텍스트마이닝 기반의 건설사고 원인 및 위험요인 분석

김슬기*, 차기춘**, 조민건***, 박승희****
*성균관대학교 글로벌스마트시티융합전공 석사과정
e-mail : seulgikim.tech@gmail.com
**성균관대학교 글로벌스마트시티융합전공 연구교수, 공학박사
e-mail : ckckicun@skku.edu
***성균관대학교 글로벌스마트시티융합전공 박사과정
e-mail : raonik6713@skku.edu
****교신저자, 성균관대학교 건설환경공학부 교수, 공학박사
e-mail : shparkpc@skku.edu

Text Mining Based Analysis of Construction Accident Causes and Risk Factors

Seulgi Kim*, Gichun Cha**, Mingeon Cho***, Seunghee Park****
*, **, ***Dept. of Global Smart City, Sungkyunkwan University
****School of Civil, Architectural & Environmental Engineering, Sungkyunkwan University

요 약

최근 국가 및 기업에서는 산업재해 예방을 위한 다양한 노력이 있음에도, 여전히 건설산업은 타 산업과 비교하면 사고·사망률이 높은 실정이다. 건설사고의 사고원인 및 위험성을 도출하기 위해서는 과거의 사례를 바탕으로 발생빈도 및 심각성을 파악하는 것이 중요하다. 본 연구에서는 건설공사 안전관리 종합정보망(www.csi.go.kr)의 건설사고 사례의 약 11,749건 중 토목 공종에 해당되는 2,557건의 데이터를 수집하였다. 그리고 비정형데이터로 기록된 사고원인 및 재발방지대책의 내용을 중심으로 텍스트마이닝 분석을 수행하였다. TF-IDF(Term Frequency-Improved Document Frequency) 및 LDA분석(Latent Dirichlet Allocation Analysis)을 통해 토목 공종 내의 18개의 세부 공종별 분석 결과를 도출하였으며, 이는 향후 건설사고 안전정책 및 재발방지대책을 수립할 수 있는 근거를 마련할 수 있을 것이라 기대한다.

1. 서론

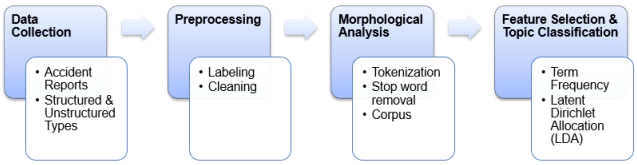
최근 국가 및 기업에서 산업재해 예방을 위한 다양한 노력이 있음에도 불구하고, 건설산업은 타 산업에 비하면 여전히 사고·사망률은 높은 실정이다. 고용노동부가 2022년 3월에 발표한 산업재해 사고 현황에 따르면, 건설업은 타 산업과 비교하면 50.4%로 사고·사망률이 넘는 것으로 나타났다 [1]. 이에 2022년 3월 국가에서는 건설사고의 저감을 위한 중대재해처벌법을 시행하였다. 사소한 재해라도 반복이 된다면 원인을 반드시 확인하여 개선할 수 있도록 하는 취지이다. 반복이 되는 원인을 찾기 위해서는 건설사고사례의 사고경위, 사고원인, 재발방지대책 등을 면밀하게 분석할 필요성이 있다. 구체적인 사고 기록 등은 정형데이터보다는 비정형데이터로 이루어져 있으며 [2], 사고 보고서를 작성하는 작성자와 텍스트의 규칙성이 상이하므로 텍스트마이닝 기법을 활용하여 건설사고 원인을 분석하고자 한다.

본 연구에서는 최근 약 3년간의 건설사고 사례를 수집하여 사고원인 및 재발방지대책의 비정형데이터를 텍스트마이닝 기법으로 분석하였다. 특히 공종별에 대한 사고빈도분석, 사고 연관성, 사고위험요소 등을 자세히 분석하고자 하였다.

2. 연구 방법 및 분석

건설현장 사고의 데이터 분석을 위한 연구방법은 데이터 수집, 데이터 전처리, 텍스트 전처리, 데이터 분석의 네 가지 단계의 순서로 진행하였으며, 그림 2에서 확인할 수 있다. 사용된 데이터는 건설안전관리통합정보망 (www.csi.go.kr) 에서 건설사고 사례의 약 11,749건 중 토목 공종에 해당되는 2,557건의 정형 및 비정형데이터이며, 웹 크롤링으로 수집하였다. 정형데이터는 날짜, 기상 조건, 사고 유형, 재정적 피해 및 사망 통계 등 정량적으로 구성되어 있는 반면, 비정형데이터는 사고경위, 사고원인, 사고 후 조치사항, 재발방지대책 등 텍스트로 구성되어 있다. 먼저, 해당 데이터셋에서 정형 및 비정형 데이터를 구별하고자 데이터를 분류하였다. 다음으로는 텍스트 전처리를 하였으며, Linux 환경에서 mecab-ko의 토

큰화(Tokenization), 불용어 처리(Stopword Removal)를 사용하였다. 마지막으로 TF-IDF(Term Frequency-Improved Document Frequency)을 통해 가장 많이 발생하는 단어를 도출하였으며, LDA분석(Latent Dirichlet Allocation Analysis)을 통해 토픽 분류 및 각 토픽에 대한 유의미한 용어들을 분석하였다.



[그림 2] 건설사고사례 텍스트마이닝 분석을 위한 연구방법

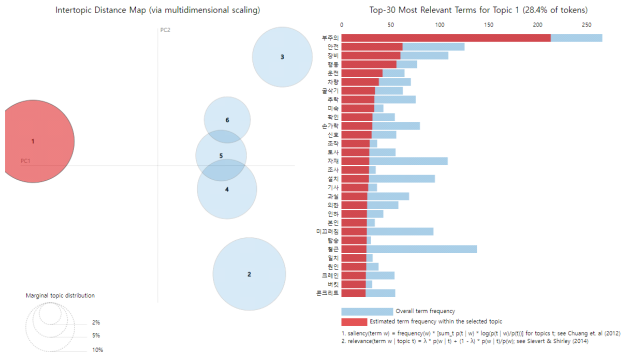
3. 분석 결과

건설 현장 사고의 구체적인 원인에 대한 Topic 분류 결과는 LDA 모델을 사용하여 도출하였다. 가장 중요한 단어는 TF-IDF를 사용하여 식별하였으며, 이는 LDA 모델의 입력으로 사용하였다. 각 주제에 대한 유의미한 용어의 워드클라우드 분석은 그림 3에 나타나 있는데, 이는 건설 현장 사고의 가장 유력한 원인을 나타낸다. 또한, 단어의 크기는 사고의 연관성을 나타낸다. 표 1은 LDA 모델의 각 주제에 대해 도출한 5개의 유의미한 단어를 보여준다. LDA 주제와 해당 단어의 시각화는 그림 4와 같다.

[그림 3] 구체적 사고원인의 워드클라우드 분석

[표 1] 구체적 사고원인의 Term-Topic 행렬 분석

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Term 1	자재	부주의	안전	철근	철근
Term 2	크레인	장비	과실	넘어짐	상부
Term 3	고정	토사	단순	자재	차량
Term 4	손가락	설치	행동	거푸집	바닥
Term 5	마흡	굴삭기	추락	안전	손가락



[그림 4] 구체적 사고원인의 LDA 토픽모델링 분석

4. 결론

본 연구에서는 건설사고사례를 바탕으로 텍스트 마이닝 기법을 적용하여 TF-IDF분석 및 LDA분석을 수행하였다. 분석 결과는 반복이 되는 위험요인을 검토할 수 있으며, 향후 안전정책 및 재발방지대책을 수립할 수 있는 근거를 마련할 수 있을 것이라 기대한다.

감사의 글

이 연구는 국토교통부/국토교통과학기술진흥원이 시행하고 한국도로공사가 총괄하는 "스마트건설기술개발 국가R&D사업(과제번호 21SMIP-A158708-02)"의 지원으로 수행되었으며, 국토교통부의 스마트시티 혁신인재육성사업으로 지원되었습니다.

참고문헌

[1] 고용노동부, "2021년 산업재해 사고·사망 현황 보도", 3월, 2022년.

[2] Cheng et al., "Text mining-based construction site accident classification using hybrid supervised machine learning", Automation in Construction, Vol. 118, pp. 103265, 10월, 2020년.