The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 5, pp.729-734, September 30, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.5.729
JCCT 2022-9-91

웹 스크래핑과 텍스트마이닝을 이용한 공공 및 민간공사의 사고유형 분석

A Study on the Analysis of Accident Types in Public and Private Construction Using Web Scraping and Text Mining

윤영근*, 오태근**

Younggeun Yoon*, Taekeun Oh**

요 약 건설업의 사고원인 파악을 위해 사고사례를 이용한 다양한 연구가 진행되고 있지만, 공공 및 민간공사의 차이점에 대한 연구는 미미한 실정이다. 본 연구에서는 발주 유형별 사고원인 분석을 위해 웹 스크래핑과 텍스트 마이닝 기술을 적용하였다. 수집된 10,000건 이상의 정형 및 비정형 데이터에 대한 통계분석과 워드클라우드 분석을 통해 공공 및 민간공사의 사고유형과 사고원인에 대한 차이가 확인되었다. 또한, 주요 사고원인들의 상관관계를 파악함으로써 향후 안전관리 대책 수립에 기여할 수 있다.

주요어 : 웹 스크래핑, 건설사고, 텍스트 마이닝, 공공 및 민간공사, 워드클라우드

Abstract Various studies using accident cases are being conducted to identify the causes of accidents in the construction industry, but studies on the differences between public and private construction are insignificant. In this study, web scraping and text mining technologies were applied to analyze the causes of accidents by order type. Through statistical analysis and word cloud analysis of more than 10,000 structured and unstructured data collected, it was confirmed that there was a difference in the types and causes of accidents in public and private construction. In addition, it can contribute to the establishment of safety management measures in the future by identifying the correlation between major accident causes.

Key words : Web Scraping, Construction Accident, Text Mining, Public and Private Construction, Word Cloud

Ⅰ. 서 론

과거부터 최근까지 건설업의 업무상사고 사망재해는 타 산업 대비 2배 이상 발생하고 있으며(그림 1), 2019년에 비해 2020년의 업무상사고 사망재해는 7% 증가한 것으로 분석된다 [1]. 건설업의 전체 사망재해 중에서 안전수칙 미준수, 근로자 부주의에 따른 떨어짐 재해의 비율이 50% 이상이며, 50억 미만의 중소규모 건설현장의 사망재해가 전체의 60% 이상을 차지하고 있다. 정부는 A형 사다리, 강관비계 등과 같은 주요 기인물에 대한 사용제한을 통해 사고를 줄이고자 노력했지만, 건설사고 발생의 다양한 원인에 대한 특성이 반영되지 않아 재해 저감 효과가 미미한 상황이다. 최근에는 중대재해처벌법이 시행되며 사업주 및 경영인의 책임과 처벌을 강화함으로써 현장 안전관리의 개선을 도모하고

*정회원, 인천대학교 안전공학과 연구원 (제1저자)
**정회원, 인천대학교 안전공학과 교수 (교신저자)
접수일: 2022년 8월 30일, 수정완료일: 2022년 9월 5일
게재확정일: 2022년 9월 9일

Received: August 30, 2022 / Revised: September 5, 2022
Accepted: September 9, 2022
**Corresponding Author: thoh@inu.ac.kr
Dept. of safety engineering, Incheon National Univ, Korea

있다. 이러한 안전관리 강화방안에도 불구하고 근로자의 낮은 인식과 다양한 원인의 융복합에 따른 재해가 지속적으로 발생하고 있으며, 이를 해결하기 위해서는 재해 발생의 원인에 기반한 근원적 예방대책과 현장 특성을 고려한 새로운 건설안전관리 방안이 요구된다.

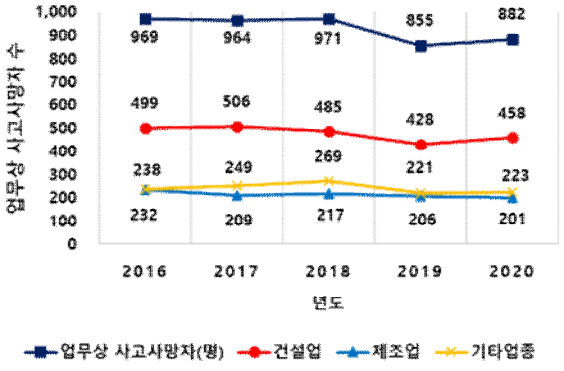

그림 1. 2016-2020 유형별 업무상 사고사망자 통계
Figure 1. Accidental death statistics by types in 2016-2020

최근에는 건설사고의 경향 파악 및 원인 분석을 위해 웹 데이터 또는 사고조사자료(근로자 정보, 현장 정보, 상태 등) 등의 범주형 데이터에 기반한 계절별 건설사고 경향 분석 및 공사금액별 사고 경향을 파악하는 연구가 진행되고 있다 [2-3]. 현재의 범주형 데이터를 이용한 분석 방법은 높은 상관관계를 가지는 변수를 이용하여 사고의 전반적인 경향을 파악할 수 있다는 장점이 있다 [3]. 하지만, 범주가 일반적인 내용으로 구분되어 있으며, 범주 내에 너무 많은 구분기준(요소)이 있어 실제 사고 발생에 대한 세부적인 원인을 파악하는 데는 한계가 있다. 따라서, 사고발생의 경위가 기록된 텍스트 데이터를 이용하여 사고원인 분석을 통해 유형별 맞춤형 안전대책을 수립할 필요가 있다. 또한, 공공 및 민간공사의 발주 및 관리 방식에 차이가 있지만, 동일한 안전대책을 수립함에 따라 유형에 맞게 적절히 안전관리가 되지 못하고 있는 실정이다. 따라서, 공공 및 민간공사의 사고유형 및 원인 분석을 통한 유형별 안전관리 대책의 수립이 필요하다.

본 연구에서는 웹 스크래핑을 이용하여 건설사고의 경위 및 조치에 대한 정형 및 비정형 텍스트 데이터를 수집하고, 수집된 데이터에 대한 텍스트 마이닝을 적용하여 공공 및 민간공사의 사고원인 분석을 통한 공사유형별 안전대책 수립에 기여하고자 한다.

II. 연구방법

본 연구에서는 건설공사 안전관리 종합정보망(CSI)에서 2019년 7월 1일부터 2022년 5월 13일까지 발생한 재해사례에 대해 웹 스크래핑 기술을 적용하여 12,122건의 데이터를 수집하였다. 수집된 데이터는 34개의 범주로 구성되며, 정형 및 비정형 데이터를 추출하여 정형 데이터에 대한 통계분석과 비정형 데이터에 대한 텍스트 마이닝을 통한 공공 및 민간공사의 유형별 재해원인에 대한 분석을 진행하였다. 구체적인 연구절차 및 방법은 그림 2와 같이 크게 4단계로 구성되며, 1) 웹 스크래핑을 통한 사고사례 데이터 수집, 2) 정형 데이터 분석, 3) 비정형 데이터 전처리, 4) 비정형 데이터의 시각화와 공공 및 민간공사의 사고유형을 비교 분석하는 절차로 진행된다. 이를 통해 공공 및 민간공사의 사고유형에 대한 경향을 파악하고 맞춤형 안전대책을 제시하고자 한다.

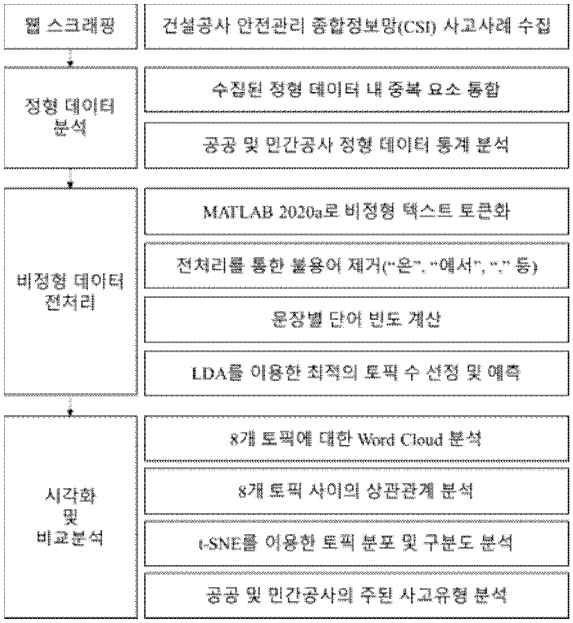

그림 2. 연구절차 및 방법
Figure 2. Research procedure and methods

III. 연구결과

1. 웹 스크래핑을 이용한 데이터 수집

건설기술진흥법 제62조 및 제67조에 따라 운영되는 건설공사 안전관리 종합정보망 (CSI)는 건설공사의 발주자를 제외한 모든 참여자가 건설사고 발생 시 국토

교통부로 신고토록하는 시스템으로 건설사고의 통계를 관리하고 정책적 개선을 위한 자료로 활용하기 위한 시스템이다. CSI의 사고사례 분류체계는 세부적으로 KOSHA GUIDE 산업재해 기록·분류에 관한 지침에 따라 구분되며, 해당 내용의 예시는 그림 3과 같다.

본 연구에서는 해당 그림 3과 같은 웹 페이지의 정보추출을 위해 셀레니움(selenium)을 활용하여 웹 스크래핑을 진행하였으며, 정형 및 비정형 데이터를 포함하는 34개의 범주를 엑셀로 추출 및 활용하였다.

그림 3. CSI 사고사례 분류 페이지
Figure 3. Accident Classification Page of CSI

2. 정형 데이터 통계분석

웹 스크래핑을 통해 수집된 정형 데이터 중 일반적인 내용인 사고명, 발생일시 등을 제외한 실제 공공 및 민간의 차이를 확인할 수 있는 사망자 수, 부상자 수, 인적사고, 물적사고, 작업 종류 및 작업자 수 범주를 활용한 통계분석을 진행하였다.

그림 4는 공공 및 민간 공사의 사망자 및 부상자 수에 대한 콤보 그래프를 보여준다.

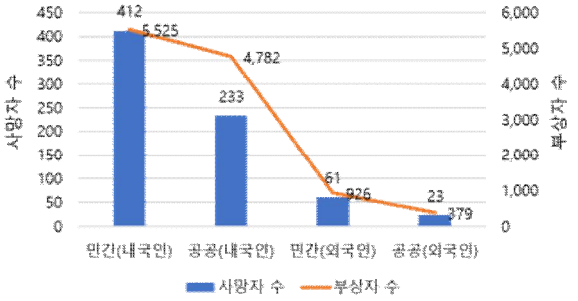

그림 4. 공공 및 민간 공사의 사망자 및 부상자 수
Figure 4. Number of fatalities and injuries in public and private

그림 4에서 2019년 7월부터 2022년 5월까지 민간공사의 내국인 사망자 수는 412명이며, 공공공사의 경우 233명의 사망자가 발생하였다. 공공 및 민간공사에 대한 전체 발주현장의 수, 근로자의 수 등과 같은 정량적 지표가 없어 세부적인 분석은 제한되지만, 부상자의 발생과 사망자 발생을 비교하였을 때 민간은 사망자 대 부상자 비율이 1:13이며, 공공은 1:21로 분석되므로 과거 재해발생이론 등에 따라 민간공사의 사망발생 비율이 높은 것으로 분석되므로 공공 및 민간공사를 구별하여 적절한 안전대책의 수립이 필요하다.

수집된 데이터 중 인적 및 물적사고에 대한 분류는 23개 이상의 요소로 구분되어 있으므로 너무 많은 요소로 인해 해석이 난해해 질 수 있다. 따라서, 동일한 재해유형(넘어짐, 떨어짐) 내의 세부 요소를 통합하여 인적, 물적사고 유형을 분석하였다. 표 1에서 공공과 민간의 인적사고 유형의 사고 발생 빈도는 넘어짐, 떨어짐, 물체에 맞음, 끼임 순으로 동일하였으며, 민간의 재해대비 공공의 재해는 약 70~80% 비율을 보였다. 물적사고의 경우 빈도가 높은 사고유형은 전도, 낙하, 충돌 붕괴 순으로 유사하였다.

표 1. 빈도가 높은 상위 4개의 인적, 물적사고 유형
Table 1. Top 4 types of human and material accidents with high frequency

인적사고	유형	넘어짐	떨어짐	물체에 맞음	끼임
	공공	1,210	1,040	833	600
	민간	1,669	1,366	1,056	702
물적사고	유형	전도	낙하	충돌	붕괴
	공공	90	64	53	21
	민간	115	117	76	48

표 2에서는 사고빈도가 높은 6개의 작업 종류가 분석되었다. 공공의 경우 설치, 해체, 이동, 운반 등의 순으로 사고빈도가 높았으며, 민간의 경우 설치, 이동, 해체, 운반 등 순으로 사고가 많이 발생하는 것을 알 수 있다. 공공과 민간의 작업별 차이점은 단순 사고발생 빈도가 민간에서 더 높다는 것 외에는 큰 차이를 식별하기 어려운 것으로 사료된다.

표 2. 사고빈도가 높은 작업 종류
Table 2. Types of work with high accident frequency

작업종류	설치	해체	이동	운반	정리	조립
공공	835	611	558	464	379	284
민간	1,224	689	746	573	422	319

그림 5는 작업자 수에 따른 사고 빈도를 보여준다. 공공은 19인 이하 공사에서 대부분의 사고가 발생하며, 작업자 수가 증가함에 따라 감소하는 경향을 보였다. 민간은 19인 이하 공사(26%), 100인~299인 공사(25%)의 사고비율이 높다. 따라서, 19인 이하 공사 및 민간의 100인~299인 공사의 안전대책을 개선할 필요가 있다.

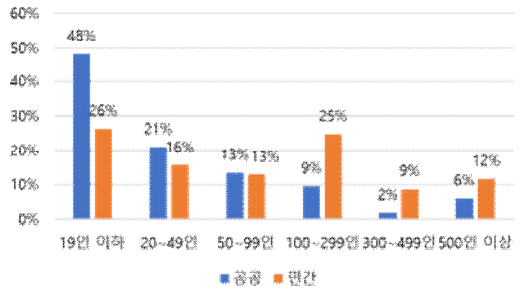

그림 5. 작업자 수에 따른 공공과 민간의 사고빈도
Figure 5. Public and private accident frequency according to the number of workers

웹 스크래핑으로 수집된 CSI의 정형 데이터를 이용하여 일반적인 통계분석을 통해 공공공사와 민간공사의 사고유형을 분석한 결과 민간의 사고빈도가 더 높고 작업자 수에 따른 차이점 외에 사고 인적·물적 사고유형 및 작업종류는 유사한 것으로 분석되었다. 하지만, 정형 데이터 분석을 통해서는 적절한 안전관리 대책을 도출하는 데는 한계가 있다. 따라서, 사고에 대한 세부적인 현장 정보, 근로자의 상태, 기인물 등을 모두 포함하고 있는 사고 경위에 대한 텍스트마이닝을 통해 공공공사와 민간공사의 사고유형을 세밀히 분석하여 안전관리 대책을 제시하는 것이 필요하다.

3. 비정형 데이터 텍스트마이닝 분석

정형 데이터 분석을 통해서는 구체적인 안전대책을 수립하기 위한 정보를 도출하기에는 제한이 있는 것으로 분석되었다. 따라서, 웹 스크래핑을 통해 수집된 비정형 데이터를 이용한 분석을 진행하였다. 그림 2의 비정형 데이터 전처리에 부문의 세부 절차에 따라 분석을 진행하였다. 비정형 데이터의 분석은 1) 문장별 토큰화, 2) 불용어 전처리, 3) 워드클라우드 시각화로 진행되었으며, 본 연구에서는 워드클라우드 시각화 전 건설사고 텍스트에 대한 최적의 토픽 수를 결정하기 위해 그림 6과 같이 문장의 혼잡도 및 시간을 계산하는 축소된 변형 베이즈, 0차 솔버(collapsed variational Bayes, zeroth order)를 사용하여 최적의 토픽 수를 8개로 결정하였다 [4-5]. 이 솔버는 실행 시간이 더 오래 걸리지만 다른 솔버 보다 정확하다는 장점이 있다.

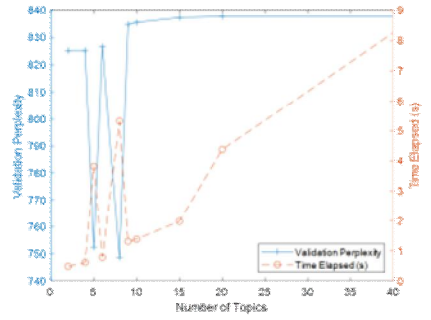

그림 6. 최적 토픽 수에 대한 혼잡도 분석
Figure 6. Perplexity analysis for optimal number of topics

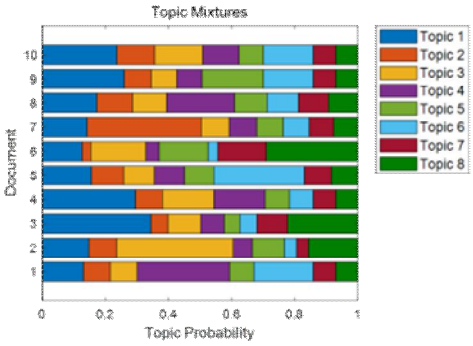

그림 7. 텍스트별 8개 토픽에 포함되는 확률 분포
Figure 7. Probability distribution of 8 topics per text

토큰화된 문장별 상위 8개 토픽의 포함 확률에 대한 예시는 그림 7과 같다. 그림 7에서 보이는 것처럼 텍스트에서 정형 데이터로 변화하였다고 하더라도 많은 정보를 포함하는 단어가 있으므로 하나의 토픽에 100%

일치하기는 어렵다. 하지만, 해당 문장의 주요 단어와 토픽 확률을 비교하여 가장 지배적인 토픽으로 문장이 할당되며, 추가로 토픽별 연관관계가 분석될 수 있다.

그림 6에서 도출된 최적의 토픽 개수 8개에 기반하여, 사고경위에 대한 워드클라우드 및 토픽별 상관관계가 그림 8 ~ 11에 제시되었다.

그림 8. 공공공사 사고에 대한 워드클라우드
Figure 8. Word Cloud for Public Construction Accidents

그림 9. 민간공사 사고에 대한 워드클라우드
Figure 9. Word Cloud for Private Construction Accidents

그림 8은 공공공사의 사고 경위에 대해 8개 토픽으로 구분된 워드클라우드 결과이다. 토픽별 분석결과, 토픽 1은 거푸집 작업(설치 또는 해체) 시 작업발판에서 미끄러져 발생하는 사고가 발생하는 것에 대한 경위를 담고 있다. 토픽 2는 자재 운반 중 부주의에 의한 넘어짐 또는 이동 중 자재에 걸려 넘어짐 사고가 발생하는 것으로 분석할 수 있다. 토픽 3은 수공수 장비 관련, 토픽 4는 철근 절단 작업 관련, 토픽 5는 중장비(굴삭기) 작업 등과 같이 주요 키워드를 통해 사고 발생에 대한 원인을 도출할 수 있다. 그림 9에서는 그림 8과 동일한 방법으로 민간공사에 대한 사고경위가 분석되었다. 가장 주된 사고원인인 토픽 1 및 3은 공공과 민간이 유사한 것으로 분석되었다. 토픽 2는 안전조치 미흡 및 부주의로 인한 사고, 토픽 4는 배관 등 자재 해체 시 손가락 끼임 등 사고로 분석될 수 있다. 공공 및 민간공사의 사고경위에 대한 워드클라우드 분석결과 발주 형태별 사고발생 원인에 차이가 있음이 확인되었다.

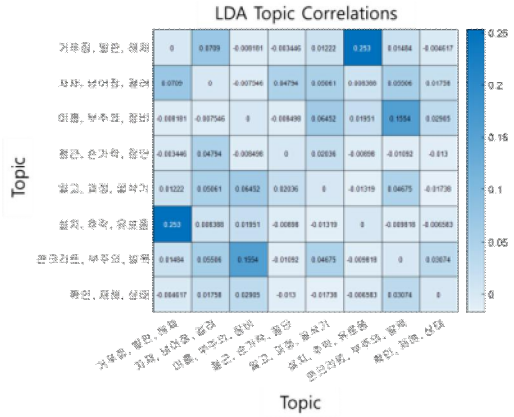

그림 10. 공공공사 토픽별 주요 키워드에 대한 상관관계
Figure 10. Correlation of keywords by public construction topic

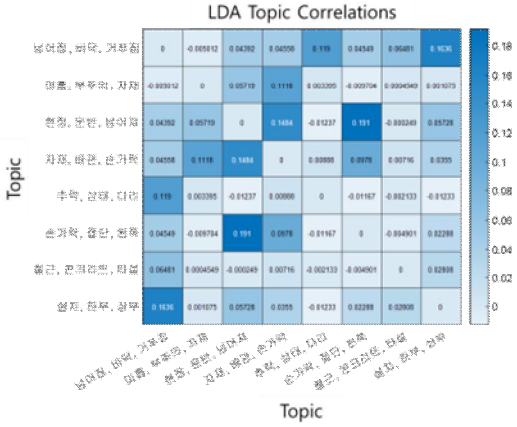

그림 11. 민간공사 토픽별 주요 키워드에 대한 상관관계
Figure 11. Correlation of keywords by private construction topic

그러나, 건설사고는 단 하나의 원인으로 발생하는 것이 아닌 여러 원인 사이의 융복합으로 인해 발생한다. 이는 그림 7에서 문장이 여러 개의 토픽에 확률적으로 분포하는 것으로 설명될 수 있다. 따라서, 그림 10 및 11에서 공공 및 민간공사의 주요 토픽별 상관관계를 분석하였다. 공공공사의 경우 토픽 1(거푸집, 발판, 해체)과 토픽 6(설치, 추락, 유로폼)의 높은 상관관계에 따라 거푸집(유로폼) 설치 및 해체 시 추락사고가 많이 발생하므로 거푸집 작업 시 안전대 착용 등의 안전관리를 강화해야 한다는 결론을 도출할 수 있다. 민간공사의 경우 토픽 3(현장, 운반, 넘어짐)과 토픽 6(손가락, 절단,

왼쪽) 및 토픽 1(넘어짐, 바닥, 거푸집)과 토픽 8(설치, 하부, 상부)의 상관관계 높게 분석되었다. 민간의 경우도 공공과 유사하게 원인을 융합한 분석이 가능하다.

그림 12에서 t-SNE를 이용한 재발방지대책 분석결과 공공 및 민간 모두 안전교육, 현장관리, 안전점검, 관리감독 및 확인 강화로 유사하였다. 유사한 사고에 동일한 재발방지대책을 적용하기 때문에 구분이 대체로 잘되며, 향후 딥러닝을 통한 사고예측모델 개발에 활용이 가능할 것으로 예상된다.

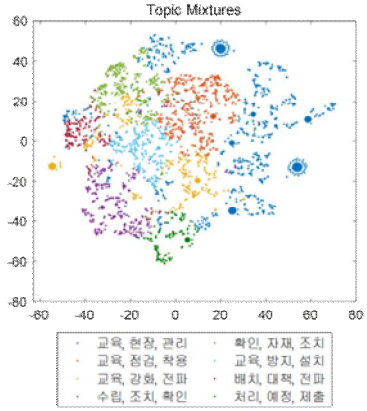

그림 12. 재발방지대책에 대한 t-SNE 결과
Figure 12. Results of t-SNE on relapse prevention measures

Ⅳ. 결 론

현재의 건설사고 분석은 고용노동부의 자료에 기반하여 건설공사 전체에 대한 통계분석을 통해 공사금액 또는 계절별 요인 등 특정 요건 및 상황에 대한 안전관리 대책을 제시하는 데 그치고 있다. 본 연구에서는 건설공사 안전관리 종합정보망에 등록된 사고사례에 대한 웹 스크래핑 기술을 적용하여 10,000건 이상의 빅데이터를 확보하고, 수집된 비정형 데이터에 대한 텍스트 마이닝을 통해 공공 및 민간공사의 사고유형을 분석하였으며, 결과를 요약하면 다음과 같다.

● 웹 상에는 건설관련 사고사례가 무수히 존재하나 활용도가 낮은 실정이다. 본 연구의 웹 스크래핑 기술과 같은 최신 기술의 적용을 통해 빅데이터를 수집하여 건설사고 분석의 수준을 향상시킬 필요가 있다.

● 비정형 데이터의 전처리를 위해 주관적, 비형식적으로 작성된 텍스트에 대한 불용어 처리 사전을 설정하였으며, 이를 통해 주요 키워드를 도출하여 공공 및 민간공사의 사고 경향을 파악할 수 있었다.

● 웹 스크래핑을 통해 수집된 정형 데이터와 텍스트 마이닝으로 도출된 정형 데이터는 향후 딥러닝 등을 이용한 사고예측모델 개발에 활용될 수 있다.

● 사고발생은 하나의 원인이 아닌 여러 원인의 융복합으로 발생하므로 도출된 주요 토픽 사이의 상관관계 분석 결과에 따라 향후 2개 이상의 원인에 대한 공통 안전대책의 수립으로 안전성 향상에 기여할 수 있다.

본 연구를 통해 비정형 데이터에 대한 텍스트 마이닝을 통해 공공 및 민간공사의 사고에 대한 다양한 분석이 가능하며, 사고유형과 사고원인에 차이가 있음을 확인할 수 있었다. 하지만, 비정형 데이터에 대한 분석 결과 현장에서 작성된 텍스트는 작성 형식이 없고 주관적으로 작성되어 일관성이 결여됨을 확인하였다. 발생하는 모든 건설사고는 건설공사 안전관리 종합정보망에 신고하도록 되어 있으므로 현재 설정된 정형 데이터에 대한 기준 외에도 비정형 텍스트 입력 시 최소한의 입력 양식을 제공함으로써 향후 텍스트 마이닝을 통한 사고분석의 일관성 및 신뢰성을 향상시킬 필요가 있다.

References

[1] Ministry of Employment and Labor, "'2016-2020 Industrial Accident and Death Statistics", 2021.
[2] K. C. Park, H. K. Kim, "Analysis of Seasonal Importance of Construction Hazards Using Text Mining", KOREAN SOCIETY OF CIVIL ENGINEERS, Vol. 41, No. 3, pp. 305-316, 2021.
[3] Y. G. Yoon, J. Y. Lee, T. K. Oh, "Text mining-based Data Preprocessing and Accident Type Analysis for Construction Accident Analysis", Journal of the Korean Society of Safety, Vol. 37, No. 2, pp. 18-27, 2022.
[4] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101, no. suppl 1 (2004): 5228-5235.
[5] Asuncion, Arthur, Max Welling, Padhraic Smyth, and Yee Whye Teh. "On smoothing and inference for topic models." In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 27-34. AUAI Press, 2009.

※ 이 논문은 2021년 및 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2021R1A6A3A01086763, No. 2022R1I1A1A01061658).