

토픽 모델링을 이용한 건설현장 추락재해 분석

류한국
삼육대학교 건축학과 교수

Falling Accidents Analysis in Construction Sites by Using Topic Modeling

Hanguk Ryu
Professor, Department of Architecture, Sahn Yook University

요 약 본 연구는 기계학습 기법 중 토픽 모델링을 활용하여 건설현장에서 발생하는 추락재해에 대한 토픽을 분류하고 각 토픽에 따른 재해요인을 분석하였다. 잠재 디리클레 할당 기반의 토픽 모델링을 적용하기 위해 텍스트 데이터의 전처리를 하였고 Perplexity 점수로 평가하여 모형의 신뢰성을 높였다. 각 토픽에서 공통으로 도출된 추락재해의 대부분은 소규모 사업장에 속한 일용직 작업자들에게 발생하였다. 추락재해의 대부분의 원인은 안전장비 미착용, 현장 정리·정돈 미흡, 안전장비의 성능 및 착용 상태로 인해 제대로 작동하지 않은 것으로 판단되었다. 추락재해를 예방하고 절감하기 위해서는 소규모 사업장에 맞는 안전교육과 작업장의 정리·정돈과 개인 안전장비의 적절한 착용 상태 및 성능을 확인하는 것이 중요한 것으로 도출되었다.

주제어 : 건설안전, 추락재해, 빈도 분석, 토픽 모델링, 잠재 디리클레 할당

Abstract We classify topics on fall incidents occurring in construction sites using topic modeling among machine learning techniques and analyze the causes of the accidents according to each topic. In order to apply topic modeling based on latent dirichlet allocation, text data was preprocessed and evaluated with Perplexity score to improve the reliability of the model. The most common falling accidents happened to the daily workers belonging to small construction site. Most of the causes were not operated properly due to lack of safety equipment, inadequacy of arrangement and wearing, and low performance of safety equipment. In order to prevent and reduce the falling accidents, it is important to educate the daily workers of small construction site, arrange the workplace, and check the wearing of personal safety equipment and device.

Key Words : Construction Safety, Fall Accident, Frequent Analysis, Topic Modeling, Latent Dirichlet Allocation

1. 서론

1.1 연구의 배경 및 목적

건설구조물이 고층화, 대형화됨에 따라 공정이 매우 복잡해지고 고소작업이 많아짐에 따라 현장에서 일하던 작업자가 높은 곳에서 떨어지는 안전재해가 빈번하게 발생하고 있다[1,2]. 공사 전에 가설 구조물을 설치하거나,

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2017R1A2B4003722).

*Corresponding Author : Hanguk Ryu(ryuhanguk333@gmail.com)

Received May 13, 2019

Revised June 21, 2019

Accepted July 20, 2019

Published July 28, 2019

본 공사 완료 후 해체하는 작업에서 추락재해가 많이 발생하고 있으며 추락재해는 사망에 이르는 확률이 높다.

다양한 개인안전장비, 안전난간, 추락 방지용 안전망 등의 시설과 관련법과 규정에 따라 추락재해를 방지하기 위하여 노력하고 있으나 여전히 안전구조물 설치와 작업자 통제가 형식적으로 이행되는 경우가 많고 안전교육 미흡과 낮은 안전의식 수준으로 인해 추락 재해가 발생하고 있는 실정이다[3]. 비교적 대형 건설현장에서는 안전관리자가 업무상 재해 예방과 안전의식의 제고를 위하여 안전관리 교육을 집중적으로 하고 작업자가 안전 시설물을 체계적으로 구축한 이후에 작업을 진행하도록 하고 있다. 그러나 중·소규모 건설현장은 상대적으로 열악한 작업환경에 노출되어 있어 안전 재해의 발생 빈도가 높다. 중·소규모 건설현장은 시공능력이 부족하거나 재정이 열악한 소규모 업체가 시공하고 있는 현상이 대부분으로 안전보다는 원가와 공사 기간과 같은 결과에만 편중되어 있어 안전관리는 불필요한 낭비로 인식하는 경향이 있다[4,5].

건설현장에서 발생하는 빈도가 높은 업무상 재해 유형은 추락, 진도·진복, 충돌, 낙하물 등 있으며, 추락재해는 건설업뿐만 아니라 전체 산업 분야에서 15.92%를 차지할 정도로 진도·진복 다음으로 빈도가 높은 재해 유형 중 하나이다. 이러한 추락재해에 관해 많은 연구가 진행되어 왔으나, 추락으로 인해 발생하는 재해지수는 크게 줄어들지 않고 있다[2]. 업무상 재해 유형 중 추락은 건설현장뿐만 아니라 전체 산업에서도 중점적으로 연구되고 관리되는 재해 중 하나이다. 건설안전에서 추락재해에 관한 많은 연구가 진행되어왔으나 추락으로 인해 발생하는 재해는 크게 줄어들지 않고 있다.

따라서 빈번히 발생하는 추락재해의 패턴을 도출하고 직관적으로 이해하고 건설현장에 적용할 수 있도록 하는 것이 중요하다. 본 연구는 추락사고로 산재보상을 받은 재해자들의 정보를 바탕으로 안전사고의 주요한 주제를 도출한다. 기계학습 기법 중 잠재 디리클레 할당(latent Dirichlet allocation, LDA) 기반의 토픽 모델링(topic modeling)을 적용하여 재해위험 요소와 세 개의 토픽 시나리오를 중심으로 분석한다.

1.2 연구의 방법 및 범위

본 연구에서는 추락으로 발생한 재해자들의 정보를 효과적으로 분석하기 위해 토픽 모델링을 적용하였다. 토픽 모델링을 적용한 본 연구의 수행 절차는 Fig. 1과 같다.

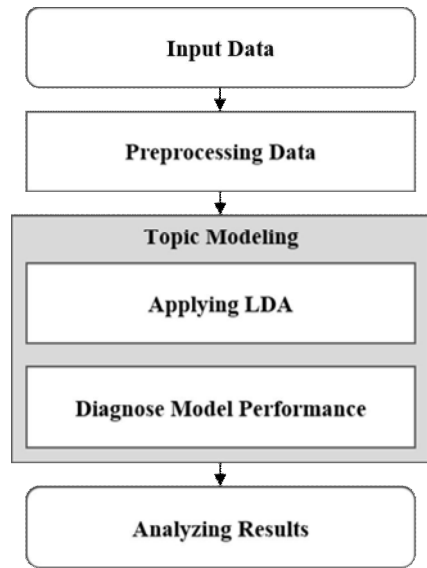


Fig. 1. Research process using topic modeling

첫째, 한국산업안전보건공단의 데이터를 바탕으로 추락재해로 인해 발생한 재해자들의 정보를 수집하였다. 둘째, 수집된 텍스트 데이터를 컴퓨터가 이해할 수 있도록 전처리 과정을 진행하였다. 셋째, LDA 기반의 토픽 모델링을 적용하고 해당 모형의 성능을 평가하였다. 마지막으로 모형에서 도출된 결과를 분석하여 향후 활용방안을 제시하였다.

2. 토픽 모델링과 잠재 디리클레 할당

2.1 토픽 모델링

토픽 모델링은 자연어 처리(natural language processing, NLP) 분야에서 문서 집합의 추상적인 토픽을 발견하기 위한 통계적 모형으로 비정형 데이터 중 보고서, 잡지, 논문, 소셜, 설문지 등과 같이 다양한 문서에서 숨겨진 의미를 발견하는 기법이다. 특정 토픽에 대한 문서에서 자주 등장하는 단어들은 유사한 의미를 지니는 데 이를 잠재적인 토픽으로 정의하고 또 다른 토픽으로 묶는 모형을 만드는 것이다. 토픽 간 비중이 어떤지 문서 내의 단어 통계를 수학적으로 분석하기 때문에 확률적 토픽 모형이라고 불리며 유의미한 구조를 발견하는 데 유용하게 사용되며 연구되고 있다[6]. Hur et al. [7]은 디지털 웹 문서가 증가함에 따라 기존의 방법론으로 정보를 탐색할 수 없는 비정형 데이터의 문서를 분류하기 위해 기계학

습 기반의 자연어 처리 모형을 적용하여 분류 모형을 개발하였다.

토픽 모델링을 통해 이슈를 분석하는 다양한 분야에서 연구가 활발하게 진행되고 있다. Parimi와 Caragea [8]는 SNS(social network service) 데이터를 활용하여 연결된 친구 링크를 예측하고 추천하는 문제를 다루었다. 온톨로지(ontology) 접근법으로 다루기엔 계산 비용이 많이 들기 때문에 LDA 기법을 적용하여 사용자 관심을 모델링하고 이를 활용하여 토픽 분포를 기반으로 링크를 예측하는 문제를 실험하였으며, 실험 결과 친구 관계를 예측하기 위한 LDA가 유용하다고 도출되었다.

DiMaggio, Nag와 Blei [9]은 토픽 모델링을 활용하여 문화에 대한 사회학적 관점 간 동질성의 활용을 미국 정부 예술기금에 관한 8,000개의 신문기사에서 정책 영역을 분석하였다. 텍스트 분석 방법을 설명하고 올바른 모형 선택과 결과해석, 토픽 모형 솔루션의 유효성 검증 방법과 통계기법을 적용하여 설명하였다. Nie, et al. [10]은 위치기반 SNS의 데이터를 활용하여 데이터가 생성된 위치를 기반으로 공간의 의미를 추출하기 위해 토픽 모델링을 활용하였다. SNS 사용자가 장소를 쉽게 이해하고 경험을 얻을 수 있게 의미론적 토픽을 통합하고 경험을 제공하는 토픽 모델링에 기초하여 제안된 프레임워크의 효과를 검증하는 시맨틱 기반의 장소 요약물 제시하였다.

Bae, Han과 Song [11]은 모바일 기기와 SNS의 결합으로 수많은 데이터를 생성하며 사회적으로도 큰 영향을 미치는 SNS 데이터를 토픽 모델링을 활용하여 트위터 사용자간에 이야기하는 주제를 찾고자 하였다. 트위터에 어떤 주제가 있었는지 추출하고 이를 웹상에 시각화하는 트위터 이슈 트래킹 시스템을 설계하였다. Na et. al. [12]은 문헌조사 및 전문가 평가와 같은 정성적인 연구방법으로는 많은 비용과 시간이 소요되고 주관적 관점이 반영되고 있는 문제점을 지적하고 토픽 모델링을 이용하여 시뮬레이션 연구 동향을 분석하였다. 그 결과 시뮬레이션을 가장 활발하게 활용되는 분야는 전기·전자 분야이며 사회과학 분야에서는 교육 및 오락 중심으로 연구 수행된다고 하였다.

Lee [13]는 540개의 건설 안전재해 관련 뉴스 데이터를 수집하여 10가지 토픽 내에서 10가지 키워드를 통해 주요 이슈를 도출하고, 월별 시계열 분석을 통해 향후 토픽에 대한 이슈를 예측하였다. Kim과 Kim [14]은 토픽 모델링을 이용한 국내 헬스케어 학술연구 동향을 분석하였다. DBpia 학술 DB에서 논문 데이터를 수집하여 수

집된 텍스트 데이터를 DMR(Dirichlet multinomial regression) 모형을 활용하여 40개의 토픽을 도출하고, 시계열 분석을 통해 헬스케어 연구 동향을 제시하였다. Jang과 Kim [15]은 한국패션디자인학회지의 논문의 주제어와 초록을 수집하고 토픽 모델링 기법을 적용하여 국내 패션디자인의 연구 동향을 파악하였다. 연구주제 및 트렌드 추이 파악으로 패션 디자인 연구 내용과 향후 방향성을 제시하였다. Yang [16]은 갑상선암 관련 연구 논문을 대상으로 LDA 기반의 토픽 모델링을 적용하여 논문에서 내재된 4개의 공통된 토픽을 도출하였다. 이처럼 토픽 모델링은 다양한 연구 분야에서 특정 토픽과 토픽에 연관된 요인들을 찾기 위해 활발히 사용되고 있다.

2.2 잠재 디리클레 할당

토픽 모델링은 수집된 문서에서 나타나는 단어를 기반으로 토픽별 단어의 분포와 문서별 토픽의 분포를 추정하는 알고리즘이다. 잠재 표현된 텍스트 데이터 안에서 의미 있는 토픽과 연관된 잠재요인을 추출해주는 확률모형으로 주로 사용되고 있는 기법은 LDA, LSA(latent semantic indexing), HDP(hierarchical dirichlet process) 등이 있으며 본 연구에서는 대표적인 기법인 LDA 기반의 토픽 모델링을 적용하였다. LDA는 수집된 데이터에서 각 데이터에 어떤 토픽들이 존재하는지에 대한 확률모형이다[17].

LDA는 특정 토픽에 특정 단어가 나타날 확률을 만들어주는데 이는 특정 토픽을 가진 문서를 만들기 위해 토픽을 선택하고 그 토픽에 맞는 글을 작성하면서 어떤 단어를 선택하고 하나의 문서를 만들어 가는 과정을 그대로 표현한 것이다. 즉, 문서가 만들어지는 과정을 확률모형으로 모형화한 것이다. LDA는 의미적으로 일관성 있는 토픽을 생성한다는 장점을 가지고 있기 때문에 토픽 모델링에서 가장 좋은 모형으로 사용되고 있다[18]. LDA의 문서를 생성하는 확률과정과 결합확률은 Fig. 2와 같다.

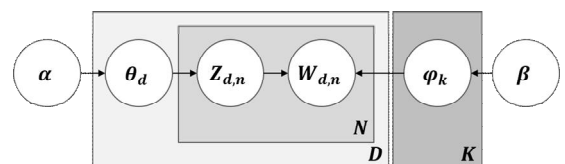


Fig. 2. LDA(latent Dirichlet allocation) Model

N	d 번째 문서의 단어 수
D	말뭉치 전체의 문서의 수
K	전체 토픽의 수
α	문서별 토픽 분포 생성을 위한 디리클레 분포 파라미터
θ_d	d 번째 문서가 가진 토픽의 비중
$Z_{d,n}$	해당 토픽을 문서 d 의 n 번째 단어에 할당
$W_{d,n}$	문서 d 의 n 번째 등장하는 단어를 할당
φ_k	k 번째 토픽에 해당하는 벡터
β	토픽별 단어 분포 생성을 위한 디리클레 분포 파라미터

각 노드는 변수를 나타내며 화살표가 시작되는 노드는 조건, 화살표가 끝나는 노드는 결과이며 네모 칸은 각 토픽, 문서, 단어의 수만큼 반복하는 것이다. α 와 β 노드를 제외한 모든 잠재 변수(latent variables)를 확률적으로 추정하는 것이다. 잠재 변수는 문서의 토픽이 무작위로 혼합되어 있고 문서의 단어가 분산되어 있는지를 결정하며 LDA의 목표는 관찰된 단어를 사용하여 숨겨진 토픽의 구조를 추론하는 것이다[19,20]. 본 연구에서는 Python 프로그래밍 언어로 LDA를 구현하였으며, 모형을 구축하기 전 평가 지표인 Perplexity로 모형의 성능을 평가하고 최적의 토픽 수를 결정하였다.

3. 추락 재해 데이터와 빈도 분석

3.1 추락재해 데이터와 토큰화

건설현장에서 발생하는 업무상 재해 유형인 추락을 분석하기 위하여 한국산업안전보건공단의 2008년부터 2014년까지 수집한 데이터를 활용하였다. 분석에 사용된 데이터는 총 4,319건으로 텍스트로 이루어진 데이터이다. 수집 데이터는 2008년도부터 산재보험에 승인된 재해자를 대상으로 전수조사로 수집된 사망 사고 데이터와 모집단을 대상으로 10% 표본추출로 수집한 부상 사고 데이터로 데이터의 신뢰성이 높다.

텍스트 기반의 비정형 데이터를 토픽 모델링에 적용하기 위해서 다음과 같은 전처리 과정을 진행하였다. 각 문서의 문장을 가장 작은 단위인 단어로 토큰화(tokenization)하여 전체적인 형태를 관찰하고 단어의 의미를 그대로 활용하기 위한 전치사, 관사, 기호는 사용하고 불필요한 용어는 제거하였다. 그리고 유의어 사전을 구축하여 의미

는 같지만 다르게 표현된 단어들을 한 단어로 변환하였다. 예를 들면, 추락, 추락사고, 떨어짐과 같은 단어를 추락으로 표준화하였다. 하나의 단어만으로 의미를 표현하지 못하는 것은 문장으로 다시 연결하였다.

이를 컴퓨터가 이해할 수 있도록 문서-단어 행렬(document-word matrix)로 표현하였다. 이는 모든 문서에서 나타나는 단어의 사전을 구축하여 각 단어를 위치와 사전의 단어를 매칭하여 벡터화 과정이다.

이렇게 구성된 단어들을 빈도를 분석한 결과는 Fig. 3과 같다. 세로축은 추락재해와 관련된 단어이며, 가로축(Ratio)은 전체 문서에서 도출되는 각 단어의 비율이다.

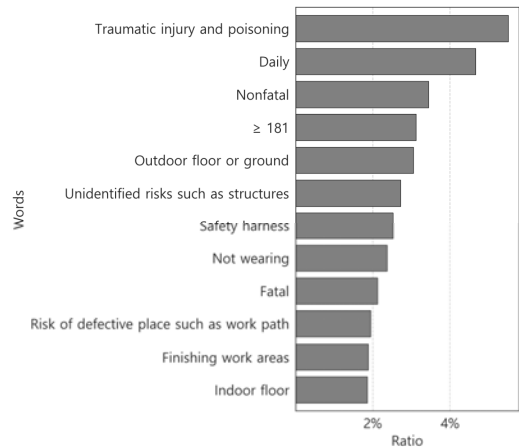


Fig. 3. Frequent of factors

3.2 추락재해 데이터의 빈도 분석

추락재해에서 나타난 단어들의 빈도를 분석하여 상위 20개의 재해요인을 요약하면 Table 1과 같다. n 은 전체 문서에서 나타난 단어의 횟수, %는 전체 문서에서 해당 단어가 나오는 비율이다.

외상성 손상 및 중독(4,283/5.5%), 일용직(3,623/4.6%), 부상(2,673/3.4%), 근무 손실일수 181일 이상(2,424/3.1%), 추락으로 인해 떨어진 장소는 옥외바닥 및 지표면(2,372/3.0%)이다. 불안정한 행동은 구조물 등 그 밖의 위험방지 및 미확인(2,115/2.7%)이다. 작업자가 재해 당시 착용하고 있던 개인안전장비는 안전대(1,957/2.5%)이다. 또한, 안전장비 미착용(1,839/2.4%), 사망(1,646/2.1%), 작업 통로 등 장소 불량 및 위험(1,509/1.9%), 공정은 마감 공사지역·공정(1,460/1.9%), 옥내바닥(1,444/1.9%) 등으로 도출되었다.

Table 1. Summary of frequent factors

Factor names	<i>n</i>	%
Traumatic injury and poisoning	4283	0.055
Daily	3623	0.047
Nonfatal	2673	0.034
≥ 181	2424	0.031
Outdoor floor or ground	2372	0.031
Unidentified risks such as structures	2115	0.027
Safety harness	1957	0.025
Not wearing	1839	0.024
Fatal	1646	0.021
Risk of defective place such as work path	1509	0.019
Finishing work areas	1460	0.019
Indoor floor	1444	0.019
Scaffolding or work step	1379	0.018
< 5	1366	0.018
Under 1 month	1347	0.017
Soma(body)	1326	0.017
Head	1291	0.017
Improper protection method	1288	0.017
Steel structure or concrete construction work areas	1096	0.014
Improper use of protective equipment	1026	0.013

4. 토픽 모델링 결과

4.1 최적의 토픽 수 도출

본 연구는 적합한 토픽의 수에 따른 모형의 성능을 평가하고 이를 토대로 모형을 구축하였다. 모형 성능 평가에 사용되는 평가 지표로는 Perplexity로 불순도 정도를 나타내며 구축한 모형이 얼마나 정확하게 예측하는지 평가하여 가능한 경우의 모형을 구축하고 전체 모형 중 Perplexity 점수가 가장 낮은 토픽의 수를 사용하여 모형을 구축한다.

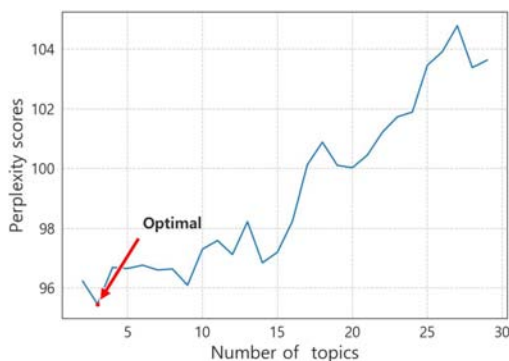


Fig. 4. Choosing optimal number of topics

토픽의 수에 따라 모형의 성능 오차가 발생하기 때문에 토픽의 개수를 2에서 100까지 10씩 증가시켜 모형을 구축하여 전반적인 점수를 평가하고 결과를 다시 반영하여 다시 구간을 2에서 30구간으로 1씩 증가하여 모형을 구축한 다음 세부적으로 평가한 결과는 Fig. 4와 같다.

토픽의 수에 따라 모형의 성능이 상이하며 가장 우수한 평가 결과를 나타내는 토픽 수는 세 개로 Perplexity 점수는 95.46로 가장 낮다. 기계학습에서 정답이 없는 비지도 학습(unsupervised learning) 기법은 현재 다양한 분야에서 훌륭한 성능을 나타내고 있지만 이를 수치상으로 평가하기 위해서는 아직 많은 한계가 존재한다. 이를 수치로 평가하기 위해 다양한 평가 지표가 개발되고 있으며 본 연구에서는 의미론적으로 일관성이 있는지 그리고 신뢰성이 있는지 확인하기 위해 Newman [19,20]의 연구에 기초하여 일관성(topic coherence)을 평가하였다.

일관성 평가는 인간의 해석 능력을 기반으로 토픽 모델링과 비교하는 방법으로 사람이 직접 평가한 토픽 적합성에 따라 뉴스는 약 80%, 책은 약 51% 정도의 차이로 일치한다. 선행연구를 통해 그 외 데이터에서 도출되는 일관성 점수는 약 30%로 기준으로 하여 평가하였으며 본 모형의 일관성 점수는 평가 기준에 근사한 29.72%이다. 토픽 모델링 결과를 바탕으로 문서를 분류한 결과는 Table 2와 같이 토픽 1에서 1,595건, 토픽 2는 1,569건, 토픽3은 1,159건으로 분류되었다.

Table 2. Topic and documents number by topic modeling

Topic number	Documents number
1	1,595
2	1,569
3	1,155

4.2 세 개의 토픽 시나리오 분석

구축한 모형에서 도출된 각 토픽에 따른 재해요인의 상위 20개를 정렬하여 요약한 결과는 Table 3과 같다.

건설현장에서 발생한 업무상 재해 유형 중 추락에 대한 토픽은 세 가지 시나리오로 나타났다. 토픽 1은 추락 재해로 인해 가벼운 부상, 토픽 2는 재해로 인해 심각한 부상 그리고 토픽 3은 사망 사고로 분석된다.

Table 3. Factors of three topics by topic modeling

Topic scenario	Factors
Topic 1	nonfatal, traumatic injury and poisoning, daily, not wearing, unidentified risks such as structures, risk of defective place such as work path, architectural wood worker, stairway or ladder, outdoor floor or ground, ladder, steel structure or concrete construction work areas, scaffolding or work step, soma(body), 29-90, 2m-3m, <5, lower limb, form assembly and dis-assembly work, indoor floor, under 1 month...
Topic 2	finishing work areas, traumatic injury and poisoning, daily, nonfatal, not wearing, scaffolding or work step, unidentified risks such as structures, outdoor floor or ground, under 2m, under 1 month, indoor floor, so-ma(body), <5, risk of defective place such as work path, 29-90, ≥181, negligence in the course of work, improper protection method, lower limb, 91-180...
Topic 3	≥181, fatal, traumatic injury and poisoning, safety harness, daily, head, outdoor floor or ground, improper use of clothing and protective equipment, improper protection method, other structures, 5m-10m, unidentified risks such as structures, defects of wearing and function with protective gear, safety helmet, indoor floor, over 20 years, 10m-20m, steel structure or concrete construction work areas, <5, under 1 month...

4.2.1 토픽 1의 시나리오(추락으로 인한 부상)

토픽 1에서 도출된 재해요인들을 중심으로 시나리오를 분석하면, 추락으로 발생한 질병은 외상성 손상 및 중독이다. 안전장비 미착용 및 불안정한 행동은 구조물 등 그 밖의 위험방지 및 미확인이다. 불안정한 상태는 작업통로 등 장소 불량 및 위험이다. 직업은 건축 목공, 기인물은 계단 및 사다리, 5인 미만 사업장, 부상 신체 부위는 하지이며, 1개월 미만의 경력을 가진 작업자이다.

토픽 1의 시나리오와 빈도 분석을 종합적으로 살펴보면, 심각하지 않은 부상은 대부분 개인안전장비 미착용, 작업장 내 위험물 방치 그리고 작업장 정리정돈 미흡 등으로 인하여 하지를 다쳐 외상성 손상 및 중독으로 도출되었다. 이런 작업환경에서 다치는 대부분 작업자는 1개월 미만의 경력을 가진 작업자이며 이러한 유형의 추락 재해를 예방은 개인안전장비 착용과 작업현장의 정리정돈 및 위험물 관리를 교육하고 관리한다면 절감될 것으로 나타났다.

4.2.2 토픽 2의 시나리오(추락으로 인한 중상)

토픽 2에서 도출된 재해요인들을 중심으로 시나리오를 분석하면, 작업공정은 마감공사, 추락으로 인한 질병은 외상성 손상 및 중독, 근무형태는 일용직, 안전장비는 미착용, 기인물은 비계 및 작업 발판이다. 불안정한 행동은 구조물 등 그 밖의 위험방지 및 미확인과 작업수행 중

과실, 가해물은 옥외바닥 및 지표면, 추락 높이는 2미터 미만, 1개월 미만의 경력, 다친 부위는 체간이다. 불안정한 상태는 작업통로 등 장소 불량 및 위험과 방호조치의 부적절이다. 추락재해로 인해 발생한 근무 손실일수는 29일에서 90일 또는 181일 이상이며 심각한 부상으로 연결된다.

토픽 2의 시나리오와 빈도 분석을 종합적으로 살펴보면, 토픽 1과 같은 안전장비 미착용과 작업현장 내 위험물 방치 및 미관리로 나타났다. 또한, 부적절한 작업 절차나 작업자의 작업수행 중 과실로 인해 발생하는 것으로 나타났다. 추락을 예방하려는 방호조치는 미흡하거나 불량인 사례로 추락재해가 발생하며 181일 이상의 근무 손실일수가 도출되었다.

토픽 1과 토픽 2의 시나리오는 대부분 소규모 작업장에서 발생하며 안전교육 및 관리·감독의 부실로 인해 재해가 발생하며 이를 지속해서 권고하거나 교육하지 못하면 소규모 작업장에서 발생하는 안전재해를 예방하거나 절감하지 못하는 것으로 판단된다.

4.2.3 토픽 3의 시나리오(추락으로 인한 사망)

토픽 3의 경우, 181일 이상, 사망, 외상성 손상 및 중독, 안전장비는 안전대와 안전모, 일용직이고 다친 부위는 머리이다. 불안정한 행동은 복장, 보호장비 부적절한 사용, 구조물 등 그 밖의 위험방지 및 미확인과 보호구 착용 상태 및 성능 불량, 근무 경력은 20년 이상, 추락 높이는 5미터에서 이상 20미터 이하로 도출되었다. 불안정한 상태는 방호조치가 부적절한 것으로 나타났다.

특히, 토픽 3의 경우 개인안전장비인 안전대와 안전모를 착용하고도 안전 사고시 사망으로 연결되는 것은 추락의 높이가 토픽 1과 2와 다르게 5~20미터로 높았다. 즉, 개인안전장비 자체가 불량하거나 부적절한 착용으로 인해 추락재해가 발생하는 것으로 나타났다. 5인 미만의 소규모 사업장의 고소에서 안전장비를 착용하지만 장비의 결함 여부를 확인하지 않고 사용하였으며 안전장비의 유지·관리가 매우 미흡한 것으로 판단된다.

토픽 3의 중요한 재해요인은 근무 경력이 20년 이상으로 작업자들이 고소에서 많이 작업하고, 오래된 현장 경험으로 인하여 안전불감증과 같은 개인의 심리 요인이 크게 작용하고 있는 것으로 판단된다. 따라서, 근무경력이 많더라도 안전에 있어서는 초심을 잃지 않고 경각심을 갖고 안전 장비의 유지·관리를 하여야 한다. 작업 전 안전장비 성능과 착용 상태를 확인하고 고소 작업을 수행하여야 한다. 토픽 모델링 결과에서 대부분 소규모 사

업장에 속한 일용직 작업자들이 추락재해의 대부분을 차지하였으므로 소규모 작업장에 대한 지속적인 안전 관리와 교육이 중요한 것으로 판단된다.

4.2.4 세 토픽 시나리오의 종합 분석

추락재해에서 도출된 세개의 토픽과 빈도를 종합적으로 검토하여 재해요인에 대한 의미와 발생 시나리오를 분석하면 다음과 같다.

첫째, 세 토픽의 추락재해는 공통적으로 소규모 사업장에 속한 일용직 작업자들에게 발생하였다. 각 토픽은 업무상 재해 정도의 심각성으로 토픽 1은 가벼운 부상, 토픽 2는 심각한 부상, 토픽 3은 사망으로 의미가 분석되었다.

둘째, 토픽 1과 토픽 2의 부상의 정도에서 발견되는 공통요인으로는 근무 경력이 1개월 미만인 작업자들에게 발생하며, 토픽 3의 사망의 경우 근무 경력이 많은 20년 이상 작업자들에게 발생하는 것으로 나타났다. 낮은 높이에서 작업하는 경우 안전장비 미착용으로 추락재해가 발생하며, 높은 고소작업의 경우 안전장비를 착용하지만, 안전장비의 성능 및 착용 상태로 인해 제대로 작동하지 않은 것으로 판단되었다.

셋째, 추락재해를 예방하고 절감하기 위해서는 소규모 사업장에 맞는 안전교육과 현장에서의 감독과 관리가 매우 중요한 것으로 나타났다. 특히, 1개월 미만 작업자들의 교육과 20년 이상의 작업자들의 안전불감증 해소 등의 노력이 필요하다. 건설현장에서의 관리는 작업장의 정리-정돈과 위험물 관리 그리고 개인안전장비의 적절한 착용 상태 및 성능 확인과 같은 것으로 도출되었다.

5. 결론

본 연구는 기계학습 기법 중 토픽 모델링을 활용하여 건설현장에서 발생하는 추락재해에 대한 토픽을 분류하고 각 토픽에 따른 재해요인을 분석하였다. 텍스트 데이터에 LDA 기반의 토픽 모델링 적용하기 위해 데이터의 전처리를 하였고 Perplexity 점수를 통해 모형을 평가하고 토픽의 수를 결정하여 결과에 대한 신뢰성을 높였다.

본 연구는 토픽 모델링 기법을 통한 기존 텍스트 분석에서 사용하는 워드 클라우드 시각화나 주제어 분석과는 다르게 기계학습 방법을 적용하여 토픽별 재해요인을 도출 및 분류하여 기존 안전재해 연구에서 다루었던 분석

과 차별성을 두었다.

본 연구는 LDA 기반의 토픽 모델링을 텍스트 데이터에 적용하여 데이터 내 의미가 있는 세 가지 토픽을 추출하고 각 토픽에서 나타나는 재해요인을 분석하였다. 토픽 모델링 결과에서 나타나는 추락재해 대부분은 소규모 작업장의 미흡한 현장관리와 안전교육 그리고 개인안전장비에 대한 결함과 착용 상태 미흡으로 나타났으며 이를 예방하기 위한 안전과 교육 정책에 대해서 좋은 방향을 제시할 것으로 기대한다. 또한, 본 연구의 결과로서 제시되는 추락재해에 관련된 안전재해 토픽을 바탕으로 향후 건설안전 정책 및 연구 방향에 기여할 수 있을 것으로 판단된다.

REFERENCES

- [1] J. H. Jo. (2012). A Study on the Causes Analysis and Preventive Measures by Disaster Types in Construction Fields. *Journal of the Korea safety management & science*, 14(1), 7-13.
- [2] Ministry of Employment and Labor. (2017). *Analysis of Industrial Accident Status*.
- [3] Korea Occupational Safety and Health Research Institute (KOSHA). (2014). *Safety and Health Research Trends*.
- [4] Ministry of Employment and Labor. (2016). *Korea Occupational Safety & Health Agency Evaluation Report for Prevention of Accident Prevention*.
- [5] S. W. Paik, H. J. Kim & D. H. Choi. (2012). A Study of Decreasing Critical Disastrous Accident in Small Construction Sites. *Journal of the Korean Society of Agricultural Engineers*, 54(6), 121-131. DOI: 10.5389/KSAE.2012.54.6.121
- [6] M. Steyvers & T. Griffiths. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7), 424-440.
- [7] Y. A. Hur, D. Y. Lee, K. K. Kim, W. H. Yu & H. S. Lim. (2017). A System for Automatic Classification of Traditional Culture Texts. *Journal of the Korea Convergence Society*, 8(12), 39-47. DOI: 10.15207/JKCS.2017.8.12.039
- [8] R. Parimi & D. Caragea. (2011). Predicting friendship links in social networks using a topic modeling approach. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. (pp. 75-86). Springer, Berlin, Heidelberg.
- [9] P. DiMaggio, M. Nag & D. Blei. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*,

41(6), 570-606.

- [10] W. Nie, X. Wang, Y. L. Zhao, Y. Gao, Y. Su & T. S. Chua. (2013). Venue semantics: Multimedia topic modeling of social media contents. *In Pacific-Rim Conference on Multimedia*. (pp. 574-585). Springer : Cham.
- [11] J. H. Bae, N. G. Han & M. Song. (2014). Twitter issue tracking system by topic modeling techniques. *Journal of intelligence and information systems*, 20(2), 109-122.
- [12] S. T. Na, J. H. Kim, M. H. Jung & J. E. Ahn. (2016). Trend Analysis using Topic Modeling for Simulation Studies. *Journal of the Korea Society for Simulation*, 25(3), 107-116.
DOI: 10.9709/JKSS.2016.25.3.107
- [13] S. G. Lee. (2018). A Study on the Trends of Construction Safety Accident in Unstructured Text using Topic Modeling. *Journal of the Korea Academia-Industrial Cooperation Society*, 19(10), 176-182.
- [14] Y. H. Kim & Y. S. Kim. (2019). Trend Analysis of Healthcare Research in Korea using Topic Modeling. *Journal of Wellness*, 14(1), 253-262.
DOI: 10.21097/ksw.2019.02.14.1.253
- [15] N. K. Jang & M. J. Kim. (2017). Research Trend Analysis in Fashion Design Studies in Korea using Topic Modeling. *Journal of Digital Convergence*, 15(6), 415-423.
DOI:https://doi.org/10.14400/JDC.2017.15.16.415
- [16] J. Y. Yang. (2019). Convergence Study on Research Topics for Thyroid Cancer in Korea. *Journal of the Korea Convergence Society*, 10(2), 75-81.
DOI: 10.15207/JKCS.2019.10.2.075
- [17] D. M. Blei, A. Y. Ng, & M. I. Jordan. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, 3, 993-1022.
- [18] D. Mimno & A. McCallum. (2008). Topic Models Conditioned on Arbitrary Features with Dirichlet-Multinomial Regression. *The 24th Conference on Uncertainty in Artificial Intelligence*. (pp. 411-418).
- [19] M. Hoffman, F. R. Bach & D. M. Blei (2010). Online learning for latent Dirichlet allocation. *In advances in neural information processing systems*. 856-864.
- [20] D. Newman, J. H. Lau, K. Grieser & T. Baldwin. (2010). Automatic Evaluation of Topic Coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. (pp. 100-108).

류 한 국(Hanguk Ryu)

[정회원]



- 1997년 2월 : 건국대학교 건축공학과(공학사)
- 2003년 2월 : 서울대학교 건축학과(공학석사)
- 2007년 2월 : 서울대학교 건축학과(공학박사)
- 2007년 8월 ~ 2019년 2월 : 국립창원대학교 건축학부 교수
- 2019년 3월 ~ 현재 : 삼육대학교 건축학과 부교수
- 관심분야 : 건설경영 및 시공, 건설안전, 데이터 사이언스, 사물인터넷
- E-Mail : ryuhanguk333@gmail.com