ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.1284
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Improving Explainability of Generative Pre-trained Transformer Model for Classification of Construction Accident Types: Validation of Saliency Visualization

Byunghee YOO[1]*, Yuncheul WOO[2], Jinwoo KIM[3], Moonseo PARK[4], Changbum Ryan AHN[5]

[1] *Department of Architecture & Architectural Engineering, Seoul National University, Republic of Korea*, E-mail address: pikaybh@snu.ac.kr
[2] *Department of Architecture & Architectural Engineering, Seoul National University, Republic of Korea*, E-mail address: whkhwy@snu.ac.kr
[3] *Department of Architectural Engineering, Gachon University, Republic of Korea*, E-mail address: jinwoo@gachon.ac.kr
[4] *Department of Architecture & Architectural Engineering, Seoul National University, Republic of Korea*, E-mail address: mspark@snu.ac.kr
[5] *Department of Architecture & Architectural Engineering, Seoul National University, Republic of Korea*, E-mail address: cbahn@snu.ac.kr

Abstract: Leveraging large language models and safety accident report data has unique potential for analyzing construction accidents, including the classification of accident types, injured parts, and work processes, using unstructured free text accident scenarios. We previously proposed a novel approach that harnesses the power of fine-tuned Generative Pre-trained Transformer to classify 6 types of construction accidents (caught-in-between, cuts, falls, struck-by, trips, and other) with an accuracy of 82.33%. Furthermore, we proposed a novel methodology, saliency visualization, to discern which words are deemed important by black box models within a sentence associated with construction accidents. It helps understand how individual words in an input sentence affect the final output and seeks to make the model's prediction accuracy more understandable and interpretable for users. This involves deliberately altering the position of words within a sentence to reveal their specific roles in shaping the overall output. However, the validation of saliency visualization results remains insufficient and needs further analysis. In this context, this study aims to qualitatively validate the effectiveness of saliency visualization methods. In the exploration of saliency visualization, the elements with the highest importance scores were qualitatively validated against the construction accident risk factors (e.g., "the 4m pipe," "ear," "to extract staircase") emerging from Construction Safety Management's Integrated Information data scenarios provided by the Ministry of Land, Infrastructure, and Transport, Republic of Korea. Additionally, construction accident precursors (e.g., "grinding," "pipe," "slippery floor") identified from existing literature, which are early indicators or warning signs of potential accidents, were compared with the words with the highest importance scores of saliency visualization. We observed that the words from the saliency visualization are included in the pre-identified accident precursors and risk factors. This study highlights how employing saliency visualization enhances the interpretability of models based on large language processing, providing valuable insights into the underlying causes driving accident predictions.

Key words: Construction safety, Large language model, Black-box model, Saliency visualization