https://doi.org/10.21289/KSIC.2024.27.3.581

텍스트 마이닝을 활용한 건설안전사고 빅데이터 분석

Big Data Analytics of Construction Safety Incidents Using Text Mining

서정욱[1], 송지훈[2*]

Jeong Uk Seo[1], Chie Hoon Song[2*]

〈Abstract〉

This study aims to extract key topics through text mining of incident records (incident history, post-incident measures, preventive measures) from construction safety accident case data available on the public data portal. It also seeks to provide fundamental insights contributing to the establishment of manuals for disaster prevention by identifying correlations between these topics. After pre-processing the input data, we used the LDA-based topic modeling technique to derive the main topics. Consequently, we obtained five topics related to incident history, and four topics each related to post-incident measures and preventive measures. Although no dominant patterns emerged from the topic pattern analysis, the study holds significance as it provides quantitative information on the follow-up actions related to the incident history, thereby suggesting practical implications for the establishment of a preventive decision-making system through the linkage between accident history and subsequent measures for reccurrence prevention.

Keywords : Text Mining, Big Data, Topic Modeling, Pattern Mining, Construction Safety

1 제1저자, 석사과정, 경상국립대학교 대학원 기술경영학과
E-mail: jwseo@gnu.ac.kr
2* 교신저자, 조교수, 경상국립대학교 대학원 기술경영학과
E-mail: chsong01@gnu.ac.kr

1 First author, Graduate Student (Master's. program), Gyeongsang National University, Department of Management of Technology
2* Corresponding author, Assistant Professor, Gyeongsang National University, Department of Management of Technology

1. 서 론

산업안전은 산업재해를 예방하고 근로자의 안전과 건강을 보호하며 “인간 존중”의 이념을 실현하기 위한 기본 요소 중 하나이다. 특히, 건설 현장과 같은 산업현장에서 산업재해로 인해 발생할 수 있는 재해의 위험을 최소화하고 안전성을 확보하도록 하는 데 그 목적을 둔다[1]. 대한민국은 급속한 산업화 과정을 거치며 산업적 활용도가 높은 기술 발전과 경제성장을 이루었으나[2], 그에 반해 작업 환경과 근로자의 근로조건은 정부의 지속적인 개혁 노력에도 크게 개선되지 않았으며 산업재해는 끊임없이 발생하고 있다. 고용노동부 「2022년 산업재해 현황분석」 보고서에 의하면 대한민국의 ‘22년 산업재해 사고사망자의 수는 전년 대비 46명 증가한 874명, 사고사망만인율은 0.43으로 이는 경제협력개발기구(OECD) 회원국 38개국 중 34위에 해당하는 수치이다[3]. 2022년 중대재해처벌법이 시행되면서, 기업은 경영책임자를 중심으로 안전보건관리체계를 구축하고 이행해야 한다[4]. 나아가 정부는 선진국의 성공사례 벤치마킹과 현장 중심의 중대재해 감축 정책 효과성 제고를 위해 「자기규율 예방체계」로의 전환을 중점 과제로 삼아 산업안전 선진국으로 도약하기 위한 「중대재해 감축 로드맵」을 발표하였다. 해당 로드맵의 중점 과제는 사업주와 근로자의 자율안전점검 및 안전의식 전환 개선을 통해 사고사망만인율을 OECD 평균 수준까지 감축하는 것이다[5]. 이는 기업이 스스로 사업장 내 유해 및 위험 요인을 파악하여 현장 안전과 보건 관리 의무를 강화해야 함을 의미한다. 정부는 산업재해 중 많은 사고사망자가 발생하는 건설산업 분야에서의 산업재해예방을 위한 다양한 지원책을 펼쳐왔다. 「건설현장 안전관리체계 개선방안」과 「건축물 안전종합대책」 등과 같은 안전관리제도 관련법의 제·개정을 통해 발주청과 인허가기관으로 현장점검 기능을 확대하였고, 설계 안전성 검토 의무화와 사업장 위험성 평가를 통해 사업주가 현장 중심의 유해 위험 요소를 파악하고 평가해 관리할 수 있는 규정을 마련하는 등과 같은 개선 노력을 펼치고 있다[6]. 이와 같은 산업재해 감소 노력에도 불구하고 산업현장에서의 산업안전감독은 근본적인 사고원인의 개선보다 처벌을 위한 위반사항 적발 위주의 산업안전보건감독이 실시되고 있다[7].

건설사고 저감 정책의 하나로 건설안전사고 발생 시 건설공사 참여자는 발주청과 인·허가기관의 장에게 통보하도록 하고 있다. 통보된 사고내역은 국토교통부와 국토안전관리원에서 구축한 건설공사 안전관리 종합정보망을 통해 축적되고 있으며, 건설안전사고 사례 데이터는 공공데이터 포털을 통해 주기적으로 공개 및 업데이트되고 있다. 국토안전관리원은 건설 현장 사고를 예방하기 위한 건설공사 안전관리 종합정보망 데이터를 근거로 건설안전사고 통계를 주기적으로 분석해 건설안전 현황을 공공 및 민간에게 제공하고 있다. 무엇보다 데이터 기반 행정 활성화 정책에 따른 신뢰성 있는 데이터와 통계자료 제공 및 건설공사 현장 안전 분야의 디지털 전환을 위해 기존의 통계적 접근을 통한 정보제공 외 실효성 있는 재발방지 대책 마련을 위한 효과적인 연구 분석이 요구되고 있다. 공공데이터 포털을 통해 다양한 건설안전사고에 관한 데이터가 제공되고 있지만, 데이터 활용보다는 수집과 민간 개방에 중점을 두고 있어 데이터 품질과 일관성에 한계가 있다. 특히, 텍스트 형태인 비정형 데이터의 효과적인 활용을 위해서는 자연어처리 기술과 건설안전 분야 전문가 지원이 요구되기에 관련 연구가 심도 있게 다루어지지 않고 있다. 관련 선행연구를 살펴보면 대체로 키워드 분석과 토픽모델링 기법을 복합적으로 활용해 건설 현장에서의 재해 사고와 안전 문제를 분석하고, 이를

통해 발생하는 재해의 원인과 패턴에 대한 이해를 주된 목적으로 한다[8-12]. 이처럼 텍스트 마이닝 기법을 적용한 선행연구들은 주로 주요 사고원인 규명에 초점이 맞추어져 있지만, 사고와 이후 후속 조치 간 연계성을 파악한 연구는 다소 미비하다고 볼 수 있다. 이러한 맥락에서 공공데이터 포털을 통해 개방된 건설안전사고 사례 데이터는 사고의 경위, 사고 발생 후 조치사항 및 재발방지 대책을 포함하는 포괄적인 자료로 체계적인 분석이 이루어질 시 「자기규율 예방체계」 의사결정 지원하기 위한 실증 자료로 활용될 수 있다.

본 연구에서는 데이터 마이닝과 텍스트 마이닝을 활용해 전반적인 건설안전사고 사례에 대한 기초통계를 제시하고 텍스트 마이닝 기법 중 LDA 기반 토픽모델링을 활용해 사고의 경위, 사고 발생 후 조치사항 및 재발방지 대책에 대한 주요 토픽 분석을 통해 사고 내역에 대한 체계화 및 사고 예방을 위한 재발방지 대책 매뉴얼 수립에 기여하고자 한다. 이는 향후 건설재해 사망자 수 감소를 위한 효과적이고 현실적인 안전대책, 재발방지, 현장 안전교육 가이드 마련 등을 위한 기초 정보를 제공하는 데 기여할 수 있을 것으로 본다.

본 논문의 구성은 다음과 같다. 2장에서는 연구에 사용되는 데이터 및 전반적인 분석 과정에 관해 기술하고, 3장에서는 건설안전사고 사례 데이터의 텍스트 마이닝 분석 결과와 이를 기반으로 한 패턴 분석을 통해 체계화된 건설사고 관리와 대응을 위한 전략적 시사점을 제공한다. 마지막 4장에서는 연구의 결론과 한계점 및 향후 연구 방향에 관해 기술한다.

2. 데이터 및 연구 방법론

본 연구는 아래 Fig. 1에 기재된 데이터 분석 프로세스를 따라 진행되었다. 분석 프로세스는 총 4단계로 구분되며, 첫 번째로는 건설안전사고 사례 데이터를 공공데이터 포털을 통해 수집하였다. 두 번째로는 수집한 데이터의 전처리를 통해 비정형 데이터에 대한 가공을 진행하였다. 세 번째로는 가공한 텍스트 데이터에 토픽모델링을 적용해 분석 카테고리별 주제를 도출하였다. 마지막에는 도출된 카테고리 간 주제 패턴 분석 수행을 통해 건설안전사고 대응을 위한 전략적 인사이트를 도출하였다.

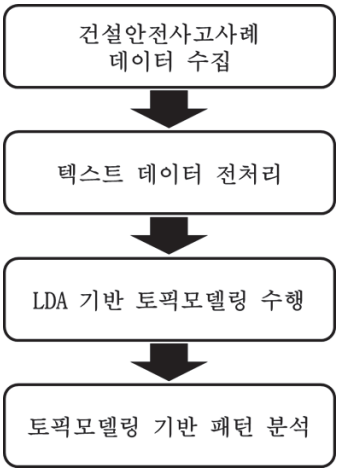

Fig. 1 Overview of the analysis framework

2.1 데이터 수집

본 연구에서는 공공데이터 포털에 공개된 국토안전관리원의 건설안전사고 사례 데이터를 수집해 분석에 활용하였다. 건설안전사고 사례 데이터는 건설기술 진흥법 제67조(건설공사 현장의 사고조사 등) 법령 시행으로, 2019년 7월 1일부터 국토교통부와 국토안전관리원이 구축해 운영 중인 건설공사 안전관리 종합정보망에 등록되는 데이터로, 공공데이터 포털에 연 1회 주기로 업데이트된다. 건설안전사고 사례 데이터 항목은 민간, 공공시설물의 분류별 인적 사고 통계(사망자 성별, 부

상자 성별, 내국인 사망자 및 부상자, 외국인 사망자 및 부상자, 사망자 연령, 부상자 연령), 사고객체, 피해현황(금액, 내용), 재발 방지대책, 사고원인, 공사예산 등으로 구성되어 있다. 데이터 중 개인정보, 민간정보, 미집계 데이터 등은 공란 처리되어 제공된다. 연구에서 사용된 데이터 범위는 2019년 7월 1일부터 2023년 4월 13일까지의 사고 데이터이며, 텍스트 마이닝 분석 대상인 컬럼은 “사고경위”, “사고발생 후 조치사항”, “재발방지 대책”의 3개 항목이다. 원본 데이터로는 총 17,682건의 사고기록을 획득하였다.

2.2 텍스트 데이터 전처리

본 단계에서는 분석 결과의 정확성 향상을 위한 데이터 전처리 작업을 수행하였고, 먼저 결측치가 있는 데이터 행을 제거하였다. 아울러 분석 데이터에서 텍스트의 길이가 사고기록마다 편차가 큼을 확인하였다. 예를 들어 “사고경위” 컬럼의 평균 길이는 100자였지만, “사고”, “추락”, “넘어짐”과 같은 2~3자 형태의 단답형 사고기록이 존재하였다. 반면 가장 긴 사고경위의 경우 1200자를 넘겼다. 더 명확하고 해석 가능한 주제를 도출하기 위해 단문으로 작성된 사고기록은 제외하였다. 분석 대상인 3개 컬럼에서 관련 기록이 40자 이상인 경우에만 텍스트 전처리 과정에 포함시켰고, 그 결과 총 3,276건의 사고사례 데이터가 남았다. 텍스트 전처리 과정에서는 먼저 정규화 작업을 통해 불필요한 문장부호 및 기타 특수문자를 제거하였다. 그리고 유사한 의미를 지니는 단어의 통일화를 위한 수정 작업을 진행하였다(예: 함마드릴→해머드릴). 이후 명사만을 추출하기 위해 파이썬 기반 KoNLpy 라이브러리의 OKT(Open Korean Text) 형태소 분석기를 활용하였다. 일반적으로 명사나 명사로 묶인 단어들은 문서의 내용을 가장 잘 대표하는 단어들로 형용사나 동사보다 더 구체적인 함축성을 지니는 경우가 많기에 문서를 구조화하는 토픽모델링과 같은 분석 수행에 적합하다[13]. 그 후 불용어 지정을 통해 문맥적으로 중요하지 않은 단어들을 제거한 후, bigram 분석을 통해 단일 용어뿐만 아니라 구문형 키워드를 추출하였다. 마지막으로 단어사전 구축에 있어 10번 미만으로 등장하거나 전체 문서에 50% 이상 등장하는 단어는 불포함시켰다.

2.3 LDA 기반 토픽모델링

토픽모델링은 일종의 비지도 학습의 한 방법으로, 문서 내에 잠재된 의미 구조(semantic structure)를 파악하는 데 활용된다. 특히, 대량의 문서를 수동으로 확인하기 어려운 환경에서 자주 등장하는 주제에 대한 개요를 파악하는 데 유용하게 사용될 수 있다. 토픽모델링의 주된 목표는 각 토픽을(주제를) 상위 t개의 관련 용어로 구성된 순위 목록으로 표현하여 k개의 토픽을 정확하게 식별하는 것이다[14]. 본 연구에서는 토픽모델링을 위한 여러 알고리즘 중 널리 사용되고 있는 LDA (Latent Dirichlet Allocation) 기반 알고리즘을 적용하며, 이는 확률 분포에 기반하고 각 문서를 다양한 토픽의 혼합 그리고 각 토픽은 단어에 대한 확률 분포로 표현됨을 가정한다[15]. 문서의 주제 분포를 이해하면 문서를 더 효과적으로 분류하고 문서 간 연관성을 밝혀내기에 더 용이하다. LDA의 단점으로는 사용자가 직접 토픽의 수를 지정해야 함에 있는데, 일반적으로 일관성 점수(coherence score) 산출을 통해 토픽 내 단어들이 얼마나 서로 의미론적으로 관련성이 높은지 확인하는 과정을 거친다. 본 연구에서는 파이썬 기반의 gensim 라이브러리(버전 4.3.2)를 적용해 분석을 수행하였다.

일관성 점수(c_v)의 산출은 총 5단계로 구분할 수 있다[16]. 각 단계에 요구되는 수식은 (1)부터

(5)에 나열하였다. 첫 번째 단계에서는 단어 쌍 집합 S를 생성한다. 특정 주제에 할당된 상위 N개의 단어 집합 $W_T=\{w_1,w_2,...,w_N\}$를 기반으로, W_T에 포함된 모든 단어로부터 가능한 단어 쌍을 생성하여 집합 S를 구성한다.

$$S = \{(w_i,w_j) \mid w_i,w_j \in W_T,\ i \neq j\} \quad (1)$$

두 번째 단계에서는 각 단어 쌍(w_i,w_j)에 대한 NPMI(Normalized Pointwise Mutual Information)을 산출한다. NPMI는 PMI를 정규화한 지표로 수식(2)로 정의된다. $P(w_i)$와 $P(w_j)$는 각 단어 w_i와 w_j의 등장 확률을, $P(w_i, w_j)$는 두 단어가 동시에 등장할 확률을 나타낸다. ϵ는 로그함수가 0으로 정의되지 않기 위해 더해주는 작은 상수값을 의미한다.

$$NPMI = \frac{\log \frac{P(w_i, w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)+\epsilon} \quad (2)$$

세 번째 단계에서는 각 단어 쌍에 대해 컨텍스트 벡터 u와 w를 생성한다. 컨텍스트 벡터는 단어들이 전체 어휘(vocabulary)와의 동시 출현 통계를 기반으로 생성되며, 수식(3)으로 정의된다.

$$u(W') = \left\{ \sum_{w_i \in W'} NPMI(w_i, w_j) \right\}_{j=1,..,|W|} \quad (3)$$

네 번째 단계에서는 컨텍스트 벡터에 대한 코사인 유사도 $\phi_{S_i}(u,w)$를 산출한다. 이는 동일한 주제 내에서 단어들이 얼마나 자주 같이 등장하며 어떤 관계를 맺는지를 정량화한 과정이다.

$$\phi_{S_i}(u,w) = \frac{\sum_{i=1}^{|V|} u_i \cdot w_i}{\| u \|_2 \cdot \| v \|_2} \quad (4)$$

마지막 단계에서는 주제에 대한 일관성 점수를 모든 단어 쌍에 대한 평균 코사인 유사도로 나타낸다. 여기서 $|S|$는 단어 쌍의 수를 나타낸다.

$$C_v = \frac{1}{|S|} \sum_{i=1}^{|S|} \phi_{S_i}(u,w) \quad (5)$$

2.4 토픽모델링 기반 패턴 분석

본 분석단계에서는 각 컬럼에서 도출된 주요 토픽(dominant topic)을 바탕으로 패턴 분석을 진행한다. 즉, "사고경위"를 대표하는 주제가 어떻게 "사고발생 후 조치사항"과 "재발방지대책"으로 연결되는지의 연관성을 파악하여 건설안전사고 예방체계의 구축을 위한 인사이트를 도출하고자 한다. Fig. 2는 이러한 패턴 분석에 대한 개념적 개요를 예시 기반으로 나타낸다. 해당 예시에서는 분석에 활용된 개별 컬럼이 하나의 대표성을 띠는 주제로 표현된다. 이후 각 주제가 동시에 출현하는 빈도수를 기준으로 빈번히 나타나는 패턴을 식별한다. 식별된 패턴에 대한 해석을 통해 건설안전사고 유

사고경위	사고발생 후 조치사항	재발방지 대책
피해자는 오전 10시경에 작업자 3명이 […]	사고발생 후 바로 병원이송 조치하고 […]	근로자 특별안전교육 실시, 작업자가 […]
타워크레인 4호기 해체 […]	지정병원에 X-Ray, CT 촬영 후 […]	타워크레인 기사와 해체팀 상호 신호체계 […]
재해자는 당일 07시04분 출근하여 […]	2023.04.03.(월) -11:00 : 사고 발생 […]	- 위험성평가 실시 및 안전사고(재발방지) 조사 […]
…	…	…

↓

사고경위 (대표 주제)	사고발생 후 조치사항 (대표 주제)	재발방지 대책 (대표 주제)
Topic 1	Topic 4	Topic 2
Topic 1	Topic 4	Topic 2
Topic 5	Topic 2	Topic 3
…	…	…

↓

패턴	패턴 빈도
Topic 1 → Topic 4 → Topic 2	350
Topic 5 → Topic 2 → Topic 3	250
…	…

Fig. 2 Conceptual overview of pattern analysis based on topics

형에 대한 보다 심도 있는 이해와 체계화된 후속 조치를 위한 전략 수립에 기여하고자 한다.

3. 분석 결과

3.1 기술통계 분석

이번 장에서는 분석에 활용한 건설안전사고 사례의 주요 기술통계를 제시한다. 원본 데이터 중 텍스트 전처리 과정을 거친 3,276건에 대한 연도별 사고사례 건수 변화의 추이, 시설물 및 사고원인에 대한 유형별 분포 현황을 시각화하였다. Fig. 3에 의하면 연도별 사고사례 건수의 변화추이는 2019년부터 2022년까지 꾸준히 증가추세를 보이다 2023년 급감한 것으로 나타나는데, 이는 2023년 발생한 사고사례에 대한 수집이 4월로 마무리

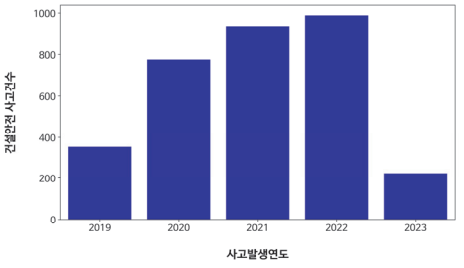

Fig. 3 Construction safety accident statistics by year

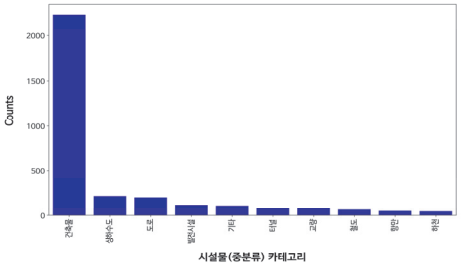

Fig. 4 Construction safety accidents statistics by facility classification

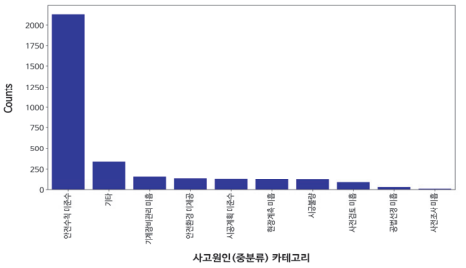

Fig. 5 Construction safety accidents statistics by accident cause classification

되었기 때문이다. 따라서 2023년 실제 사고사례는 더 많을 것으로 추정할 수 있다. 이는 전처리 전 원본 데이터의 변화 추세와 비교하더라도 비슷한 양상을 나타낸다. Fig. 4는 시설물(중분류) 별 건설사고 발생 현황을 나타내는데, 일반 건축물 유형에서 압도적으로 많은 사고사례가 발생함을 확인하였다. Fig. 5에 의하면 주요 사고원인 유형으로 “안전수칙 미준수” 비중이 가장 높았으며, 그 외 “기타”, “기계장비관리 미흡”, “안전환경 미제공” 등의 순으로 나타났다. 이와 같은 분석 결과는 건설현장에서 여러 규정 미준수 및 관리 미흡으로 인한 건설안전사고가 지속적으로 발생하고 있음을 의미한다.

3.2 토픽모델링 분석 결과

본 연구에서는 앞서 언급한 일관성 점수를 비교해 LDA 기반 토픽모델링을 위한 최적의 토픽 수를 선정하였다. 점진적으로 토픽의 수를 3에서 11까지 증가시켜가며 의미론적 일관성이 높게 나타나는 모델을 선택해 분석을 수행하였다. Table 1은 분석 컬럼별 최적의 토픽 개수와 이에 상응하는 일관성 점수를 나타낸다. “사고경위”의 경우 5개의 토픽, “사고발생 후 조치사항”과 “재발방지 대책”의 경우 각기 4개의 토픽이 도출되었다. Table 2부터 Table 4에는 토픽모델링 분석 결과를 정리하였고, 각 토픽을 대표하는 상위 키워드

15개를 나열하였다. Table 5는 개별 토픽에 대한 레이블값을 부여한 결과를 나타낸다. “사고경위”로부터는 Topic 3을 제외하고 건설현장 작업 중 발생 가능한 여러 사고의 유형이 도출되었다. “사고발생 후 조치사항” 컬럼에서는 사고 후 병원으로의 이송 및 진료, 이에 따른 행정적 절차 그리고 작업 안전관리 교육의 시행 등이 주요 주제로 파악되었다. “재발방지 대책”으로는 전반적으로 안전교육 관련 토픽이 도출되었으며, 세부적으로는 안전한 작업환경 조성을 위한 예방조치, 가설물 설치 및 위험평가 관리 등이 확인되었다.

Table 1. Determination of the optimal topic number

분석 컬럼	토픽 개수	Coherence score
사고경위	5	0.3974
사고발생 후 조치사항	4	0.4078
재발방지 대책	4	0.3730

Table 2. Topic modeling results for incident history

구분	토픽 키워드 (Top 15)
Topic 1	추락, 설치, 상부, 비계, 높이, 발판, 지하, 해체, 바닥, 하부, 구간, 고정, 골절, 발, 작업발판
Topic 2	굴삭기, 장비, 적재, 차량, 공사, 토사, 자재, 구간, 지게차, 에이치빔, 파일, 충돌, 인양, 빔, 도로
Topic 3	확인, 진료, 통증, 건설, 이송, 실시, 수술, 상황, 본인, 진행, 입원, 동료, 이상, 퇴근, 근로복지공단
Topic 4	바닥, 배관, 자재, 골절, 운반, 절단, 지상, 발, 부분, 외부, 손, 콘크리트, 세대, 사용, 갱폼
Topic 5	철근, 거푸집, 파이프, 설치, 하부, 손가락, 합판, 유로폼, 손, 자재, 고정, 절단, 망치, 좌측, 조립

Table 3. Topic modeling results for post-incident measures

구분	토픽 키워드 (Top 15)
Topic 1	병원_이송, 병원, 조사, 컴퓨터단층_촬영, 치료, 진단, 엑스레이_촬영, 수술, 소견, 건설_신고, 입원, 결과, 확인, 촬영, 도착
Topic 2	병원, 병원_이송, 치료, 이송, 조치, 즉시, 응급_조치, 수술, 병원_후송, 후송, 입원, 도착, 검사, 진료, 실시
Topic 3	보고, 병원, 확인, 병원_진료, 병원_방문, 진단, 통보, 시공사, 근로자, 치료, 퇴근, 통증, 조사, 본인, 요청
Topic 4	실시, 안전, 근로자, 설치, 조치, 교육_실시, 관리감독자, 제거, 시행, 재발_방지, 안전교육_실시, 해당, 교육, 구간, 관리

Table 4. Topic modeling results for preventive measures

구분	토픽 키워드 (Top 15)
Topic 1	안전, 금지, 사용, 장비, 철저, 해체, 신호수, 배치, 조치, 사전, 관리감독, 설치, 위험, 강화, 진행
Topic 2	안전, 시행, 재발_방지, 관리, 공사, 예정, 철저, 지시, 방지, 점검, 조치, 예방, 안전관리, 진행, 안전사고
Topic 3	설치, 안전, 자재, 구간, 조치, 고정, 통로, 운반, 철근, 바닥, 철저, 상부, 주의, 금지, 정리정돈
Topic 4	안전, 점검, 보호구_착용, 관리감독, 위험_요인, 철저, 방법, 위험_요소, 비계, 강화, 제거, 전파, 위험성_평가, 고소, 해당

Table 5. Topic labeling results

구분	Topic	주제
사고경위	1	비계설치 및 해체 중 추락사고
	2	건설기계 적재 작업 중 발생한 사고
	3	사고발생 현장 대응 및 조치
	4	바닥 및 자재 작업 중 발생한 사고
	5	철근 및 건설 자재 작업 중 발생한 사고
사고발생 후 조치 사항	1	병원 이송 및 정밀진단
	2	응급 상황 및 병원 이송 처리
	3	병원진료 및 사고 행정처리
	4	작업 안전관리 및 교육 시행
재발방지 대책	1	작업 안전 강화 및 관리 조치
	2	안전점검 및 예방조치 관리
	3	안전 가설물 설치 및 자재고정 관리
	4	작업 안전 및 위험평가 관리

3.3 패턴 분석 결과

본 단계에서는 토픽모델링의 결과값을 활용해 동시 출현하는 토픽 패턴에 대한 분석을 수행하였다. Table 6은 가장 출현 빈도가 높은 상위 10개의 토픽 패턴을 나타낸다. 토픽 패턴은 2.4장에서 언급한 바와 같이 “사고경위”→“사고발생 후 조치사항”→“재발방지 대책” 순으로 구성된다. 분석 결과에 의하면 압도적으로 높은 비율을 차지하는 지배적 패턴은 발견되지 않았으나, 전반적으로 “사고경위” 컬럼에서는 “비계설치 및 해체 중 추락사고”와 “철근 및 건설 자재 작업 중 발생한 사고”의 비중이 높게 나타났으며, “사고발생 후 조치사항”으로는 “작업 안전관리 및 교육 시행”이 주를 이루었다. 특히, 추락사고는 주로 발 골절로 이어지며 건설 자재 취급에 있어 발생한 사고에서는 손 또는 손가락 부위의 상해를 유추할 수 있었다. “재발방지 대책”에서는 “작업 안전 및 위험평가 관리”가 상위 두 패턴에서 부각되었다. 이는 사고 발생 후 부상자에 대한 병원 이송과 더불어 관리감독자에 의한 재발방지 안전교육이 동시에 시행됨을 시사한다. 이후 후속 조치로 작업자의 안전을 위협하는 위험요소 제거나 안전 가설물 설치를 통해 전반적인 관리 조치를 강화하는 형태로 사고사례에 대한 처리가 완료됨을 확인할 수 있었다. 이러한 연구 결과는 특정 사고경위에 따른 연계된 후속 조치를 정량화해 제공함으로써, 사고발생 시 요구되는 유형별 대응 매뉴얼 수립을 보조할 수 있을 것으로 전망된다. 특히, 재발 방지를 위한 관리강화에 있어 보호구 착용, 신호수 배치, 위험요소 제거, 위험성 평가와 같은 예방적 조치를 통한 리스크 관리가 체계적으로 수행되고 있음을 강조한다.

Table 6. Pattern analysis based on derived topics

No.	패턴			패턴 빈도 및 비율
1	Topic1	Topic4	Topic4	119 (3.6%)
2	Topic5	Topic4	Topic4	107 (3.3%)
3	Topic5	Topic4	Topic3	95 (2.9%)
4	Topic5	Topic4	Topic1	86 (2.6%)
5	Topic1	Topic4	Topic3	81 (2.5%)
6	Topic5	Topic4	Topic2	80 (2.4%)
7	Topic2	Topic4	Topic1	79 (2.4%)
8	Topic2	Topic2	Topic2	75 (2.3%)
9	Topic4	Topic4	Topic4	73 (2.2%)
10	Topic4	Topic4	Topic3	72 (2.2%)

4. 결 론

최근 건설공사 기술력 발전과 더불어 안전을 중점으로 한 설계와 시공이 강화되고 현장 안전시설 및 안전교육이 체계화되고 있지만, 여전히 안전을 위한 작업절차 미준수와 형식적인 안전교육 수행으로 인해 예방 가능한 안전사고가 반복해서 발생하고 있다. 정부는 산업재해를 감축하고자 다양한 정책을 펼쳐왔으며, 텍스트 마이닝을 접목시킨 선행연구에서는 주로 사고원인 규명 및 사고유형 분석에 초점이 맞추어져 있었다. 본 연구에서는 공공데이터 포털에 공개된 건설안전사고 사례 데이터를 중심으로 LDA 기반 토픽모델링을 적용해 사고경위를 세부적인 특성에 따라 구분하였고, 이어지는 후속 조치와의 연계성을 패턴화해 제시하였다. 그 결과 사고경위 관련 5개의 토픽, 사고발생 후 조치와 재발방지 대책 관련 각 4개의 토픽을 획득하였다. 토픽 간 연계성을 패턴화한 결과 사고 발생 후 작업안전에 대한 교육이 시행되었으며, 재발방지 대책으로 안전점검을 강화하고 위험성을 평가하는 과정을 찾아볼 수 있었다.

정부 정책의 기조가 처벌보다 예방을 우선시하

는 방향으로 변화하고 있는 가운데, 본 연구의 결과는 현재 건설안전사고 기록이 어떻게 정의되고 관리되는지 모니터링하는데 긍정적 기여를 할 수 있을 것으로 기대한다. 나아가 향후 사고기록 데이터 관리뿐만 아니라 예방적 의사결정 체계 구축을 위한 실무적 시사점 제시를 위한 초기 학술연구로 의의가 있을 것으로 본다. 이는 사고 발생 시 시나리오별 세부 행동 절차를 마련하는 데 기여할 수 있을 것이다.

본 연구의 한계점으로는 텍스트 전처리 과정을 거치며 원본 데이터의 약 20%만이 실질적 분석에 활용되어 더 다양한 토픽 패턴 도출이 어려웠다는 점이다. 향후 사고사례 작성 기준을 재정립해 단답형 답변을 지양하는 형태로 데이터의 수집이 이루어진다면 더 의미론적으로 풍요로운 시사점 제공이 가능할 것으로 여겨진다. 이를 위해서는 토픽모델링 기법 적용에 적합한 최소 텍스트 길이를 결정하는 실험이 요구된다. 또한, 이번 연구에서 명사만을 추출해 분석에 활용했다면, 향후 연구에서는 동사나 형용사 등 다양한 품사를 포함시켜 보다 포괄적인 분석을 수행할 필요가 있다. 마지막으로 최신 거대언어모델(LLM)을 활용해 사고사례를 임베딩하는 형태로 문맥 정보를 반영한다면 텍스트가 지니는 의미적 유사성을 더 효과적으로 반영할 수 있을 것으로 기대한다.

사 사

본 논문은 산업통상자원부의 ‘융합기술사업화 확산형 전문인력 양성사업’의 지원을 받아 수행된 논문임.

참고문헌

[1] 심규범, 건설현장의 산업안전 효과 제고 방안, 한국건설산업연구원, p.3-105, (2007).
[2] 신장철, 해방 이후의 한국경제와 초기 경제개발 5 개년계획-원조경제의 탈피와 수출드라이버 정책의 채택을 중심으로, 한일경상논집, 66, p.3-24, (2015).
[3] 고용노동부, 2022년 산업재해현황분석(산업재해보상보험법에 의한 업무상 재해를 중심으로), p.15-20, (2023).
[4] 중대재해 처벌 등에 관한 법률, 제4조 1항.
[5] 고용노동부, 「중대재해 감축 로드맵」 및 관련자료 게시 [보도자료], <https://www.moel. go.kr/policy/policydata/view.do?bbs_seq=20221201442>, viewed 29 February (2024).
[6] 사업장 위험성평가에 관한 지침 제7조.
[7] 고용노동부. 위험성평가 특화점검 등의 본격 실시를 위한 “2023년도 산업안전보건감독 종합계획” 발표 [보도자료], < https://www.moel.go.kr/news/enews/report/enewsView.do?news_seq=14585>, viewed 29 February (2024).
[8] 박기창, 김형관. 텍스트마이닝을 이용한 건설공사 위험요소의 계절별 중요도 분석. 대한토목학회논문집, 41(3), p.305-316, (2021).
[9] 김하영, 이준성, 장예은, 건설 재해사례 보고서의 텍스트 마이닝을 통한 복합사고 패턴 분석, 대한건축학회논문집, 38(4), p.237-244, (2022).
[10] 김진국, 양충헌, 박수빈, 텍스트 마이닝을 이용한 터널 교통안전 연구동향 분석, 디지털콘텐츠학회논문지, 23(10), p.2075-2083, (2022).
[11] 이상규, 비정형 텍스트 기반의 토픽 모델링을 이용한 건설 안전사고 동향 분석. 한국산학기술학회 논문지, 19(10), p.176-182, (2018).
[12] 신승현, 원정훈, LDA 토픽모델링과 네트워크 분석을 이용한 중대규모 건설현장 사고동향 분석, 빅데이터서비스학회 논문집, 1(2), p.65-78, (2023).
[13] 유제원, 송지훈, 슈퍼앱 리뷰 토픽모델링을 통한 서비스 강화 방안 연구. 한국산업융합학회 논문집, 27(2), p.343-356, (2024).

[14] Belford, M., Greene, D. Ensemble topic modeling using weighted term co-associations, Expert Systems with Applications, 161, 113709, p.1-13, (2020).
[15] Blei, D. M., Ng, A. Y., Jordan, M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(Jan), p.993-1022, (2003).
[16] Syed, S., Spruit, M. Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation, In 2017 IEEE International Conference on Data Science and Advanced Analytics, p.165–174, (2018).

(접수: 2024.05.07. 수정: 2024.06.03. 게재확정: 2024.06.07.)