction workers' walking-related accidents, which refer to incidents occurring while construction workers transition between work areas rather than during task execution, account for 31% of construction site incidents and often result from workers' lack of awareness of site hazards. Despite their high frequency, existing accident classification methods primarily focus on site characteristics, accident types, or worker attributes, making it difficult to analyze these accidents sufficiently. Since these methods are designed to assess incidents that occur during task execution, they often fail to capture the distinct nature of walking-related accidents.

This study proposes a structured methodology for the automated classification of walking-related accidents using Natural Language Processing (NLP) techniques. We define classification criteria tailored to walking-related accidents on construction sites and apply a supervised learning approach using Bidirectional Encoder Representations from Transformers (BERT) fine-tuning and machine learning classifiers. This method enables precise identification and categorization of walking-related accidents from textual accident data.

To validate this approach, we conducted a case study using the Korean construction accident dataset. The study confirmed that the proposed classifier model achieved a high precision score of 0.93. Through 1,360 labeled cases, we identified 7,720 walking-related accident cases from 23,469 reports, demonstrating the effectiveness of our classification framework in extracting walking-related accident case data.

This study contributes to improving construction site safety management by shifting the focus from conventional task-centered accident classification to the specific classification of walking-related accidents. Furthermore, it enhances the automation of the classification process, enabling more efficient handling of large-scale accident case data and addressing the limitations of expert-dependent classification methods. The classified walking-related accident data supports worker-friendly safety education, strengthens risk assessment for walking hazards, and provides valuable insights for developing proactive accident prevention strategies.

Keywords: construction safety, natural language processing, supervised learning, text-based analysis, walking-related accident.

Peer-review under responsibility of the scientific committee of the Creative Construction Conference 2025.

1. Introduction

The construction industry has consistently recorded high rates of occupational accidents and fatalities. Despite ongoing efforts to improve site safety, a substantial number of incidents continue to occur, indicating limitations in the prevailing safety management frameworks. These observations suggest a need for more refined and activity-specific approaches to accident analysis.

Conventional construction accident classification systems typically categorize incidents by work type, accident mechanism, or worker attributes. Although these frameworks are effective for analyzing task-related accidents, they often fail to address accidents that occur while workers are walking within the site, outside the context of active task execution. Walking-related accidents, which refer to incidents that

Corresponding author email address: cucumber15@naver.com

occur during intentional walking between areas on a construction site, account for a considerable proportion of all accidents but are frequently overlooked in standard classification systems [1].

The exclusion of walking-related accidents from formal safety analyses presents a significant gap in risk management. Workers often encounter hazards such as uneven surfaces, temporary walkways, or obstructed routes during walking, yet these risks are not adequately covered by task-focused safety protocols [2]. Without explicit recognition of these hazards, both hazard identification and preventive planning may remain incomplete.

To address this issue, the present study proposes a classification framework that distinguishes walking-related accidents from task-related ones. As illustrated in Figure 1, the framework classifies accidents based on whether the incident was caused by walking activity itself or occurred during the execution of a construction task. This distinction allows for the systematic identification of walking-related cases from unstructured accident narratives.

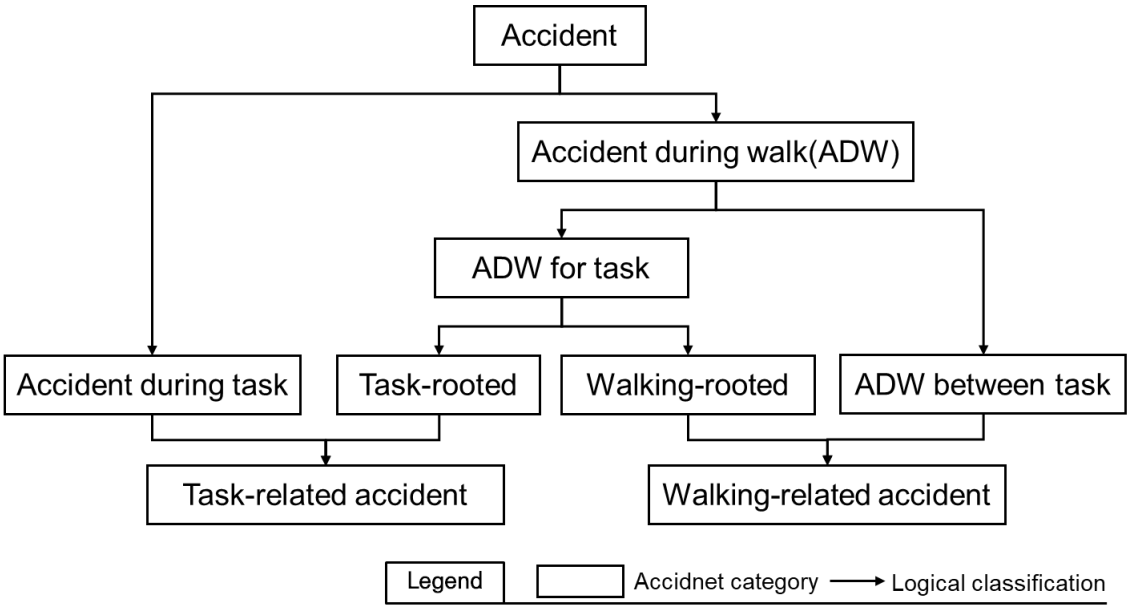

Fig. 1. Definition of walking and task related accidents.

Recognizing walking as an independent behavioral context also has practical implications for construction safety training. Instruction that emphasizes walking-related hazards, such as navigating unstable surfaces or narrow passageways, can reflect site conditions more accurately than generalized task-based guidelines [2]. Training programs aligned with workers' actual walking behaviors may enhance situational awareness and reduce the likelihood of accidents.

This study contributes to both theoretical and practical advancements in construction safety by identifying walking as a distinct category of accident causation. The proposed classification enables targeted assessment of walking-related risks, supports mobility-conscious site planning, and offers a scalable method for analyzing large volumes of accident case data. Furthermore, the findings provide a basis for developing worker-oriented safety interventions that complement existing task-centered strategies [3,4,5].

.

2. Literature Review

Recent advancements in natural language processing (NLP) and text mining have significantly influenced construction safety research, especially in accident analysis [2,3,4,7]. Techniques such as transformer-based models such as BERT have enabled the extraction of meaningful patterns from unstructured accident narratives. These methods facilitate efficient categorization and interpretation of large-scale construction accident data.

Most prior studies, however, focus on traditional labels like accident type (e.g., falls, electrocution) or site characteristics, limiting the behavioral and situational context captured—particularly for transitional activities such as walking between work areas. Consequently, walking-related hazards are often underrepresented in construction safety analyses.

Latent Class Clustering Analysis (LCCA) has been applied to reveal hidden patterns in construction accident data. One study identified a cluster resembling walking-related accidents, accounting for approximately 31% of total incidents [1]. However, clustering relies on probabilistic similarity rather than explicit labeling, making definitive identification of walking-related accidents difficult. This highlights the need for classification approaches that directly and accurately isolate walking-related incidents.

Classification frameworks like the Occupational Injury and Illness Classification System (OIICS) include categories such as slips, trips, and falls (STF), particularly falls on the same level (STFL) [6]. While STFL captures incidents on flat surfaces, it excludes vertical transitions such as stairs or ladders and does not specifically differentiate walking as an intentional activity. In contrast, this study defines walking-related accidents as those involving intentional worker relocation—whether horizontal or vertical—thus encompassing a broader and more context-sensitive range of incidents.

Overall, prior research demonstrates the promise of text-based classification for construction accident analysis but also reveals significant gaps. Addressing these gaps requires a walking-centered classification framework that can accurately capture the nuanced contexts of worker mobility on construction sites.

3. Research Goals and Objectives

This study aims to address limitations in existing construction accident classification systems that have traditionally focused on task-related activities, leaving walking-related incidents largely unexamined. It proposes a walking-centered perspective by developing a dataset that highlights walking-related accidents—those that occur during intentional worker relocation on-site.

The core objective is to automatically extract walking-related accidents from construction accident narratives and construct a labeled dataset. This reflects the recognition that intentional site walking, though frequent, is a poorly documented risk factor. By systematically identifying such cases, the research contributes to more practical safety analysis and training applications.

To realize this goal, the study first establishes a clear definition of walking-related accidents and a labeling criterion that considers both the intent behind the worker's walking and the situational context of the incident. In this study, accidents were classified based on whether the cause and context were primarily rooted in task execution or in worker walking. As illustrated in Figure 1, task-related accidents include both those occurring directly during task performance and those caused by task-related factors. Walking-related accidents, in contrast, refer to those rooted in the act of intentional site walking—either due to walking-related factors or occurring during transitions between tasks. This classification served as the foundational labeling criterion for the supervised learning approach.

The study then applies a fine-tuned BERT model to embed the textual accident narratives and uses various machine learning classifiers to evaluate the performance of supervised classification. By comparing model performance, the research examines whether supervised learning can reliably distinguish walking-related accidents from other types.

This framework provides a methodological foundation for recognizing walking-related safety issues in construction, with potential applications in mobility-aware safety training, risk evaluation, and the development of targeted intervention strategies.

4. Methodology

This study utilized accident narrative data from Korea's "CSI Accident Cases" database, which provides detailed textual descriptions of construction site incidents. A sequential research process was followed to construct a walking-related accident dataset through automated classification techniques.

Accident narratives were first collected using dynamic web crawling techniques. A crawler was developed to extract textual content from publicly available archives. After collection, preprocessing was performed to remove noise and standardize text, including the elimination of special characters, correction of spacing errors, and exclusion of irrelevant or insufficiently detailed cases.

Recognizing that construction narratives often blur the distinction between work and walking, a three-class labeling scheme was adopted: "working-related accident," "walking-related accident," and "unclassifiable." Labeling was conducted manually using defined criteria, focusing on whether intentional site walking meaningfully contributed to the accident. For instance, slipping while carrying rebar was considered walking-related, while injuries caused solely by the material were not.

The labeled dataset, comprising 1,360 cases, was randomly divided into training and test sets at a 7:3 ratio. To embed the narratives, the "bert-base-multilingual-cased" model was fine-tuned on the labeled data, capturing domain-specific linguistic patterns.

Various machine learning classifiers were then trained on the BERT embeddings, including linear models (SVM, LR), tree-based models (RF, GB, XGB, LGBM, CB, EX, HGBC, DT), a neural network model (MLP), a distance-based model (KNN), a Bayesian model (NB), and a dimension-reduction model (LDA). Default hyperparameters were used to maintain comparability across models.

Performance evaluation prioritized precision, particularly for the walking-related accident class, to ensure that predicted cases genuinely reflected walking-related risks. The classifier with the highest precision—Naive Bayes—was applied to the broader database, resulting in the construction of a large-scale walking-related accident dataset.

5. Presentation of Research

To evaluate the proposed classification approach, multiple machine learning models were trained using BERT-based embeddings of labeled accident narratives. Performance was assessed using precision, recall, F1-score, and accuracy, with emphasis on the precision of the walking-related class due to its relevance to this study. Table 1 summarizes the classification performance of all tested models using BERT-based embeddings. Among these, the Naive Bayes classifier achieved the highest precision for walking-related accident detection.

Table 1. Table 1. Evaluation results of trained classifiers.

	Accuracy	$Precission_w$	$Recall_w$	$F1\text{-}Score_w$
SVM	79.16667	92.56757	78.28571	84.82972
LR	78.92157	91.94631	78.28571	84.5679
RF	78.43137	92.61745	78.85714	85.18519
GB	78.67647	91.44737	79.42857	85.01529
XGB	78.92157	91.44737	79.42857	85.01529
LGBM	78.67647	92	78.85714	84.92308
CB	78.67647	92	78.85714	84.92308
EX	78.67647	92.56757	78.28571	84.82972
HGBC	78.43137	92.61745	78.85714	85.18519
DT	79.65686	93.24324	78.85714	85.44892
MLP	78.43137	91.94631	78.28571	84.5679
KNN	77.69608	91.89189	77.71429	84.21053
NB	73.03922	93.70629	76.57143	84.27673
LDA	79.16667	90.32258	80	84.84848

Among all models, Naive Bayes achieved the highest precision in identifying walking-related accidents. Despite its simplicity, it consistently detected walking-related language patterns in the text. Based on this result, the NB classifier was selected and applied to the full dataset of 23,469 accident cases. Through this classification, 7,720 cases were identified as walking-related, forming a comprehensive dataset for walking-focused safety analysis.

To visualize the classification landscape, t-distributed stochastic neighbor embedding (t-SNE) was applied to the BERT embeddings. The resulting two-dimensional plot showed that the three classes—working-related, walking-related, and unclassifiable—were reasonably well-separated. Walking-related cases formed a distinguishable cluster, indicating that the classifier captured distinct semantic features associated with worker walking. Figure 2 visualizes the classification results using t-SNE, illustrating the distribution of BERT embeddings across the three classes. The clear separation suggests that the model effectively learned class-distinct features.

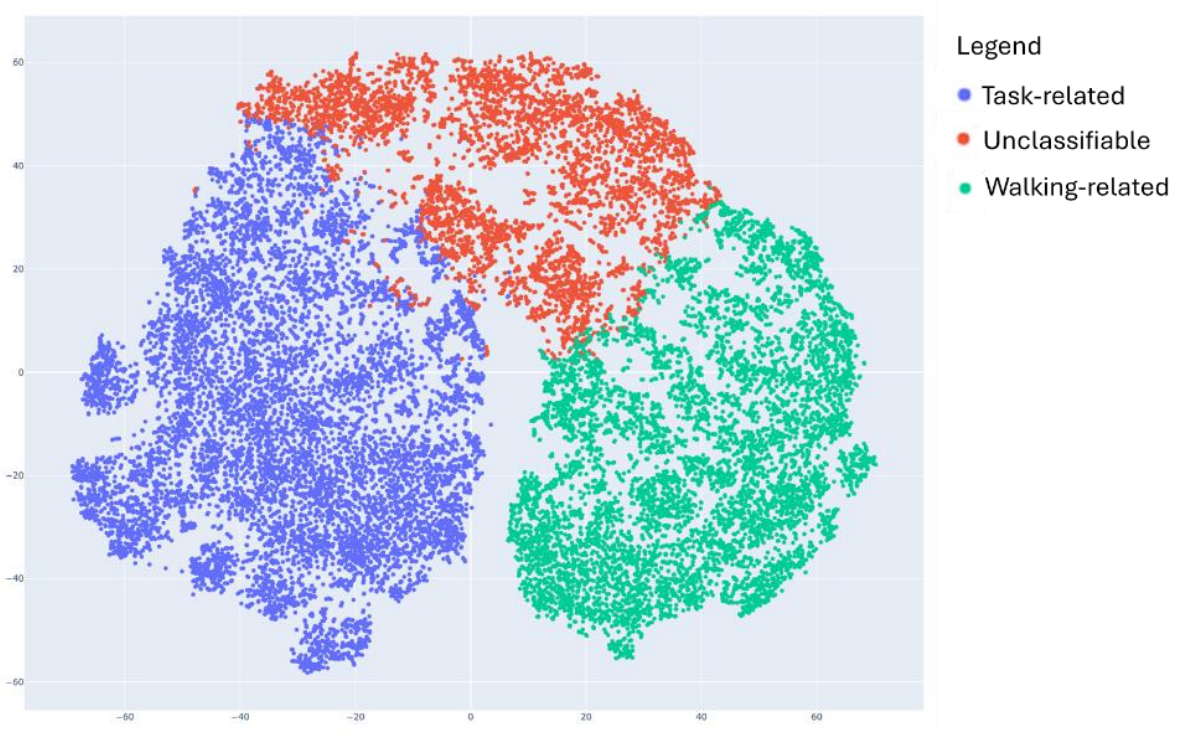

Fig. 2. t-SNE Visualization of Classification Results

These results support the validity of the classification framework and highlight the practical value of embedding-based text analysis for construction safety research. The visual and quantitative evidence suggests that walking-related accidents can be reliably isolated as a distinct category within unstructured accident data.

6. Discussion

This study demonstrates the feasibility of classifying walking-related accidents from unstructured construction narratives using a walking-centered approach. Traditional classification systems have largely focused on task-related contexts, leaving transitional hazards underexplored. By identifying walking-related incidents as a distinct category, this research contributes a complementary framework to existing taxonomies.

The classification approach has practical implications for safety training and site planning. By highlighting walking-related risks, it supports more targeted interventions, including hazard recognition during navigation and layout planning that considers safe walking paths. Academically, it broadens the analytical perspective of construction safety by introducing behavioral context into accident analysis.

While the study showed strong performance using Naive Bayes and BERT-based embeddings, it was limited to a single embedding model and a relatively simple classifier. Future work could apply more advanced models or compare multiple embeddings to improve robustness.

Additionally, the dataset constructed in this study opens avenues for future research on mobility hazards. These include spatial risk analysis, walking-specific risk assessment tools, and studies of vulnerable worker groups. As the dataset grows, it may also support predictive modeling for walking-related accident prevention.

7. Conclusion

This study proposed a classification framework to identify walking-related accidents in construction narratives using a walking-centered approach, fine-tuned BERT embeddings, and supervised machine learning. By recognizing worker walking as a meaningful context for accidents, the study expands on task-focused classification schemes and addresses overlooked transitional risks.

A labeled dataset was created through manual annotation, and a Naive Bayes classifier achieved the highest precision in identifying walking-related incidents. Applied to over 23,000 construction accident records, the model extracted 7,720 walking-related cases, demonstrating the effectiveness of this approach.

This classification scheme has potential applications in mobility-aware safety training, risk assessment, and site planning. While the approach used a single embedding model and a basic classifier, it lays the foundation for more advanced applications, including embedding comparisons and predictive modeling.

Ultimately, the study contributes a practical method for improving the understanding and management of walking-related risks in construction, offering a new perspective for accident prevention efforts.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2021R1A2C1013188).

References

[1] B. U. Ayhan and O. B. Tokdemir, “Accident Analysis for Construction Safety Using Latent Class Clustering and Artificial Neural Networks,” Journal of Construction Engineering and Management, vol. 146, no. 3, Mar. 2020, doi: 10.1061/(asce)co.1943-7862.0001762.

[2] W.-R. Chang, S. Leclercq, T. E. Lockhart, and R. Haslam, “State of science: occupational slips, trips and falls on the same level,” Ergonomics, pp. 1–23, Mar. 2016, doi: 10.1080/00140139.2016.1157214.

[3] Qiao, C. Wang, S. Guan, and L. Shuran, “Construction-Accident Narrative Classification Using Shallow and Deep Learning,” Journal of Construction Engineering and Management, vol. 148, no. 9, Sep. 2022, doi: 10.1061/(asce)co.1943-7862.0002354.

[4] J. Li and C. Wu, “Deep Learning and Text Mining: Classifying and Extracting Key Information from Construction Accident Narratives,” Applied Sciences, vol. 13, no. 19, p. 10599, Sep. 2023, doi: 10.3390/app131910599.

[5] Q. Shuang, X. Liu, Z. Wang, and X. Xu, “Automatically Categorizing Construction Accident Narratives Using the Deep-Learning Model with a Class-Imbalance Treatment Technique,” Journal of Construction Engineering and Management, vol. 150, no. 9, Sep. 2024, doi: 10.1061/jcemd4.coeng-14515.

[6] P. Duan, J. Zhou, and Y. M. Goh, “Spatial-temporal analysis of safety risks in trajectories of construction workers based on complex network theory,” Advanced Engineering Informatics, vol. 56, p. 101990, Apr. 2023, doi: 10.1016/j.aei.2023.101990.

[7] A. Shamshiri, K. R. Ryu, and J. Y. Park, “Text mining and natural language processing in construction,” Automation in Construction, vol. 158, p. 105200, Feb. 2024, doi: 10.1016/j.autcon.2023.105200.