EXTRACTING INFORMATION FROM CONSTRUCTION SAFETY REQUIREMENTS USING LARGE LANGUAGE MODEL

Si Van-Tien Tran, Nasrullah Khan, Emmanuel Charles Kimito, Akeem Pedro, Mehrtash Sotani, Rahat Hussain & Taehan Yoo
Department of Architectural Engineering, Chung-Ang University, Seoul 06974, Korea

Chansik Park
Department of Architectural Engineering, Chung-Ang University, Seoul 06974, Korea

ABSTRACT: *The construction industry has long been recognized for its complex safety regulations, which are essential to ensure the well-being of on-site employees. However, navigating these regulations and ensuring compliance can be challenging due to the volume and complexity of the documents involved. This study proposes a novel approach to extracting information from construction safety documents utilizing Large Language Models (LLM), called CSQA, to provide real-time, precise answers to queries related to safety regulations. The approach comprises three modules: (1) the construction safety investigation module (CSI) collects safety regulations for building the information needed. By leveraging a collection of safety regulation PDFs, the system follows a process of text extraction, preprocessing, and global indexing for efficient search. (2) The safety condition identification module (SCI) retrieves the CSI database; after that, the LLM, with its extensive training, processes user queries, searches the indexed regulations, and retrieves pertinent information. (3) the safety information delivery (SID) would provide the answer to the user and incorporate a feedback mechanism to further refine system accuracy based on user responses. Preliminary evaluations reveal the system's superior performance over traditional search engines, owing to its ability to grasp query context and nuances. The CSQA presents a promising method for accessing safety regulations, with potential benefits including reduced non-compliance incidents, enhanced worker safety, and streamlined regulatory consultations in construction.*

KEYWORDS: *Construction safety document, extraction, LLM.*

1. INTRODUCTION

Safety has consistently been seen as a vital concern within the construction industry. Workplace safety catastrophies can result in major loss of life and damage to property with severe repercussions (S. V.-T. Tran et al., 2023; S. V. T. Tran et al., 2021). According to the latest statistics from the Occupational Safety and Health Administration (*OSHA Fatality Report*, n.d.), the construction industry witnessed an annual total of 1,008 fatalities in 2020. Notably, falls from elevated positions constituted around thirty-three percent of these. According to data from Statistics Korea (*Construction Work | Statistics Korea*, n.d.), the construction business in South Korea accounted for more than 50% of all fatal accidents within the industry. To prevent accidents at construction sites, several scholars and professionals have demonstrated that implementing enhanced safety measures in the workplace might reduce and prevent accidents (Bao et al., 2022; S. V. Tran et al., 2022; S. V. T. Tran et al., 2022). Therein, field compliance checking is a crucial endeavor to identify non-compliance with construction safety standards, with the primary objective of safeguarding employees against potential safety events (Jeong et al., 2023; Kang et al., 2023).

Analyzing construction safety documents with natural language processing (NLP) techniques enables automatic information extraction of safety requirements. For instance, Feng and Chen (Feng & Chen, 2021) proposed a framework based on deep learning to extract event-related information (e.g., date, location, and type of accident) from accident news reports for construction safety management. Rupasinghe and Panuwatwanich (Rupasinghe & Panuwatwanich, 2021) proposed a rule-based technique for extracting information about hazards from accident reports. Baker et al. (Baker et al., 2020) suggested employing NLP (a collection of text patterns) to uncover injury precursors. These works together focused on either the study of injury and accident records or the extraction of hazard variables. Despite these studies, there is a dearth of research aimed at automatically extracting requirements from construction safety rules in order to enable field compliance. Besides, the information extraction should provide users with precise and timely responses to their inquiries within human natural language.

Large Language Models (LLM) have emerged as a game-changing technology, displaying extraordinary ability in natural language processing jobs. Incorporating LLMs into construction safety provides a distinct benefit in its capacity to customize to particular, project-centric data. This is especially important given the vast volumes of

Referee List (DOI: 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Si Tran, Nasrullah Khan, Emmanuel Charles Kimito, Akeem Pedro, Mehrtash Sotani, Rahat Hussain, Taehan Yoo, Chansik Park, *Extracting Information from Construction Safety Requirements Using Large Language Model*, pp. 761-767, © 2023 Author(s), CC BY NC 4.0, DOI 10.36253/979-12-215-0289-3.76

private paperwork that projects often require. Every project in the construction environment is unique, with its own blueprints, safety regulations, and vendor-specific rules, often encased inside PDFs and other digital forms. LLMs have the capacity to be trained or fine-tuned on project-specific datasets. Once a company uploads its confidential documents, the LLM can absorb this data, guaranteeing that when queries are posed, the solutions are general and suited to the context of that specific project's data.

This research proposes CSQA approach, a unique method for extracting construction safety documentation using Large Language Models (LLM), to answer real-time safety regulatory questions and fill the knowledge gap for industry experts. The method has three parts: (1) The construction safety investigation module (CSI) gathers building safety rules. The system uses safety regulation PDFs for text extraction, preprocessing, and global indexing for efficient search. (2) The safety condition identification module (SCI) obtains the CSI database, then the LLM analyzes user queries, examines the indexed rules, and retrieves relevant information with its thorough training. (3) Safety information delivery (SID) would address the user and offer feedback to improve system accuracy depending on user replies. Section 2 discusses the current state of construction safety information retrieval and LLM. Section 3 will present the recommended approach. The authors produce case scenarios in Section 4 to validate the approach. Subsequently, the discussion and conclusions of the study are presented.

2. LITERATURE REVIEW

2.1 Current state of construction safety information retrieval and extraction

Over the years, the construction industry, renowned for its complex projects and the resulting safety imperatives, has accrued a vast repository of safety regulations, guidelines, and best practices. Traditionally, retrieving and extracting relevant safety information was primarily a manual process (Zhong et al., 2020). Professionals frequently find themselves navigating through extensive physical binders or digital documents. This approach, while exhaustive, is fraught with difficulties. Due to the time-consuming nature of manual searches and the possibility of human error, there are frequent voids in the incorporation of vital safety directives(S. V. T. Tran et al., 2021). Moreover, the dynamic nature of construction projects, with their distinct challenges and parameters, necessitates a customized understanding of safety regulations, which manual searches cannot provide efficiently(Wu et al., 2022).

Efforts have been made since the advent of the digital age to expedite this procedure (S. V. T. Tran et al., 2021). Initially, safety information was migrated to digital databases, enabling keyword-based searches. Even though this change facilitated the retrieval process to some degree, it was not without limitations. Keyword searches frequently return many results, necessitating additional sorting to locate relevant information. The lack of contextual comprehension and the static nature of these databases provided a wealth of information without the nuanced interpretation required for specific project scenarios. For instance, Feng and Chen (Feng & Chen, 2021) proposed a framework based on deep learning to extract event-related information (e.g., date, location, and type of accident) from accident news reports for construction safety management. Rupasinghe and Panuwatwanich (Rupasinghe & Panuwatwanich, 2021) proposed a rule-based technique for extracting information about hazards from accident reports. Baker et al. (Baker et al., 2020) suggested employing NLP (a collection of text patterns) to uncover injury precursors. This context paves the way for investigating more sophisticated AI-driven methodologies capable of efficient information retrieval and contextual comprehension and understanding.

2.2 Information extraction using Large Language Model

Natural language processing allows a computer to interpret and process natural language text similarly to a person. Information extraction (IE) is a branch of natural language processing that obtains needed information from text sources. In general, there are two techniques for information extraction [11]: (1) machine learning (ML) and (2) rule-based approaches. However, research has focused on using rule-based techniques because training samples are few. Large Language Models (LLM) have transformed natural language processing, providing a game-changing answer to this problem.

Within the realm of construction, safety stands as a paramount pillar, with documentation and guidelines serving as the backbone to ensure the welfare of all stakeholders. Extracting relevant, actionable information has been a persistent challenge with the sheer volume and complexity of safety documentation. The potency of LLMs in safety information extraction lies in their ability to discern the context of a query and retrieve information that is not just relevant but also actionable. For instance, when asked about safety protocols for handling specific machinery, an LLM can sift through a vast repository of safety guidelines, pinpointing the exact procedures, precautions, and best practices.

Besides, construction safety may benefit from Large Language Models (LLMs) since they can be tailored to project-specific data, particularly given the vast volumes of private paperwork projects frequently include. Every construction project has plans, safety regulations, and vendor-specific rules, frequently in PDFs. LLMs may be trained or fine-tuned using project-specific datasets. The LLM may integrate confidential materials uploaded by a company to provide project-specific solutions to inquiries. Project secrecy and relevance are greatly affected by LLMs' project-specific customization. Traditional search engines and databases may provide general results or need substantial human labeling to identify project-specific data. LLMs automatically comprehend the context after being fine-tuned on a project's papers, ensuring that every answer meets the project's particular characteristics and criteria. This improves information retrieval accuracy and relevance and keeps sensitive project data in that context, protecting private project information.

3. METHOD

The primary purpose of developing an approach of extracting construction safety requirements using large language model. The structure and key features of the system are shown in **Figure 1**, which comprises three modules. (1) The construction safety investigation module (CSI) gathers building safety rules. The system uses safety regulation PDFs for text extraction, preprocessing, and global indexing for efficient search. (2) The safety condition identification module (SCI) obtains the CSI database, then the LLM processes user queries, examines the indexed rules, and retrieves relevant information with its thorough training. (3) Safety information delivery (SID) would address the user and offer feedback to improve system accuracy depending on user replies.

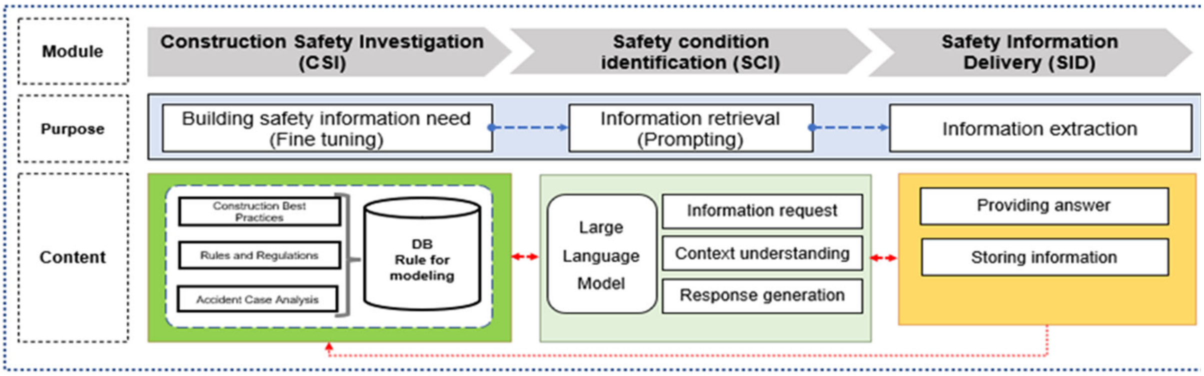


Fig. 1: Proposed approach of Extracting Construction Safety Requirements using Large Language Model

The Construction Safety Investigation (CSI) module is the foundational block in this approach, concentrating on the meticulous collection of safety regulations essential for creating the requisite information database. This module predominantly handles a variety of PDF safety regulation documents and serves as the entry point for raw safety data. The CSI module has multiple functions, including text extraction, preprocessing, and global indexing. Text extraction is crucial, as it converts the information in PDFs into a structured format. Preprocessing then entails cleaning and normalizing the extracted text to prepare it for the subsequent phases.

The Safety Condition Identification (SCI) module serves as the interface between the foundational database created by the CSI module and the user-facing delivery module in this approach. The primary responsibility of the SCI module is to interact with the CSI database and retrieve pertinent safety information based on user queries. The Large Language Model (LLM) incorporated into this module plays a crucial role, utilizing its extensive training to process and comprehend user queries in real time. The LLM examines the indexed regulations in the CSI database and retrieves relevant information, considering the context and subtleties of the user's query. Incorporating LLM into this module ensures that the retrieval process is accurate, context-aware, and efficient, providing instantaneous responses to user queries.

This approach also includes the Safety Information Delivery (SID) module, which focuses on delivering the retrieved and processed safety information to the end-user. It serves as the user interface, providing plain, concise, and pertinent responses to user queries. Beyond merely delivering information, the SID module includes a feedback mechanism that allows users to rate the accuracy and relevance of the provided answers. This user feedback is crucial for refining the system's precision and improving dependability. By perpetually incorporating user feedback, the SID module ensures that the system evolves and adapts to the users' changing requirements and preferences, maintaining its relevance and effectiveness in delivering precise construction safety information.

3.1 Prototype development

Figure 2 depicts prototype development process and tool uses for the proposed approach. The authors used Langchain, an open-source Python library for building LLM-powered applications. Utilizing LLMs with vector indexing via embedding provides a foundation for the solution. Initially, a comprehensive safety regulation database is accessed and processed to collate information predominantly housed in PDF formats. Subsequently, this information is extracted, followed by a data cleaning procedure to omit redundant elements, such as punctuation, commas, and line spaces. For this operation, a smaller LLM from the Spacy library is deployed. The information is segmented into manageable chunks to facilitate efficient filtering, aligned with the embedding model's chunk size within the embedding space. After the initial processes, the refined information is fed into a text embedding model to formulate and archive the information in a vector database, commonly referred to as a vector store, pivotal for advanced information retrieval mechanisms. The essence of the embedding model is to transmute high-dimensional textual data into a more condensed representation, aligning with the operational frameworks of LLMs, as we can't put our whole PDF textual information in user query or Prompt. For this critical transformation, the OpenAI text embedding model is employed. The formulated vector database harboring safety regulation information is then integrated into the pipeline, allowing LLMs to perform advanced retrieval of information pertinent to user queries.

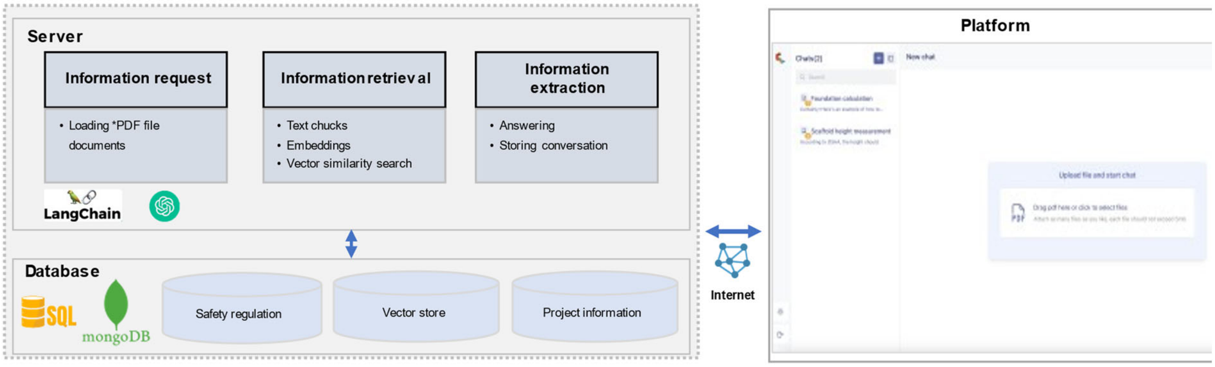


Fig. 2: System architecture

The creation of the vector store is facilitated using FAISS for efficient similarity search and clustering analysis of high-dimensional vector databases. Upon establishing the vector database, it is intricately interwoven within the operational pipeline, enabling LLMs to execute sophisticated information retrieval and interaction, utilizing OpenAI's GPT-4 through Langchain, a versatile open-source framework for building AI apps and Chatbots. The streamlined process integrates FAISS, user queries, and LLMs responses in a seamless flow. When a user initiates a question, it is directed to FAISS's sophisticated similarity search algorithm, which extracts relevant information from the vector database used by LLMs in embedding form.

4. CASE STUDY

The authors performed a case study of safety information extraction related to scaffolding during construction by implementing the CSQA approach, as illustrated in Figure 3. The extraction of safety regulations, specifically OSHA 1926 Subpart L, is pivotal in maintaining a high level of safety in construction environments where scaffolding is utilized. To do this, the authors download A Guide to Scaffold Use in the Construction Industry as a PDF file and then upload it to the CSQA prototype system. After that, the safety managers query the information related to their needs. By meticulously extracting and implementing each safety provision laid out by OSHA, construction companies can significantly mitigate the risk of scaffold-related incidents, protecting workers from falls, structural collapses, and falling objects. This process of extracting and adhering to OSHA's stringent safety regulations is essential in fostering a culture of safety within the construction industry, emphasizing the importance of the well-being of every individual on the construction site and ensuring the successful and safe completion of construction projects.

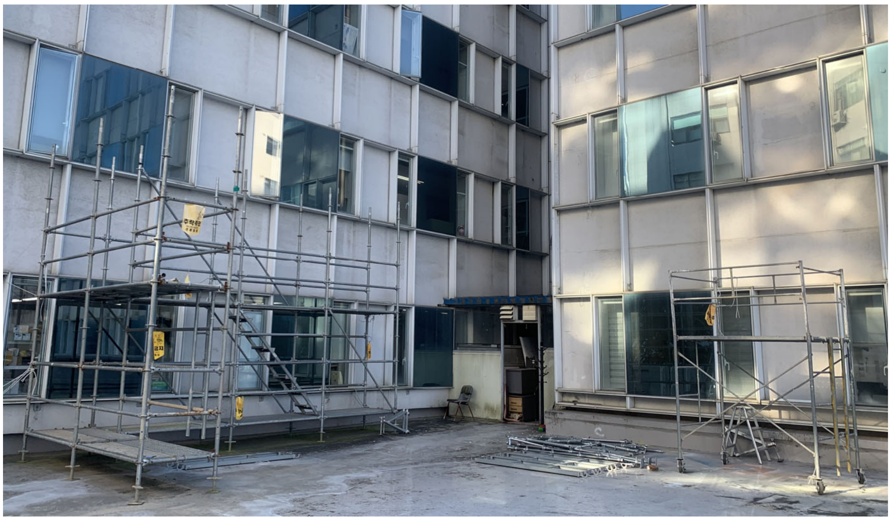
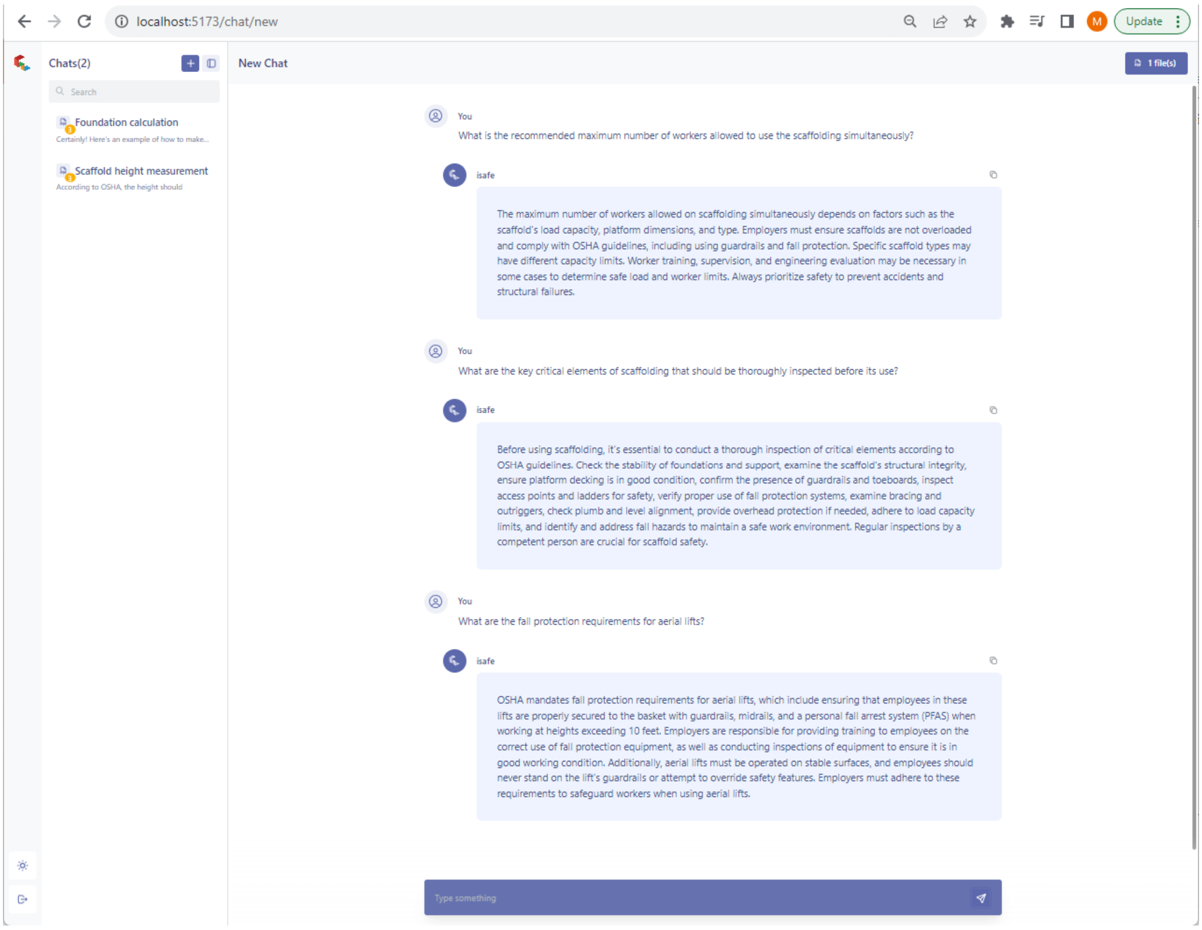


Fig. 3: The construction jobsite using both fixed and mobile scaffolding

In the case study, both fixed and mobile scaffolding were used at the construction jobsite (as illustrated in Fig. 3). To prepare for the safety inspection process, the safety manager considers some potential hazards situation that may occur during using scaffolding system. The results of the prompting process were illustrated in Fig. 4. In the scenario, the safety manager would request information about the maximum number workers allowed to use the scaffolding simultaneously. After prompting, the results of extractions were described in Fig. 4.



localhost:5173/chat/new
Update
Chats(2)
Search
Foundation calculation
Certainly! Here's an example of how to make...
Scaffold height measurement
According to OSHA, the height should
New Chat
1 file(s)
Type something

You

What is the recommended maximum number of workers allowed to use the scaffolding simultaneously?

isafe

The maximum number of workers allowed on scaffolding simultaneously depends on factors such as the scaffold's load capacity, platform dimensions, and type. Employers must ensure scaffolds are not overloaded and comply with OSHA guidelines, including using guardrails and fall protection. Specific scaffold types may have different capacity limits. Worker training, supervision, and engineering evaluation may be necessary in some cases to determine safe load and worker limits. Always prioritize safety to prevent accidents and structural failures.

You

What are the key critical elements of scaffolding that should be thoroughly inspected before its use?

isafe

Before using scaffolding, it's essential to conduct a thorough inspection of critical elements according to OSHA guidelines. Check the stability of foundations and support, examine the scaffold's structural integrity, ensure platform decking is in good condition, confirm the presence of guardrails and toeboards, inspect access points and ladders for safety, verify proper use of fall protection systems, examine bracing and outriggers, check plumb and level alignment, provide overhead protection if needed, adhere to load capacity limits, and identify and address fall hazards to maintain a safe work environment. Regular inspections by a competent person are crucial for scaffold safety.

You

What are the fall protection requirements for aerial lifts?

isafe

OSHA mandates fall protection requirements for aerial lifts, which include ensuring that employees in these lifts are properly secured to the basket with guardrails, midrails, and a personal fall arrest system (PFAS) when working at heights exceeding 10 feet. Employers are responsible for providing training to employees on the correct use of fall protection equipment, as well as conducting inspections of equipment to ensure it is in good working condition. Additionally, aerial lifts must be operated on stable surfaces, and employees should never stand on the lift's guardrails or attempt to override safety features. Employers must adhere to these requirements to safeguard workers when using aerial lifts.

Fig. 4:The results of safety information extraction

5. DISCUSSION AND CONCLUSION

The study aimed to improve construction safety by proposing an approach to extracting safety requirements using a large language model. A thorough literature review highlighted the significance of safety information retrieval and extraction. Accordingly, the safety requirements were collected and tailored for building the database, which is contained in the safety investigation module (CSI). The system uses safety regulation PDFs for text extraction, preprocessing, and global indexing for efficient search. The safety condition identification module (SCI) obtains the CSI database, then the LLM processes user queries, examines the indexed rules, and retrieves relevant information with its thorough training. Safety information delivery (SID) would address the user and offer feedback to improve system accuracy depending on user replies. Hence, the safety requirements could be extracted following the request of site employees. The authors developed the prototype of an LLM-powered application by using Langchain to validate the approach. The results show that the maximum number of workers allowed to use the scaffolding simultaneously was retrieved from the guide to scaffold use in the construction industry.

However, the research has the following limitations: (1) The study concentrates on optimizing a database of safety requirement information; however, it does not discuss the algorithm's architecture and precision. (2) The case study is only used to extract scaffolding-related information. For future studies, the authors will analyze additional accident reports and regulations to develop potential hazard situations associated with a specific activity. Additionally, the authors will concentrate on developing a system based on the proposed method. Then, we examine the effectiveness of the system with larger initiatives and more project members.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. NRF-2022R1A2B5B02002553). This research was conducted with the support of the "National R&D Project for Smart Construction Technology (No.23SMIP-A158708-04)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation.

REFERENCE

Baker, H., Hallowell, M. R., & Tixier, A. J. P. (2020). Automatically learning construction injury precursors from text. *Automation in Construction*, *118*, 103145. https://doi.org/10.1016/J.AUTCON.2020.103145

Bao, L., Tran, S. V. T., Nguyen, T. L., Pham, H. C., Lee, D., & Park, C. (2022). Cross-platform virtual reality for real-time construction safety training using immersive web and industry foundation classes. *Automation in Construction*, *143*, 104565. https://doi.org/10.1016/J.AUTCON.2022.104565

Construction Work | Statistics Korea. (n.d.). Retrieved October 6, 2022, from http://kostat.go.kr/portal/eng/pressReleases/4/5/index.board

Feng, D., & Chen, H. (2021). A small samples training framework for deep Learning-based automatic information extraction: Case study of construction accident news reports analysis. *Advanced Engineering Informatics*, *47*, 101256. https://doi.org/10.1016/J.AEI.2021.101256

Jeong, H., Shin, & Wonsang. (2023). An Analysis on the Safety Management Level of Domestic Medium Construction Companies and Its Improvement Measures. *Korean Journal of Construction Engineering and Management*, *24*(3), 20–30. https://doi.org/10.6106/KJCEM.2023.24.3.020

Kang, Jeong, H., Chae, J., & Kang, Y. (2023). Distribution of Occupational Safety and Health Management Costs (OSHMC) by Project Size and Activity Type with the Consideration of Accident Rates. *Korean Journal of Construction Engineering and Management*, *24*(4), 44–51. https://doi.org/10.6106/KJCEM.2023.24.4.044

OSHA fatality report. (n.d.). Retrieved October 6, 2022, from https://www.osha.gov/stop-falls

Rupasinghe, N. K. A. H., & Panuwatwanich, K. (2021). UNDERSTANDING CONSTRUCTION SITE SAFETY HAZARDS THROUGH OPEN DATA: TEXT MINING APPROACH. *ASEAN Engineering Journal*, *11*(4), 160–178. https://doi.org/10.11113/AEJ.V11.17871

Tran, S. V.-T., Lee, D., Bao, Q. L., Yoo, T., Khan, M., Jo, J., & Park, C. (2023). A Human Detection Approach for Intrusion in Hazardous Areas Using 4D-BIM-Based Spatial-Temporal Analysis and Computer Vision. *Buildings 2023, Vol. 13, Page 2313*, *13*(9), 2313. https://doi.org/10.3390/BUILDINGS13092313

Tran, S. V., Bao, L. Q., Nguyen, L. T., & Pedro, A. (2022). *Development of Computer Vision and BIM-cloud based Automated Status Updating for Construction Safety Monitoring*. *Nov 2022*. cloud_based_Automated_Status_Updating_for_Construction_Safety_Monitoring

Tran, S. V. T., Khan, N., Lee, D., & Park, C. (2021). A Hazard Identification Approach of Integrating 4D BIM and Accident Case Analysis of Spatial–Temporal Exposure. *Sustainability 2021, Vol. 13, Page 2211*, *13*(4), 2211. https://doi.org/10.3390/SU13042211

Tran, S. V. T., Nguyen, T. L., Chi, H. L., Lee, D., & Park, C. (2022). Generative planning for construction safety surveillance camera installation in 4D BIM environment. *Automation in Construction*, *134*, 104103. https://doi.org/10.1016/J.AUTCON.2021.104103

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, *134*, 104059. https://doi.org/10.1016/J.AUTCON.2021.104059

Zhong, B., He, W., Huang, Z., Love, P. E. D., Tang, J., & Luo, H. (2020). A building regulation question answering system: A deep learning methodology. *Advanced Engineering Informatics*, *46*, 101195. https://doi.org/10.1016/J.AEI.2020.101195