KJCEM 22. 5. 073~085
September 30, 2021

www.kicem.or.kr (pISSN 2005-6095 / eISSN 2465-9703)
http://dx.doi.org/10.6106/KJCEM.2021.22.5.073

딥러닝 기반 분류 모델의 성능 분석을 통한 건설 재해사례 텍스트 데이터의 효율적 관리방향 제안

김하영[1] · 장예은[2] · 강현빈[3] · 손정욱[4] · 이준성[5*]

[1]이화여자대학교 건축도시시스템공학과 학사과정 · [2]이화여자대학교 건축도시시스템공학과 박사과정 · [3]이화여자대학교 건축도시시스템공학과 석사 · [4]이화여자대학교 건축도시시스템공학과 부교수 · [5]이화여자대학교 건축도시시스템공학과 교수

A Suggestion of the Direction of Construction Disaster Document Management through Text Data Classification Model based on Deep Learning

Kim, Hayoung[1], Jang, YeEun[2], Kang, HyunBin[3], Son, JeongWook[4], Yi, June-Seong[5*]

[1]Undergraduate Student, Department of Architectural and Urban Systems Engineering, Ewha Womans University
[2]Ph.D. Candidate, Department of Architectural and Urban Systems Engineering, Ewha Womans University
[3]M.S, Department of Architectural and Urban Systems Engineering, Ewha Womans University
[4]Associate Professor, Department of Architectural and Urban Systems Engineering, Ewha Womans University
[5]Professor, Department of Architectural and Urban Systems Engineering, Ewha Womans University

Abstract : This study proposes an efficient management direction for Korean construction accident cases through a deep learning-based text data classification model. A deep learning model was developed, which categorizes five categories of construction accidents: fall, electric shock, flying object, collapse, and narrowness, which are representative accident types of KOSHA. After initial model tests, the classification accuracy of fall disasters was relatively high, while other types were classified as fall disasters. Through these results, it was analyzed that 1) specific accident-causing behavior, 2) similar sentence structure, and 3) complex accidents corresponding to multiple types affect the results. Two accuracy improvement experiments were then conducted: 1) reclassification, 2) elimination. As a result, the classification performance improved with 185.7% when eliminating complex accidents. Through this, the multicollinearity of complex accidents, including the contents of multiple accident types, was resolved. In conclusion, this study suggests the necessity to independently manage complex accidents while preparing a system to describe the situation of future accidents in detail.

Keywords : Construction Safety, CNN, Deep Learning, Classification Model, Disaster Data

1. 서론

1.1 연구 배경 및 목적

2019년 우리나라에서 산업재해(작업중사고)로 사망한 건설근로자는 428명으로, 이는 전체 산업의 50.1%에 해당한다(한국산업안전보건공단, 2020). 이러한 수치는 다양한 예방대책을 활용한 민·관의 지속적 노력에도 불구하고 좀처럼 개선되지 않고 있다. 반복되는 중대 재해를 막고 이를 원천적으로 제거하기 위해서는 사고의 근본적 원인을 면밀히 파악하여 선제적으로 예방하는 것이 필요하다.

국내 건설안전 관련 부처에서는 사고를 예방하기 위해 사고의 직접적 원인(e.g., 사고 당시 공종, 작업, 환경, 행동)을 중심으로 통계적 기법을 활용하여 재해 데이터 분석 결과를 제시하고 있다. 그러나 사고의 근본적 원인을 파악하기 위해서는 직접적 원인뿐만 아니라 교육적·관리적 원인 등의 간접적 원인도 함께 분석할 필요가 있다.

한편, 빅데이터 분석 기술을 건설현장의 안전관리에 접목하고자 하는 노력이 이루어지고 있다. 센서를 통해 습득되는 실시간 데이터를 사고 예방에 활용하기 위해서는 이미 발생한 사고들에 대한 보다 근본적 분석을 바탕으로 사고와 데이터 간 유의미한 결과를 도출하는 것이 선행되어야 한다. 즉, 하나의 사고 사례와 관련된 다양한 정보를 다각도로 분

* **Corresponding author:** Yi, June-Seong, Department of Architectural and Urban Systems Engineering, Ewha Womans University, 510 Asan Engineering Building, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul, Korea
E-mail: jsyi@ewha.ac.kr
Received April 9, 2021: **revised** August 18, 2021
accepted September 2, 2021

석함으로써 사고의 직접적·간접적 원인을 파악해야 한다.

사고의 간접적 원인을 규명하기 위해서는 재해 데이터를 분석하는 과정이 선행되어야 하는데, 건설안전 분야에서 관리되는 데이터는 대부분 텍스트 기반의 비정형 문서 형태이므로 이를 통해 간접적 원인을 파악하는 것은 매우 도전적인 일이다. 이는 사고의 직접적·간접적 원인을 파악하기 위해 오랫동안 축적된 방대한 데이터 속에서 유의미한 정보를 빠르게 자동적으로 도출할 수 있는 방안이 절실함을 시사한다.

최근 딥러닝을 통한 데이터 분석 기술의 발전으로 비정형 텍스트 데이터로부터 유의미한 정보 추출 및 분석 가능성이 증대되고 있다. 이에 본 연구에서는 딥러닝을 통해 재해 데이터를 사고 유형으로 분류함으로써 보다 심층적으로 사고의 직접적·간접적 원인을 파악할 수 있는 기반을 구축하고자 한다. 특히 정량적 통계분석이 아닌 텍스트 분석을 활용하여 추락·낙하·끼임과 같은 기존의 사고 유형 분류체계를 개선할 수 있는 방향성을 제시하며 궁극적으로는 건설안전 데이터의 활용성을 높이는데 기여하고자 한다.

1.2 연구 범위 및 방법

본 연구는 딥러닝 기반의 텍스트 데이터 분류 모델의 성능 고찰을 통해 한국어 건설 재해사례의 효율적 관리방향을 제안하기 위해 다음과 같은 범위 및 방법으로 연구를 수행한다.

첫째, 산업안전보건공단에서 제공하는 건설 재해사례 보고서 중, 중대 재해에 해당하는 데이터를 분석 범위로 한다. 본 연구에서 정의하는 중대 재해는 산업재해[1] 중 사망 등 재해의 정도가 심한 것으로(산업안전보건법 제2조 제2항), 1) 사망자가 1명 이상 발생한 재해, 2) 3개월 이상 요양을 요하는 부상자가 동시에 2인 이상 발생한 재해, 3) 부상자 또는 직업성 질병자가 동시에 10인 이상 발생한 재해 중 하나에 해당하는 재해를 말한다(산업안전보건법 시행규칙 제3조). 둘째, 분류의 범주는 건설업에서 가장 대표적으로 발생하는 사고 유형인 추락, 감전, 붕괴, 낙하, 협착의 5가지로 한정한다. 데이터는 웹 크롤링을 통해 수집하며 이를 CNN 알고리즘으로 학습 및 분류한다. 상기 데이터의 형태인 한국어 비정형 텍스트 분석을 위해 한국어 형태소 분석기에 건설산업과 관련된 용어를 추가하여 모델을 개발하고, 실험 결과와 분석을 바탕으로 본 연구의 결론과 한계를 제시한다.

1) 산업안전보건법 제2조 제1항에서는 산업재해를 "노무를 제공하는 사람이 업무에 관계되는 건설물·설비·원재료·가스·증기·분진 등에 의하거나 작업 또는 그 밖의 업무로 인하여 사망 또는 부상하거나 질병에 걸리는 것을 말한다."고 규정한다.

2. 예비적 고찰

2.1 건설산업 안전관리 및 데이터 축적 현황

2.1.1 건설산업 안전관리 현황

건설산업의 안전관리는 산업안전보건법, 건설기술진흥법 등의 법규를 중심으로 이해관계자 간 협력 아래 이루어지고 있다. 정부 주도적으로 이루어지고 있는 건설 재해에 관한 데이터는 고용노동부 산하의 산업안전보건공단, 국토교통부 산하의 국토안전관리원 두 기관에 의해 이루어지고 있다. 인적 피해와 관련된 내용은 산업안전보건공단이, 물적 피해와 관련된 내용은 국토안전관리원이 관리한다.

건설안전 제도, 위험요소 감지 및 관리, 안전 교육 문제 그리고 안전 모니터링 등은 모두 안전 분야에서 중요한 연구 주제이지만 이러한 연구들의 정보원이 되는 건설 재해 데이터와 통계시스템 및 사고 정보 분석 체계가 체계적으로 구축되어 있지 않아 사고 예방을 위한 재해 정보의 활용이 어렵다(Kim & Kim, 2001).

2.1.2 건설산업 비정형 텍스트 데이터 관리 현황

기업에서 생산되는 데이터의 약 80%는 비정형 데이터로 이루어져 있으며(Marzouk & Enaba, 2019), 그중 텍스트 데이터의 비율은 매우 높아 정형 데이터의 2배에 달하는 속도로 증가하고 있다(IBM, 2015). IBM 보고서(2015)에 따르면 전 세계 데이터의 80%가 텍스트와 같은 비정형 데이터로 이루어져 있다. 특히 건설산업은 타 산업과 대비되는 고유한 특징으로 인해 매우 정보 집중적인 성격을 갖는다(Cui et al., 2018; You & Wu, 2019).

건설프로젝트는 그 규모가 방대하기 때문에 실시간으로 수많은 인력/자재/비용의 교환이 발생한다(Martinez-Rojas et al., 2016). 또한, 일회에 그치는 프로젝트의 특성상 각종 데이터가 이종(heterogeneous)의 형태인 문서, 편지(이메일), 보고서, 스케치, 도면 또는 물리적 모델로 교환된다(Ismail et al., 2018). 특히 참여자 간의 교환을 목적으로 데이터가 매순간 재생산 및 재가공되기 때문에 텍스트 문서가 큰 비중을 차지한다. Aconex Group (2018)에 따르면 메가프로젝트의 경우 생애주기 동안 평균적으로 1억 3천만 건의 이메일과 5천5백만 건의 문서 교환이 발생한다. 해당 문서들은 비정형 텍스트로 이루어져 있으며, 대부분 계약, 시방, 회의, 변경, 보고 및 정보 요청 등으로 이루어진다(Caldas et al., 2002). 국내의 경우 CERIK (2014)에서 시공능력평가 30위권 이내의 건설기업들을 대상으로 데이터베이스 구축 현황을 조사한 결과, 약 70% 이상의 기업들이 입찰 및 계약, 원가 및 재무관리, 품질관리 등의 다양한 분야에서 데이터를 꾸준히 축적하고 있는 것으로 확인되었다.

한편 건설프로젝트 데이터의 주요한 형태인 비정형 텍스

트는 정형 텍스트에 비해 분석이 어려워 그 활용성에 있어 한계를 가진다(Al Qady & Kandil, 2013; Martinez-Rojas et al., 2016). 특히 안전 분야의 경우 건설현장 및 근로자의 상황을 기록한 비정형 텍스트 데이터가 대부분을 차지하는 것으로 알려졌다. 이로 인해 E&C 기업이 보유한 데이터 중 99.5% 이상이 전혀 활용되지 않고 있으며(Xpera Group, 2017) 건설프로젝트는 여전히 비효율적인 수단을 통해 관리되고 있다(Chassiakos & Sakellaropoulos, 2008). 이는 건설산업이 첨단 ICT 기술을 도입하려는 시도에도 불구하고 타 산업에 비해 뒤떨어진 디지털화를 더욱 늦추는 요인으로 작용한다(Ahuja et al., 2010; Sardroud, 2015).

2.2 선행연구 분석

2.2.1 건설산업 안전관리 관련 연구

건설현장에서 발생하는 각종 사고를 예방하기 위해 데이터를 기반으로 한 연구와 더불어 사고 비교·분석, 설문조사, 심층 인터뷰 등을 활용한 연구들이 진행되었다.

데이터를 기반으로 건설 사고에 기인하는 요인들을 분석하고 발생을 예측하는 연구들은 다음과 같다. Yang et al. (2009)은 IRR (Index of Relative Risk) 재해 평가기법을 적용하여 2005년~2007년 사이에 발생한 건설 재해의 직종별 재해 리스크 분석을 수행하였다.

Choi (2019)는 머신러닝을 활용한 다양한 데이터 마이닝 기법을 통해 근로자의 사망을 예측하는 연구를 수행하였다. 그 결과, 랜덤포레스트(Random Forest) 알고리즘의 성능이 92%로 근로자의 사망을 예측하는 성능이 가장 뛰어난 것으로 드러났다. 또한, 중대 재해 발생에 성별, 고용규모, 근속년수, 월, 요일 등이 유의한 요소로 도출되었고, 연령, 공사유형은 유의하지 않은 요소로 결정되었다. 이는 연령 항목과 관련하여 연령이 높을수록 사망률이 높아진다는 Amiri et al. (2015), Villanueva and Garcia (2011)의 연구와는 상반되는 결과이다.

이상의 선행연구를 고찰한 내용으로 볼 때, 결과가 정형화된 데이터를 활용한 연구에서 사고의 원인은 날씨, 작업 공종, 근로자 특성 등으로 나타나는 반면, 설문조사 및 사례 분석을 통한 연구에서는 보다 근본적 원인이 도출되었다.

2.2.2 건설산업 비정형 텍스트 데이터 관련 연구

문서 내 비정형 텍스트 분석은 기존 불가능했던 새로운 지식과 통찰력을 제시할 수 있다. 이러한 이유로 최근 텍스트 마이닝을 활용하여 사고 문서를 자동으로 분류하는 것에 대한 관심이 증가하고 있다(Ubeynarayana & Goh, 2017).

건설 데이터의 규모가 점차 커짐에 따라 이를 자동으로 분류하고 분석하는 기술의 필요성이 증대되었고, 2000년 이후 다양한 도구를 활용한 텍스트 데이터 분류 및 분석에 관한 연구가 다수 진행되었다. Al Qady and Kandil (2013)과 Zou et al. (2017)은 의미론적 유사성을 기반으로 한 문서 검색 및 분류 모델을 제시하였다. Tixier et al. (2016)는 건설재해 데이터로부터 의미 있는 자료를 자동으로 추출하는 방법론을 개발하였으며 단어 간 거리를 활용하여 보고서에 등장하는 부상 유형, 신체 부위와 같은 안전사고에 영향을 미치는 요소들을 자동으로 추출하였다. Chokor et al. (2016)은 미국 OSHA에서 공개한 재해보고서를 K-Means 클러스터링을 활용하여 군집 분석하였으며, 그 결과 완전한 비지도 학습을 활용해서 추락, 충돌, 감전, 붕괴 4개의 군집을 자동적으로 이루는 것을 확인하였다.

이처럼 산업 내 텍스트 문서 분석의 중요성이 커지면서 영어 텍스트 데이터를 대상으로 한 다양한 연구들이 진행되고 있지만, 국내에서 생성된 한국어 데이터를 대상으로 한 연구는 아직 부족한 실정이다(Park et al., 2018). 정부와 기업에서 텍스트 형태의 건설 사고 데이터가 축적되고 있으나 텍스트 데이터의 구조화되지 않은 특성으로 인해 비효율적이고 일관성 없는 지식 축적만이 이루어지고 있다(Moon et al., 2018). 이러한 특성과 한국어 데이터 분석의 어려움으로 국내에서는 건설 문서의 분류체계에 관한 연구(Park et al., 2017; Ok & Seo, 2013)는 다소 진행된 반면 문서의 자동 분류 및 정보 추출 연구, 다양한 알고리즘의 적용에 대한 연구는 충분하지 못해 영어 기반 데이터에 비해 상대적으로 좋은 성능을 보여주기 힘들었다(Moon et al., 2018).

2.3 데이터 분류 방법론

전통적으로는 사람이 직접 규칙을 정의하는 규칙 기반 방법론들을 사용하여 문서를 분류하였다. 그러나 데이터가 방대해짐에 따라 기계학습 방법론들이 등장하였으며, 그중에서도 텍스트 데이터를 다루는 연구에는 비선형 기반의 서포트 벡터 머신, 트리 모델, 앙상블 기법 등이 주로 사용되었다.

최근에는 대량의 데이터 수집 및 저장과 GPU 컴퓨팅 기술의 발달로 모델 복잡도를 쉽게 늘릴 수 있는 신경망(Neural Network) 모델이 각광받게 되었다. 또한, 단어를 특정 차원의 벡터로 변환시켜주는 Word2Vec (Mikolov et al., 2013)의 개발 이후 딥러닝을 자연어처리(Natural Language Processing; NLP)에 접목한 연구가 급속도로 성과를 내기 시작했다. 자연어처리란 사람이 사용하는 언어인 자연어를 분석하여 기계가 처리할 수 있도록 하는 것으로, 음성인식, 감성 분석, 텍스트 분류 작업 등과 같이 다양한 분야에서 활용되고 있다(Liddy, 2001). 자연어는 단어나 표현 등장 순서가 중요한 데이터로, 같은 구성요소라도 순서에 따라 다르게 해석될 수 있는 특징을 가지고 있다.

자연어처리에 주로 사용되는 딥러닝 알고리즘인 CNN

(Convolutional Neural Networks)과 RNN (Recurrent Neural Network)은 모두 단순 단어의 빈도와 등장 여부에 따른 특성 파악이 아닌 위치적 특성을 기억해 단어의 문맥적 의미를 보존하는 장점을 지닌다. 기존에는 텍스트 데이터 처리를 위해 순차적 특징을 학습하는 RNN을 사용해야 한다는 인식이 있었으나, 최근 CNN이 텍스트 데이터 처리에 좋은 성능을 나타내며 그 활용이 두드러지고 있는 추세이다 (Kim et al., 2014).

데이터 분석 방법은 연구의 목적과 특성에 따라 다른 방법이 사용될 수 있다. 본 연구에서 활용하는 문서는 한국어 텍스트 데이터이므로 본 연구에서는 건설재해사례 텍스트 데이터 분석을 위해 CNN을 사용한다. CNN은 인간이 사물을 인식할 때 부분적으로 특징을 인식하고 그 특징들을 조합하여 물체를 인식하는 시신경 구조를 모방한 알고리즘이다. 특히 데이터의 전체 영역에 대해 동일한 중요도로 처리하면서 특정 범위마다 특징을 추출하기 때문에 순서와 관계없이 전체 데이터에서 특징을 찾아낼 수 있다(Jung, 2018).

한국어는 단어 생략이 많고 단어 간 순서가 바뀌어도 의미가 전달되는 반면 조사 사용 및 동사의 변형에 따라 의미가 크게 달라진다는 특징이 있다. 다음은 사고 사례에서 발견되는 해당 특징에 관한 예시이다.

> *(사례1)…풀어진 로프를 (L형강에) 묶어 두고 옥상 난간에 (몸을) 걸친 채 작업 중 …*
> *(사례2)…L형강에 묶어 둔 로프가 풀어져…*

(사례1) 및 *(사례2)* 모두 *'풀어지다', '묶다', '로프'*세 단어가 등장하지만 *(사례1)*의 경우 사고 상황 자체를 서술한 반면, *(사례2)*는 세 단어가 직접적인 사고의 원인이 된다. 이렇듯 단어와 표현의 순서에 따라 문장 사이에 의미 차이가 발생한다. 동시에 단어 순서를 바꾸어 '옥상 난간에서 풀어진 로프를 묶어 두고 몸을 걸친 채 작업 중'으로 표현해도 의미 전달에 문제가 없음을 알 수 있다. 이 때문에 한국어 데이터 분석 시 순차적 특징을 학습하는 RNN 보다는 CNN이 더 좋은 성능을 나타낸다.

2.4 CNN 알고리즘 원리

여러 레이어로 구성된 CNN 기반 문서 분류 알고리즘의 작동 원리는 다음과 같다. 첫 번째 레이어가 단어를 저차원 벡터로 임베딩하면 다음 레이어에서 모델 변수로 설정한 여러 사이즈의 필터를 활용하여 특징을 추출하고, 임베딩된 단어 벡터에 대해 합성곱 변환을 수행한다. 그 후 합성곱 레이어의 결과의 자질을 풀링이라는 서브 샘플링을 통해 축소시킨다. 이후 드롭아웃 레이어를 지나며 신경망의 일부를 비활성화해 과적합을 방지하고, 최종적으로 소프트맥스 레이어 결과를 도출하여 분류를 수행한다.

3. 데이터 수집 및 전처리

3.1 데이터의 선정 및 수집

본 연구에서는 산업안전보건공단에서 제공하는 1994년부터 2018년까지의 '건설 재해사례 보고서' 2,201건을 활용한다. 해당 기관의 데이터를 사용하는 이유는 다음과 같다. 인적 사고는 공사 중지 등의 큰 손실을 유발하는 불확실한 리스크이며, 물적 사고가 인적 사고로 이어지는 복합적 사고의 경우에는 인적 피해가 발생한 사례로 분류되므로 산업안전보건공단의 데이터가 비교적 전체 사례를 포괄할 수 있기 때문이다. 더불어 국토안전관리원의 사고 조사 보고서에는 구조적, 기술적 진단의 내용이 많은 반면, 산업안전보건공단의 데이터는 근로자가 인지하지 못한 경우, 무리한 힘을 가한 경우 등 근로자의 행위와 상태를 서술한 내용이 많아 기존의 정형화된 데이터를 바탕으로 분석하는 것에 비해 텍스트 데이터 분석의 효과가 클 것이라고 판단하였다.

본 연구에서 공공적으로 공개된 국내 건설업 재해사례 데이터를 활용한다. 따라서 작성자에 따라 간단한 서술, 상세한 서술 등 형식이 바뀌며 같은 단어임에도 불구하고 여러

data_number_13

제목: 리프트로 탑승구에서 자재운반중 추락
업종: 건축업
기인물: 리프트
피해정도: 사망 1명
공정: 리프트로 탑승구에서 자재운반
재해유형: 추락
날짜: 1993년 04월

1. 연령 : 42세
2. 지역 : 안양시 평촌

3. 사고경위
사고당일 09:10경 아파트 신축현장에서 피재자(방수공)가 21층 리프트탑승구에서 합판을 발코니 턱에 걸쳐 놓던 중, 옥상층으로부터 하강중이던 리프트(3명탑승)가 삐져나온 합판을 침으로 인해 피재자가 튕기며52M 아래 지상으로 추락 사망

4. 사고원인
ㅇ근로자의 부주의
- 합판을 리프트 승강로상까지 부주의하에 돌출시킴으로 인해 사고발생
ㅇ리프트 탑승구 문짝 설치(폐쇄상태유지)

5. 대책
ㅇ리프트 탑승구 문짝 설치(폐쇄상태 유지)
ㅇ리프트 탑승구에 인접하여 자재적치 금지

[그림] 사고현장사진
1) 피재자 추락경로
2) 사고발생위치(21층 리프트 탑승구)

TBODY 13ENDS

가지 형태로 표현된다. 예를 들어, '굴삭기'는 '백호우', 'B/H', *'Back Hoe'* 등으로 같은 의미로 사용될 수 있다.

데이터는 Python과 매크로(macro)를 활용하여 pdf, 한글 파일, 텍스트 파일 등 다양한 형태로 수집되었다. 줄글 형태의 게시글로 제공되는 데이터는 웹 크롤링을 활용하였고, 파일로만 제공되는 데이터들의 경우 매크로를 활용하여 수집하였다. 그 외 데이터의 형식이 일정하지 않은 경우에는 직접 수집하였다. 이러한 절차를 통해 수집된 데이터에서 과도한 줄바꿈과 띄어쓰기를 제거한 뒤, 사고 경위부터 원인을 문장화하는 작업을 진행하여 'data_number_13'과 같은 형태의 데이터로 정제하였다.

3.2 데이터 전처리 및 가공

3.2.1 데이터 전처리

대부분의 텍스트 마이닝 연구는 데이터 분석을 진행하기 전에 데이터 전처리 과정을 거친다. 영문 데이터의 경우, 형태소 분석, 띄어쓰기 교정, 철자 확인에 더불어 텍스트 사이에 등장하는 숫자나 특수문자, 조사 등의 제거를 위한 많은 패키지가 개발되어있다. 그러나 한국어는 조사와 어미가 다르게 발달한 전형적인 교착어로, 영어와 달리 단어가 아닌 형태소(morpheme)가 자연어 분석에서 중요한 역할을 한다(Nam & Jo, 2017; Park et al., 2018). 비록 한국어 데이터의 경우 자연어처리 기술이 상대적으로 느리게 발전하고 있지만, 형태소 분석, 문장 파싱 패키지를 제공하는 KoNLPy (Park & Cho, 2014)의 개발 및 배포로 점차 가속화되고 있다.

형태소 분석기의 대부분은 국어사전, SNS, 뉴스 등의 데이터를 기반으로 만들어졌기 때문에 건설산업과 같은 전문 산업 분야에서는 좋은 성능을 나타내지 못한다는 단점이 있다. 특히 여러 단어의 결합형 용어가 많고, 사전에 등장하지 않는 단어의 비율이 높은 건설산업의 용어를 분석하기에 어려움이 있다. 또한, 단어별로 모두 분해하여 형태소로 분석하는 것이 목적이기 때문에 명사 외에도 부수적인 조사, 어미 등이 많이 생성되게 되는데, 이 정도 수준의 형태소 분석은 각 단어의 분포를 기반으로 거리를 계산하는 것에는 방해가 될 수 있다. 따라서 본 논문의 취지인 비지도학습을 통한 단어 분석 가능성을 가늠해 보는 데에는 명사 추출을 정확하게 해내는 수준의 품사 판별만이 필요하다.

〈Table 1〉은 한국어 데이터 분석에 사용되는 라이브러리들을 패키지화한 KoNLPy (Park & Cho, 2014)에 있는 4가지 형태소 분석기의 기능을 비교한 표이다. 해당 실험에는 '아파트 옥상 슬라브 단부에서 안전난간 설치작업 중 추락'이라는 문구를 사용하였다. 〈Table 1〉의 4가지 형태소 분석기는 공통적으로 텍스트를 형태소와 명사 단위로 나눌 수 있으며 형태소의 종류까지 반환할수 있다는 특징이 있다. 다

Table 1. Morphological analyzer's performance

Text	'apateu ogsang seullabeu danbueseo anjeonnangan seolchijageobjung chulag'			
Morphological Analyzer	Kkma	Hannanum	Komoran	Okt
Result of Parsing	'apateu'	'apateu'	'apateu'	'apateu'
	'ogsang'	'ogsang'	'ogsang'	'ogsang'
	'seullabeu'	'seullabeu'	'seullabeu'	'seullabeu'
	'dah'	'danbu'	'danbu'	'danbu'
	'n'	'eseo'	'eseo'	'eseo'
	'bu'	'anjeonnangan'	'anjeon'	'anjeon'
	'eseo'	'seolchijageobjung'	'nangan'	'nangan'
	'anjeon'	'chulag'	'seolchi'	'seolchi'
	'nangan'		'jageob'	'jageobjung'
	'seolchi'		'jung'	'chulag'
	'jageob'		'chulag'	
	'jung'			
	'chulag'			

만 각 분석기마다 형태소를 분류해내는 기준이 다르기 때문에 텍스트의 종류에 따라 유의미한 성능 차이가 발생한다. 텍스트는 상황 및 목적에 따라 작성 형태가 달라지기 때문에 이러한 특징을 고려하여 데이터에 따라 형태소 분석기를 잘 선택할 필요가 있다.

건설 재해사례 텍스트 데이터는 특수 산업분야에 대한 데이터이기 때문에 합성어가 많고, 작성자에 따라 서술 형태가 달라 합성어들(e.g., '안전난간' → '안전', '난간')이 많이 분리될수록 문장의 의미가 변질될 가능성이 있다. 따라서 본 연구에서는 합성어의 의미 손실을 최소화하면서 문장을 분석할 수 있는 Hannanum을 활용하였다. 또한, 1차적으로 형태소를 분석한 후 전처리할 때 단어들이 많이 훼손되는 것보다는 분석기가 분리해내지 못한 단어들을 직접 분리하는 작업이 더 합리적이라 판단하였다.

3.2.2 데이터 가공: 벡터화

비정형 텍스트를 분석하기 위해서는 문자를 기계가 인식할 수 있는 벡터화된 숫자로 임베딩(embedding), 즉 변환해야 한다. 이 중 단어 간의 유사도를 바탕으로 단어를 벡터화하는 학습 과정을 워드 임베딩이라고 하며, 벡터값이 비슷한 단어들은 유사한 의미와 관계를 가진다고 유추할 수 있다.

본 연구에서는 워드 임베딩의 한 방법론인 워드투벡터(Word2Vec) 모델을 사용하였다. 2013년 구글에서 발표한 Word2Vec (T.Milkolov et al., 2013)은 간단한 인공신경망 모형을 기반으로 단시간에 양질의 단어 벡터를 도출한다. 이에 현재 가장 많이 사용되는 워드 임베딩 모델로, Word2Vec 개발 이후 많은 분야에서 텍스트 기반의 딥러닝 연구가 진행될 수 있었다.

〈Fig. 1〉, 〈Fig. 2〉는 전체 사고 사례를 바탕으로 단어들을 벡터화한 결과로, 장소, 나이 등 비슷한 속성을 지닌 단어들이 서로 가까이 분포되어 있음을 확인할 수 있다. 유사 단어가 많은 보고서 데이터는 텍스트의 유사도를 바탕으로 정보를 추출하면 의미가 전혀 다른 단어들이 추출될 수 있으며 오타가 있을 때는 제대로 된 값을 반환하지 못한다.

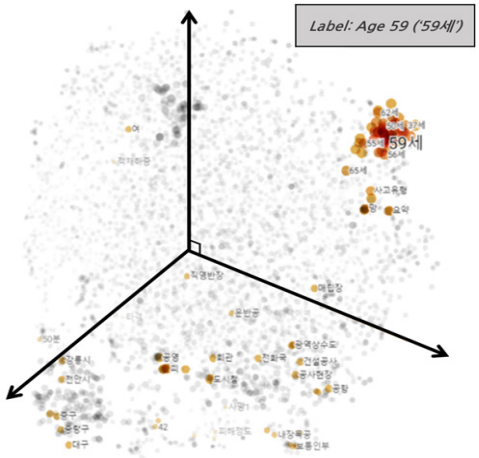

Fig. 1. Cluster of 'Age'

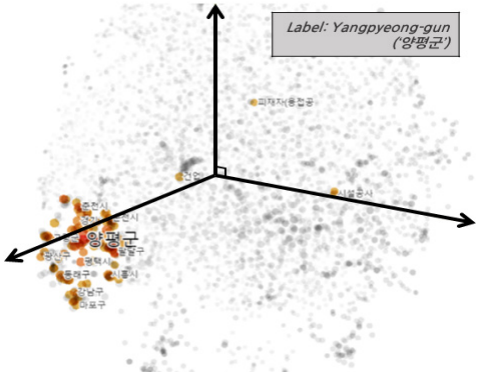

Fig. 2. Cluster of 'Area'

이에 반해 단어를 벡터화하면 그 분포와 근처 단어들을 기반으로 정보를 표현하게 되어 문맥상 같은 의미를 지닌 용어는 표현이 다르더라도 벡터 상에서는 가까이 분포하게 된다. 또한, 각각 다른 사람이 작성한 문서에 나타나는 다양한 유의어와 오타 관련 문제 등 또한 단어 벡터화를 통해 해결 가능할 것으로 기대된다.

예를 들어, '추락'과 유의한 단어들인 '추락방호조치', '추락방망' '추락재해방지' 등을 벡터화하는 경우, 해당 단어들의 표현이 다름에도 불구하고 비교적 가까이 분포하고 있는 것을 확인할 수 있다〈Fig. 3〉. 또한, '추락'을 검색하였을 때 코사인 유사도를 기준으로 추락과 가장 가까운 단어들 중 일부는 '충돌', '사망한', '중심', '실족' 등이다〈Fig. 4〉.

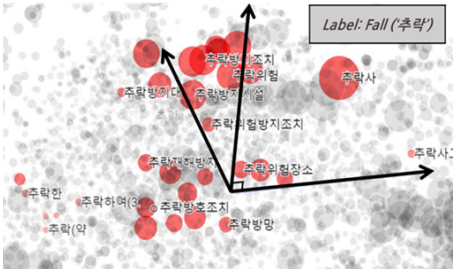

Fig. 3. Synonym of 'Fall'

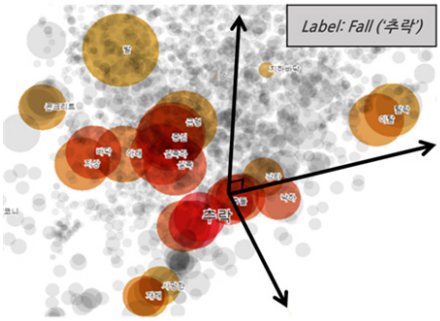

Fig. 4. Related terms with 'Fall'

이를 통해 산업안전보건공단의 건설 재해 데이터에는 근로자가 작업 중 중심을 잃어 실족한 사례가 많으며, 추락 도중 물체와 충돌하여 사망한 재해가 많이 발생한 것으로 유추된다. 이를 통해 별도의 전처리 없이도 같은 의미인 비슷한 생김새의 단어들과 그와 연관된 단어들을 학습한다는 것을 확인할 수 있다.

4. 분류 모델 구현 및 분석을 통한 방향 제안

본 장에서는 프레임워크에 따라 분류 모델링 및 분석이 진행된다〈Fig. 5〉. 먼저 CNN에 기반한 분류 모델을 구현하고, 3장에서 수집 및 전처리된 건설 사고 유형을 분류하는 실험을 수행한다. 3가지 측면에서 결과를 분석하고 개선 실험을 진행한 후, 이를 바탕으로 한국어 건설 재해 데이터의 효율적 관리를 위한 방향을 제안한다.

4.1 모델 개요

본 연구에서는 2장에서 고찰한 내용을 따라 프로그래밍을 위한 Python 및 데이터 분석을 위한 Pytorch의 nn패키지를 활용해 CNN 기반의 모델을 구현하였다. 모델의 세부 속성은 〈Table 2〉와 같다.

4.2 데이터 구성 및 분류 범주

〈Table 3〉은 3장에서 수집 및 전처리된 2,201개의 데이터 현황과 산업안전보건공단에서 산업재해를 분류 및 관리하

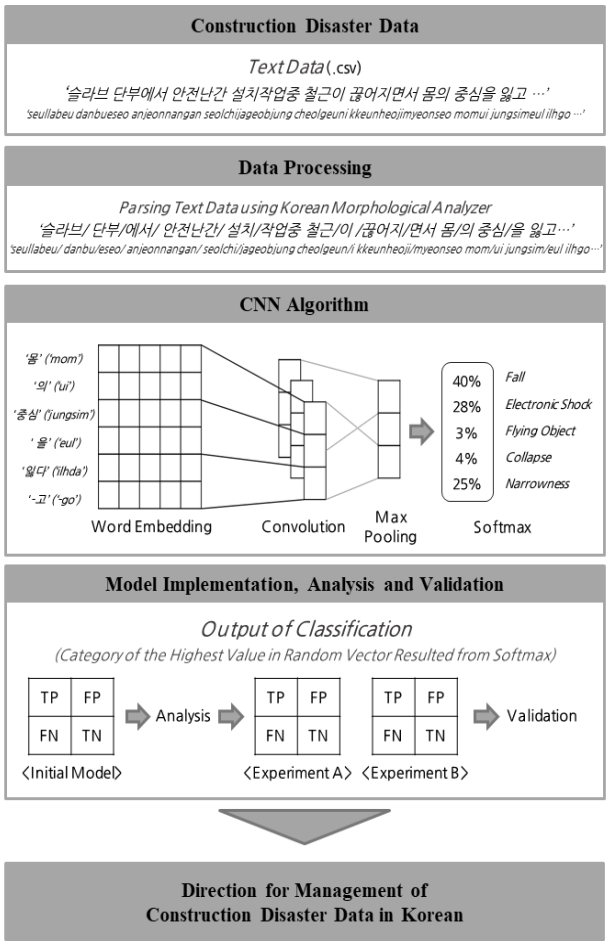

Fig. 5. Conceptual diagram modeling approach

Table 2. Hyper parameters for the model

Feature	Description
Language	Python
Package	nn Package (Pytorch)
Activation Function	ReLU
Loss-function	cross-entropy loss
Optimizer	Adam
Batch-size	64
Split-ratio	0.2
Learning rate	0.00001
Num-epochs	256
Dropout rate	0.5
Max-norm	3.0
Embedding dim	512
Kernal size	3,4,5,6,7

고 있는 13가지 종류의 사고 유형을 나타낸다. 본 연구의 분석 대상이 되는 데이터는 건설 사고이므로 추락이 큰 비중을 차지하는 반면, 익사 등은 그 비중이 상대적으로 작은 것을 알 수 있다.

Table 3. Accident type, sample and ratio of construction accident data

Accident Case	Sample	Ratio (%)
Fall	1,074	48.80%
Electric Shock	268	12.18%
Collapse	227	10.31%
Flying Object	172	7.81%
Narrowness & Winded	164	7.45%
Collision & Strike	112	5.09%
Inversion & Overturn	53	2.41%
Run over	41	1.86%
Oxygen deficiency & Suffocation	33	1.50%
Fire	20	0.91%
Explosion	19	0.86%
Others	18	0.82%
Pressure, Vibration	0	0.00%
Total	2,201	100.00%

분류 모델의 정확도는 분류 범주의 수 및 범주별 데이터 수에 영향을 데이터의 특성을 고려하여 알맞은 범주를 지정하는 것이 중요하다. 일반적으로 수학적 확률에 따라 범주의 수가 적을수록, 범주별로 데이터가 일정 수준 이상으로 고르게 분포할수록 정확도가 증가한다. 이에 따라 본 연구에서는 〈Table 3〉의 사고 유형 중 발생빈도가 가장 높은 5개 유형: 추락, 감전, 붕괴, 낙하, 협착에 한정하여 모델의 분류 범주를 구성하였다. 해당 5개 유형의 샘플은 총 1,905개로, 전체 중 86.6%를 차지한다. 다만 5개 유형에 해당하는 샘플 수가 매우 상이한 분포를 보임에 따라, 언더 샘플링을 통해 각 유형별 130개의 데이터를 무작위 추출하여 총 650개의 실험 데이터를 확보하였다.

4.3 모델 측정지표(KPI)

모델의 분류 성능 결과로부터 데이터의 질적 특성 개선방안을 제시하기 위해서는 모델의 분류 성능을 올바르게 평가할 KPI (Key Performance Index)가 설정되어야 한다. 본 연구는 이러한 지표로서 식 (1)을 활용한다.

$$Accuracy(\%) = \frac{Number\ of\ correctly\ classified\ data}{All\ classified\ data} \times 100 \qquad (1)$$

본 모델은 5개의 분류 범주(i.e., 추락, 감전, 붕괴, 낙하, 협착)에 대한 각각의 정확도 및 모델 전체의 정확도를 나타내는 총 6개의 식을 통해 평가할 수 있다. 이를 하나의 매트릭스에 나타낸 것을 정오분류표(Confusion Matrix)라 하며, 본 연구에서는 이를 통해 모델 전체의 정확도를 중심으로 각각의 정확도에 대한 결과를 분석한다.

4.4 건설산업 사고 유형 분류 실험 결과 및 분석

4.4.1 모델 초기 정확도 실험 결과

모델 학습을 위해 전처리 과정을 거쳐 유형별로 130건씩 총 650건의 데이터가 활용되었으며, 훈련된 모델 테스트를 위한 데이터는 150건의 샘플로 구성되었다. 본 연구에서는 해당 데이터를 여러 번 학습시키는 과정을 거쳐 초기 모델을 구현하고, 150건의 데이터를 통한 테스트 결과, 초기 모델의 정확도는 28.00%가 도출되었다〈Fig. 6〉.

유형별 정확도를 세부적으로 살펴보면, 추락에 대한 정확도(60.00%)가 나머지 유형의 정확도에 비해 상대적으로 높았다. 반면 감전, 낙하, 붕괴, 협착의 정확도는 비교적 낮게 도출되었는데, 특히 각 유형에 해당하는 사고를 추락으로 분류하는 경우가 많았다.

<Initial Model>

Accuracy 28.00%		Classified: Fall	Electric Shock	Flying Object	Collapse	Narrowness
True	Fall	18 60.00%	3 10.00%	3 10.00%	4 13.33%	2 6.67%
	Electric Shock	10 33.33%	4 13.33%	6 20.00%	9 30.00%	1 3.33%
	Flying Object	13 43.33%	3 10.00%	7 23.33%	0 0.00%	7 23.33%
	Collapse	10 33.33%	4 13.33%	10 33.33%	6 20.00%	0 0.00%
	Narrowness	10 33.33%	7 23.33%	0 0.00%	4 13.33%	9 30.00%

Fig. 6. Confusion matrix of the initial model

4.4.2 결과 분석

4.4.1의 초기 모델 테스트 결과 도출된 특징은 다음과 같다:

- 추락 재해의 분류 정확도가 상대적으로 높게 도출
- 타 유형을 추락 재해로 분류하는 경우가 비교적 많이 발생

이에 대한 원인 분석을 위해 4.4.1의 학습 및 테스트에 활용되지 않고 남아 있는 샘플 중에서 30개를 랜덤 추출한 후 원인 분석 실험을 수행하였다. 그 결과, 1) 구체적인 사고 유발 행동 묘사, 2) 유사한 문장 구조, 3) 여러 유형에 해당되는 복합사고가 위에 제시한 2가지 특징에 영향을 미치는 것으로 나타났다.

1) 구체적인 사고 유발 행동 묘사

원인 분석 실험 결과, 사고 유발 행동에 대한 구체적인 정보가 포함된 경우 그렇지 않은 경우에 비해 올바르게 분류될 확률이 높았다. 다음은 이와 관련된 사례이다.

(사례3) 및 *(사례4)의* 경우, '거푸집이 해체되면서', '몸의 중심을 잃고', '상단에서 이동중 실족', '배근된 철근에 걸려 넘어지면서' 등과 같이 사고 유발 행동이 비교적 구체적으로 묘사되어 있음을 알 수 있다. 반면 *(사례5)* 및 *(사례6)*은 구체적 서술이 생략되어 있다. 산업안전에 대한 기본 개념을 수립한 Heinrich (1941)의 사고발생모델에 따르면 사고의 60%가 불안전한 '행동'과 관련되어 있으며, 이후 수행된 많은 연구들(Stanton & Willenbrock, 1990; Abdelamid & Everett, 2000; Low et al., 2018; Kale & Baradan, 2020) 또한 여전히 이를 뒷받침하고 있다. 이러한 관점에서, 데이터가 사고 유발 행동 관련 서술을 포함하는 경우 모델이 이를 사고 유형을 특정할 지표로 활용함으로써 정확도가 증가하는 것으로 보인다. 상대적으로 분류가 용이했던 추락의 경우 구체적인 사고 유발 행동이 묘사된 경우가 많은 반면, 나머지 유형들은 이러한 묘사가 대부분 생략된 채 서술되었다.

구체적인 사고 유발 행동 포함 사례:

(사례3)…거푸집이 해체되면서 동시에 몸의 중심을 잃고 약 10m 아래 지상 바닥으로 추락…

(사례4)…상단에서 이동중 실족 또는 배근된 철근에 걸려 넘어지면서 5.1m 아래 Con'c 바닥으로 추락…

구체적인 사고 유발 행동 미포함 사례:

(사례5)…보형틀 설치 중 외부비계와 건물 사이 공간으로 추락…

(사례6)…천정 견출 작업중 5.6m 아래로 추락…

2) 유사한 문장 구조

서로 다른 유형임에도 불구하고 데이터의 문장 구조가 유사해 정확한 분류가 힘든 사례들이 존재한다:

(사례7)…크레인을 이용하여 고정용 앵글을 장물차에 상차하던 중 크레인의 Wire Rope가 파단되면서 앵글묶음이 낙하하여 피재자를 강타, 사망한 재해임.

(사례8)…트럭에 적재된 강관파이프 다발을 하역하기 위하여 타워크레인 인양로프를 파이프다발에 걸고 들어 올린 다음 하부에 각재를 넣고 지지하던중 각재가 부러지면서 파이프 다발이 지상으로 미끄러져 피재자가 협착 사망한 재해임.

(사례9)…이동식크레인을 이용하여 비계파이프 다발을 인양하던 중 와이어로프의 말단부 체결부위가 빠지며 비계파이프 다발이 낙하하며 피재자를 강타, 사망한 재해임.

(사례10)…덤프트럭(5ton)에 적재된 철도침목의 위치를 조정하기 위해 섬유로프를 푸는 순간 침목 묶음 상부에 올려져 있던 침목 10여개가 낙하하며 피재재를 강타, 사망한 재해임.

특히 추락, 낙하, 협착 유형은 문장 구조가 매우 유사한데

〈Fig. 7〉 물체의 파단이 발생한 뒤 사람이 떨어지면 추락, 기인물이 낙하해 사람이 맞으면 낙하, 그리고 기인물이 미끄러져 사람이 짓눌리면 협착 유형에 해당한다.

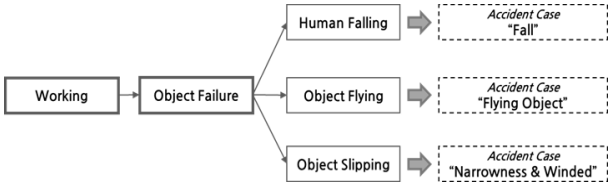

Fig. 7. Data misclassified as a fall or narrowness type

이처럼 작업, 기인물, 근로자의 행동이 유사한 경우 문장의 구조 또한 유사해 같은 사고로 분류될 수 있다. 특히 1)에서 서술한 이유로 상대적으로 추락을 잘 분류하는 모델이 다른 유형의 특징은 명확히 학습하지 못해 유사한 문장 구조를 가진 데이터를 모두 추락으로 잘못 분류하였을 가능성이 있다.

3) 여러 유형에 해당되는 복합사고

건설업의 특성상 재해는 단일 유형에 한정되는 것이 아니라 복수의 유형에 해당하는 경우가 빈번히 존재한다. 그러나 산업안전보건공단의 관리체계는 이러한 복합사고를 다음과 같이 '두 가지 이상의 유형이 연쇄적으로 발생한 산업재해기록/분류에 관한 지침'에 따라 13개의 유형 중 1개에 해당하도록 분류하고 있다(KOSHA, 2016):

> *재해의 경우는 상해결과 또는 피해를 크게 유발한 형태로 분류한다.*
>
> *재해자가 "전도"로 인하여 기계의 동력전달부위 등에 "협착"되어 신체 부위가 "절단"된 경우에는 "협착"으로 분류한다.*
>
> *재해자가 구조물 상부에서 "전도"로 인하여 "추락"되어 두 개골 골절이 발생한 경우에는 "추락"으로 분류한다.*
>
> *재해자가 전주에서 작업 중 "전류접촉"으로 "추락"한 경우 상해결과가, 골절인 경우에는 "추락"으로 분류하고, 상해결과가 전기쇼크인 경우에는 "전류접촉"으로 분류한다.*

산업안전보건공단의 건설 재해 데이터는 국내 건설 사고의 현황 파악 및 사실관계 보존을 주목적으로 생성 및 관리된다. 특히, 위의 기준은 근로자 상해에 대한 보상을 위해 수립된 것으로, 이러한 지침에 의한 분류가 실제 사고의 유형을 대변할 수 있는지에 대한 검증은 이루어진 바 없다. 따라서 복수의 유형에 해당함에도 5개 유형 중 1개로 한정되어 분류되어 있는 복합사고는 전체 데이터 학습 및 실험에 매우 부정적으로 작용하였을 가능성이 있다. 구체적으로, 복수의 유형 모두에 해당하는 복합사고는 유형 간 다중공선성을 야기함으로써 선형 독립을 해칠 우려가 있다.

4.5 분석 결과 검증: 개선 실험

4.4에서 도출된 건설 재해 데이터 분류 특성에 대한 원인 중 1) 구체적인 사고 유발 행동 및 2) 유사한 문장 구조의 경우 건설 재해 데이터 자체의 특성에 기인하는 불가피한 요인이라고 할 수 있다. 그러나 3) 여러 유형에 해당되는 복합사고의 경우 개선 실험을 통해 검증이 가능하다. 따라서 본 절에서는 복합사고의 다중공선성 문제를 2가지 방법으로 접근하여 검증하고자 한다.

초기 모델에서 나타난 특징은 결국 모델이 추락 이외의 유형에 해당하는 사고를 학습 및 분류하는 데 있어 유의미한 규칙을 찾기 어렵다는 것을 의미한다. 이에 본 연구에서는 다음과 같이 두 가지의 개선 실험을 계획하였다: 1) 복합사고 재분류 실험(Experiment A) 및 2) 복합사고 제외 실험(Experiment B). Experiment A는 산업안전보건공단에서 복합사고를 분류하는 기준을 모델에 적용함으로써 복합사고를 재분류하였을 때 정확도를 살펴보기 위함이며, Experiment B에서는 복합사고를 제외한 나머지 데이터에 대한 정확도 향상 정도를 관찰하고자 한다.

Table 4. The number of complex accidents by case

Accident Case	Total Accidents	Simple Accidents	Complex Accidents
Fall	1,074	792	282
Electric Shock	268	261	7
Flying Object	172	132	40
Collapse	227	191	36
Narrowness	164	145	19
Total	1,905	1,621	384

〈Table 4〉는 사고 유형별 데이터마다 등장하는 복합사고의 수를 나타낸 표이다. 복합사고는 모든 유형에서 유의미한 비중을 차지한다. 본 연구에서 다루는 복합사고는 한 사고 내에 두 가지 이상의 유형이 포함된 경우이다.

분류 규칙을 모델에 적용하여 재분류한 결과 초기 모델과 동일한 28.00%의 정확도가 도출되었다〈Fig. 8〉. 추락을 제외한 다른 범주들의 분류 정확도는 소폭 증가한 반면, 추락 재해의 분류 정확도는 오히려 감소하여 질(quality)은 다소 개선되었으나 분류 성능까지 개선했다고 보기는 어렵다.

두 번째 실험을 위해 사고 유형에 따른 분류에 부정적 영향을 줄 수 있다고 판단된 복합사고 데이터를 전체 데이터로부터 분리하였다. 〈Table 4〉에서 복합사고를 제외한 1,621개의 데이터 중 학습데이터는 1,395개, 테스트 데이터는 126개로 실험을 진행하였다. 분류 결과, 기존 대비 185.7% 향상된 정확도(52.00%)로 126개의 데이터 중 65개의 데이터를 분류하였다〈Fig. 9〉. 복합사고를 제거함에 따라

<Experiment A>

Accuracy 28.00%		Classified				
		Fall	Electric Shock	Flying Object	Collapse	Narrowness
True	Fall	15 50.00%	3 10.00%	3 10.00%	5 16.67%	4 13.33%
	Electric Shock	10 30.00%	7 23.3%	5 16.67%	7 30.33%	1 3.33%
	Flying Object	14 46.67%	4 13.33%	6 20.00%	0 0.00%	4 13.33%
	Collapse	9 30.00%	4 13.33%	7 23.33%	7 23.33%	3 10.00%
	Narrowness	8 26.67%	4 13.33%	4 13.33%	4 13.33%	10 33.33%

Fig. 8. Confusion matrix of Experiment A

<Experiment B>

Accuracy 52.00%		Classified				
		Fall	Electric Shock	Flying Object	Collapse	Narrowness
True	Fall	15 60.00%	2 8.00%	2 8.00%	4 16.00%	2 8.00%
	Electric Shock	3 12.00%	9 36.00%	4 16.00%	4 16.00%	5 20.00%
	Flying Object	4 16.00%	4 16.00%	15 60.00%	2 8.00%	1 4.00%
	Collapse	5 20.00%	2 8.00%	3 12.00%	12 48.00%	3 12.00%
	Narrowness	4 16.00%	2 8.00%	2 8.00%	3 12.00%	14 56.00%

Fig. 9. Confusion matrix of Experiment B

데이터의 수가 감소하였음에도 불구하고 정확도가 급격히 증가했다. 이는 데이터의 학습이 원활히 진행되었다는 것을 보여주며, 결국 복합사고가 건설 재해사례의 사고 유형별 분류 정확도를 저해하는 큰 장애 요소 중 하나라고 할 수 있다.

한편, CNN의 분류 정확도는 데이터의 수에 큰 영향을 받는다. 본 연구에서 모델 학습에 활용된 650개의 데이터는 높은 정확도를 도출하기에 다소 부족했을 수 있다. 특히 많은 수의 데이터 샘플을 전제로 하는 워드 임베딩은 정확도 테스트 과정에서 학습하지 않은 단어들이 등장하는 경우 해당 단어들을 제외시키거나 모두 같은 벡터로 인식하게 된다. 이로 인해 발생하는 오류는 모델의 정확도에 영향을 주며, 특히 데이터 샘플의 수가 이를 상쇄할 만큼 많지 않은 경우에는 그 영향이 더 클 수 있다.[2)]

4.6 한국어 건설 재해 데이터의 효율적 관리방향

4.4를 통한 분석 결과 1) 구체적인 사고 유발 행동, 2) 유사한 문장 구조, 3) 여러 유형에 해당되는 복합사고가 건설 재해 데이터의 분류에 영향을 미치는 것으로 분석되었다. 특히 4.5에서는 2가지 실험(Experiment A, Experiment B)을 통해 복합사고가 건설 재해 데이터 분류에 있어 다중공선성을 유발하는 주요 요인임을 검증할 수 있었다. 구체적으로, Experiment A에서 복합사고를 분류하는 규칙을 추가하더라도 정확도가 개선되지 않은 반면, Experiment B에서 복합사고를 제외한 데이터 분류 결과 정확도가 현저히 개선되었다. 이를 바탕으로 본 연구에서는 다음과 같이 한국어 건설 재해 데이터의 효율적 관리방향을 제안한다.

우선, 사고에 대한 상황을 상세히 서술하는 체계로 변화하는 것이 필요하다. 전술한 바와 같이, 건설 사고의 예방을 위해서는 사고의 근본적 원인을 파악하는 것이 선행되어야 한다. 즉 과거의 사고로부터 유의미한 결과를 도출하기 위해서는 당시의 상황에 대한 자세한 묘사가 필요하다. 특히 4.4.2에서 다른 유형에 비해 높은 정확도로 분류되었던 추락 사고에서 상세한 서술이 빈번히 나타났음을 고려할 때, 향후 이를 체계적으로 관리한다면 유형별 유사한 문장 구조 발생을 지양할 수 있을 뿐만 아니라 사고의 근본적 원인을 파악할 수 있는 기초 자료로 활용될 수 있을 것으로 생각된다.

두 번째로, 사고를 단일사고와 복합사고로 분리하여 관리하는 것이 필요하다. 4.5에서 복합사고를 기준에 따라 분류하는 규칙을 적용한 실험(Experiment A)에서도 정확도가 향상되지 않은 반면, 복합사고를 제외한 실험(Experiment B)에서는 정확도가 현저히 향상되었다. 이는 복합사고를 인위적으로 단일사고로 분류하는 기존 방식의 실효성에 의문을 제기할 뿐만 아니라 유형 간 다중공선성을 유발함으로써건설 재해 데이터 관리의 한계로 작용할 수 있다. 따라서 건설 사고에서 중요한 의미를 갖는 복합사고를 단일사고와 별도로 관리 및 연구하는 것이 필요하다. 이러한 체계는 비로소 건설 사고의 근본적 원인을 파악하기 위한 초석으로 작용할 수 있다.

2) 머신러닝 기반 모델의 분류 정확도는 데이터의 샘플 수 및 질적 특성에 따라 크게 달라지기 때문에 모델 정확도 결과의 타당성에 대한 표준화된 기준은 존재하지 않으며, 데이터와 함께 고려되어야 한다. 다만 유사한 데이터를 통해 수행한 연구(Khallaf & Khallaf, 2021; Zhong et al., 2020; Chi et al., 2016; Williams & Gong, 2014)의 경우 참조할 가치가 상당하다. 국내 건설 재해 분야 분류 모델의 정확도를 살펴보면 아직까지는 30% 근처에서 도출되는 것으로 보인다(Kim, 2017; Kim et al., 2017). 다만 이는 건설산업의 데이터에만 국한된 문제는 아니며(Sleeman IV & Krawczyk, 2021), 텍스트 데이터와 같이 비정형의 형태를 띌수록 정확도는 감소하는 경향성을 나타낸다. 더욱이 5개 이상의 다중유형을 분류하는 것은 그 이하의 유형으로 분류할 때보다 어렵다고 할 수 있다(Abramovich & Pensky, 2019). 이러한 맥락을 두루 고려할 때, 본 모델의 초기 및 최종 정확도는 비교적 적정한 것으로 판단된다.

5. 결론

본 연구는 산업안전보건공단의 건설재해사례 데이터를 기반으로 한국어로 작성된 건설 재해 데이터의 분류 성능을 개선하고자 하였다. 이러한 목표 아래 CNN 기반 사고 유형 분류 모델을 구현하고 실험을 진행하였다. 초기 모델 테스트 결과, 추락 재해의 분류 정확도가 상대적으로 높게 도출되며 타 유형을 추락 재해로 분류하는 경우가 많이 발생한다는 특징이 나타났다. 분석 결과 건설 중대 재해사례의 사고 유형별 분류 정확도를 저해하는 큰 원인 중 하나가 복합사고라는 결과를 도출하였으며, 복합사고를 재분류하는 방법보다 데이터를 제외하는 것이 정확도 향상에 크게 기여하였다. 따라서 건설현장에서 발생하는 사고에 대한 데이터를 수집 및 축적 시 복합사고를 독립적으로 관리할 필요가 있다.

본 연구는 복합사고에 대한 기록이 건설 재해 데이터의 분석 성능을 저해하는 요소임을 밝히고, 산업재해 보상을 위한 분류 기준이 아닌 빅데이터 시대의 데이터 분석에 유용한 건설 사고 분류 기준을 수립의 필요성을 제안하였다는 점에서 학술적, 실무적 기여도를 가진다. 향후 연구자들이 한국어 건설 재해 데이터를 분석할 때 복합사고의 정제 및 처리에 주의를 기울인다면, 안전사고 데이터 관리와 이를 바탕으로 개발되는 대책 및 시스템 등은 실무에 보다 적합한 수준에 도달함으로써 궁극적으로 안전관리에 기여할 것으로 기대된다. 마지막으로 본 연구는 한국어 데이터의 정성적 특징을 파악하고 정제하기 위해 워드 임베딩 기법을 적용하였다. 이를 통해 외래어와 약어, 영문이 혼용된 건설 문서 데이터 분석의 한계를 극복하고자 했다는 점에서 의의를 가진다.

본 연구의 결과 모델은 '추락'이라는 용어가 등장하지 않더라도 텍스트 속의 정보들을 종합하여 '떨어져 사망', 또는 '(사람이) 낙하함' 등으로 표현된 데이터 또한 '추락'으로 분류할 수 있다. 다만 CNN 알고리즘의 성능이 데이터의 양과 질에 따라 결정되는데, 이로 인해 본 연구는 정확도 측면에서 한계를 가진다. 그러나 분석이 상대적으로 어려운 건설 텍스트 데이터를 다중 유형으로 분류하고자 시도하였다는 데 그 의의가 있다. 딥러닝 모델은 추후 지속적인 데이터 축적을 통한 자가학습에 의해 성능이 점진적으로 개선될 수 있으며, 본 연구는 모델의 활용성 제고를 위해 이를 전제로 CNN을 활용하였다. 향후 개선된 모델을 통해 건설 재해 데이터를 보다 심층적으로 분류함으로써(e.g., 기인물, 부상 위치, 부상 강도) 체계적인 안전관리 연구에 기여할 수 있을 것으로 생각된다. 이는 사고 데이터에 내재된 지식이 사고 예방으로 이어질 수 없었던 현행 건설 안전관리의 한계를 극복하고, 궁극적으로 건설 재해 데이터의 활용성을 제고할 수 있을 것으로 기대된다.

감사의 글

이 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었음(과제번호 21CTAP-C152263-03).

References

Abdelhamid, T.S., and Everett, J.G. (2000). "Identifying root causes of construction accidents." *Journal of construction engineering and management,* 126(1), pp. 52-60.

Abramovich, F., and Pensky, M. (2019). "Classification with many classes: challenges and pluses." *Journal of Multivariate Analysis,* 174, p. 104536.

Ahuja, V., Yang, J., and Shankar, R. (2010). "Benchmarking framework to measure extent of ict adoption for building project management." *Journal of Construction Engineering and Management,* 136(5), pp. 538-545.

Al Qady, M., and Kandil, A. (2015). "Automatic Classification of Project Documents on the Basis of Text Content." *Journal of Computing in Civil Engineering,* 29(3), p. 04014043.

Amiri M., Ardeshir, A., Fazel Zarandi, M.H., and Soltanaghaei, E. (2016). "Pattern extraction for high-risk accidents in the construction industry: a data-mining approach." *International Journal of Injury Control and Safety Promotion,* 23(3), pp. 264-276.

Caldas, C.H., Soibelman, L., and Han, J. (2002). "Automated Classification of Construction Project Documents." *Journal of Computing in Civil Engineering,* 16(4), pp. 234-243.

CERIK (2014). CERIK Research Report, 2014.

Chassiakos, A., and Sakellaropoulos, S. (2008). "A web-based system for managing construction information." *Advances in Engineering Software,* 39(11), pp. 865-876.

Chi, N.W., Lin, K.Y., El-Gohary, N., and Hsieh, S.H. (2016). "Evaluating the strength of text classification categories for supporting construction field inspection." *Automation in Construction,* 64, pp. 78-88.

Choi, J.K. (2019). "A Prediction Model for Fatal Accidents among Construction Workers using Machine Learning." MS thesis, Sungkyunkwan Univ., Korea.

Chokor, A., Naganathan, H., Chong, W.K., and Asmar, M.E. (2016). "Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning." *Procedia Engineering,* 145, pp. 1588-1693.

Cui, T., Wu, Y., and Tong, Y. (2018). "Exploring ideation and implementation openness in open innovation projects: IT-enabled absorptive capacity perspective." *Information & Management,* 55(5), pp. 576-587.

Famous, G. (2018). "Three Technology Trends Shaping the Future of Design and Construction in 2018." *Aconex Group,* 〈https://blogs.oracle.com〉 (Feb. 20, 2021)

Heinrich, H.W. (1941). *Industrial Accident Prevention: A Scientific Approach,* 2nd ed.

Hill, B.L. (2017). "Digging for the Big Data Gold in Today's Construction Projects." *Xpera Group,* 〈https://www.xperagroup.com〉 (Feb. 20, 2021)

IBM (2015). IBM Annual Report, 2015.

Jung, J.M. (2018). "A study of improvement of deep learning performance for document classification using the word class." MS thesis, Korea Univ., Korea.

Kale, Ö.A., and Baradan, S. (2020). "Identifying factors that contribute to severity of construction injuries using logistic regression model." *Teknik Dergi,* 31(2), pp. 9919-9940.

Kang, H.B., and Yi, J.S. (2018). "An Analysis of Public Text Data in Construction Disaster Cases using Word2Vec-based Data Visualization." *Proceedings of the 2018 Architectural Institute of Korea Conference,* 38(2), pp. 567-570.

Khallaf, R., and Khallaf, M. (2021). "Classification and analysis of deep learning applications in construction: A systematic literature review." *Automation in Construction,* 129, p. 103760.

Kim, D.C., and Kim, H.J. (2001). "A Plan of the Accident Classification System for the Analysis of Disaster Information in Construction Projects." *Journal of the Architectural Institute of Korea: Structure & Construction,* 17(11), pp. 139-145.

Kim, Y. (2014). "Convolutional Neural Networks for Sentence Classification." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* EMNLP. pp. 1746-1751.

Kim, Y.C. (2017). "A study on safety accident prediction using data mining technique at domestic construction site." MS thesis, Kyonggi Univ., Korea.

Kim, Y.C., Yoo, W.S., and Shin, Y. (2017). "Application of Artificial Neural Networks to Prediction of Construction Safety Accidents." *Journal of the Korean Society of Hazard and Mitigation,* 17(1), pp. 7-14.

KOSHA (2020). Statistics of Industrial accident 2019, 2020.

Levy, O. and Goldberg, Y. (2014). "Neural word embedding as implicit matrix factorization." In *Advances in neural information processing systems,* pp. 2177-2185.

Liddy, E.D. (2001). *Natural language processing Encyclopedia of Library and Information Science.* NY: Marcel Decker. Inc.

Low, B.K.L., Man, S.S., and Chan, A.H.S. (2018). "The risk-taking propensity of construction workers—An application of Quasi-expert interview." *International journal of environmental research and public health,* 15(10), p. 2250.

Martínez-Rojas, M., Marín, N., and Vila, M.A. (2013). "A preliminary approach to classify work descriptions in construction projects." *IFSA World Congress and NAFIPS Annual Meeting, 2013 Joint, IEEE,* Washington, DC, pp. 1090-1095.

Marzouk, M., and Enaba, M. (2019). "Text analytics to analyze and monitor construction project contract and correspondence." *Automation in Construction,* 98, pp. 265-274.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013). "Distributed Representations of Words and Phrases and their Compositionality." A*dvances in Neural Information Processing Systems,* pp. 3111-3119.

Moon, S., Kim, T., Hwang, B.G., and Chi, S. (2018). "Analysis of construction accidents based on semantic search and natural language processing." *2018 International Symposium on Automation and Robotics in Construction and International AEC/FM Hackathon: The Future of Building Things.*

Moon, S., Kim, T., Hwang, B.G., and Chi, S. (2018). "Document Management System Using Text Mining for Information Acquisition of International Construction." *Journal of Civil Engineering,* KSCE, 22(12), pp. 4791-4798.

Nam, G.I., and Jo, E.G. (2017). *Korean Text Sentiment Analysis,* Communication-Books.

Ok, H. Kim, S.J., and Seo, M.B. (2013). "A Study on the Improvement of the Domestic Construction Information Classification System." *Proceedings of the 2013 Korean Institute of Information Scientists and Engineers Conference,* pp. 25-27.

Park, E.J., and Cho, S.Z. (2014). "KoNLPy: Korean natural language processing in Python." *26th Annual Conference on Human and Language Technology,* pp. 133-136.

Park, H.J, Song, M.C., and Shin, K.S. (2018). "Sentiment Analysis of Korean Reviews Using CNN - Focusing on Morpheme Embedding -." *Journal of Intelligence and Information Systems,* (24)2, pp. 59-83.

Park. T.Y, Han H.J., Kim, Y., and Kim, S.J. (2017). "A Study on the Analysis and Improvement of Classifications for Integrated Management of Disaster and Safety Information." *Korean Biblia Society for Library and Information Science,* 28(3), pp. 125-150.

Sardroud, J.M. (2015). "Perceptions of automated data collection technology use in the construction industry." *Journal of Civil Engineering and Management,* 21(1), pp. 54-66.

Sleeman IV, W.C., and Krawczyk, B. (2021). "Multi-

class imbalanced big data classification on Spark." *Knowledge-Based Systems,* 212, p. 106598.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I., and Lin, K.Y. (2008). "Management and analysis of unstructured construction data types." *Advanced Engineering Informatics,* 22(1), pp. 15-27.

Stanton, W.A., and Willenbrock, J.H. (1990). "Conceptual framework for computer-based, construction safety control." *Journal of Construction Engineering and Management,* 116(3), pp. 383-398.

Tixier, A.J.-P., Hallowell, M.R., Rajagopalan, B., and Bowman, D. (2016). "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports." *Automation in Construction,* 62, pp. 45-56.

Ubeynarayana, C.U., and Goh, Y.M. (2017). "An Ensemble Approach for Classification of Accident Narratives". *ASCE International Workshop on Computing in Civil Engineering* 2017, pp. 409-416.

Villanueva, V., and Garcia, A.M. (2011). "Individual and occupational factors related to fatal occupational injuries: a case-control study." *Accident Analysis and Prevention,* 43(1), pp. 123-127.

Williams, T.P., and Gong, J. (2014). "Predicting construction cost overruns using text mining, numerical data and ensemble classifiers." *Automation in Construction,* 43, pp. 23-29.

Yang Y.S., Park J.H., and Lee C.S. (2009). "Accident Risk Analysis of Construction Workers by Occupation." Journal of the Architectural Institute of Korea: *Structure & Construction,* 25(10), pp. 149-156.

Yi, K.J. (2005). "Construction Workers' Occupational Risk of On-Site Travelling Activities." *Korean Journal of Construction Engineering and Management,* KICEM, 6(3), pp. 120-127.

You, Z., and Wu, C. (2019). "A framework for data-driven informatization of the construction company." *Advanced Engineering Informatics,* 39, pp. 269-277.

Zhong, B., Pan, X., Love, P. E., Ding, L., and Fang, W. (2020). "Deep learning and network analysis: Classifying and visualizing accident narratives in construction." *Automation in Construction,* 113, p. 103089.

Zou, Y., Kiviniemi, A., and Jones, S.W. (2017). "Retrieving similar cases for construction project risk management using Natural Language Processing techniques." *Automation in construction,* 80, pp. 66-76.

요약 : 본 연구는 딥러닝 기반의 텍스트 데이터 분류 모델의 성능 고찰을 통해 한국어 건설 재해사례의 효율적 관리방향을 제안한다. 이를 위해 비정형 텍스트 문서인 건설 재해 보고서를 활용해 건설 사고의 대표적 유형인 추락, 감전, 낙하, 붕괴, 협착의 5개 범주로 분류하는 딥러닝 모델을 구현하였다. 초기 모델 테스트 결과, 추락 재해의 분류 정확도가 상대적으로 높게 도출되며 타 유형을 추락 재해로 분류하는 경우가 많이 발생한다는 특징이 나타났다. 원인 분석 결과, 1) 구체적인 사고 유발 행동, 2) 유사한 문장 구조, 3) 여러 유형에 해당되는 복합사고가 위의 특징에 영향을 미치는 것으로 분석되었으며, 이 중 추가 실험을 통해 검증이 가능한 복합사고에 대한 두 가지 정확도 개선 실험을 진행하였다: 1) 재분류, 2) 제외. 실험 결과, 복합사고 제외 시 분류 성능이 185.7% 향상되었으며, 이를 통해 여러 사고 유형에 대한 내용을 동시에 포함하는 복합사고의 다중공선성(multicollinearity)이 해소되었음을 알 수 있다. 결론적으로 본 연구에서는 향후 사고에 대한 상황을 상세히 서술하는 체계를 마련함과 동시에 복합사고를 독립적으로 관리할 필요성을 시사한다.

키워드 : 건설 안전, CNN, 딥러닝, 분류 모델, 재해 데이터