KSCE Journal of Civil and Environmental Engineering Research
Vol. 42, No. 1: 127-134/ February, 2022
DOI: https://doi.org/10.12652/Ksce.2022.42.1.0127

ISSN 1015-6348 (Print)
ISSN 2287-934X (Online)
www.kscejournal.or.kr

Information Technology **정보기술**

건설현장 정형·비정형데이터를 활용한 기계학습 기반의 건설재해 예측 모델 개발

조민건* · 이동환** · 박주영*** · 박승희****
Cho, Mingeon*, Lee, Donghwan**, Park, Jooyoung***, Park, Seunghee****

Development of Machine Learning-based Construction Accident Prediction Model Using Structured and Unstructured Data of Construction Sites

ABSTRACT

Recently, policies and research to prevent increasing construction accidents have been actively conducted in the domestic construction industry. In previous studies, the prediction model developed to prevent construction accidents mainly used only structured data, so various characteristics of construction sites are not sufficiently considered. Therefore, in this study, we developed a machine learning-based construction accident prediction model that enables the characteristics of construction sites to be considered sufficiently by using both structured and text-type unstructured data. In this study, 6,826 cases of construction accident data were collected from the Construction Safety Management Integrated Information (CSI) for machine learning. The Decision forest algorithm and the BERT language model were used to train structured and unstructured data respectively. As a result of analysis using both types of data, it was confirmed that the prediction accuracy was 95.41 %, which is improved by about 20 % compared to the case of using only structured data. Conclusively, the performance of the predictive model was effectively improved by using the unstructured data together, and construction accidents can be expected to be reduced through more accurate prediction.

Key words : Construction accident, Prediction model, Machine learning, BERT, Decision forest

초 록

현재 국내 건설업에서는 꾸준히 증가하는 건설재해를 예방하기 위해 다양한 정책적 노력과 연구가 활발하게 진행되고 있다. 기존 연구에서 건설재해 예방을 위해 개발한 예측 모델의 경우, 주로 정형데이터만을 활용하였기에 건설현장의 다양한 특성을 충분히 고려하지 못한 예측 결과가 도출되었다. 따라서, 본 연구에서는 정형데이터와 텍스트 형식의 비정형데이터를 동시에 활용하여 건설현장의 특성을 충분히 고려할 수 있는 기계학습 기반 건설재해 사전 예측 모델을 개발하였다. 본 연구는 기계학습을 위해 건설공사 안전관리 종합정보망(CSI)의 최근 3년간 건설재해 데이터 6,826건을 수집하였다. 수집된 데이터 중 정형데이터의 학습은 5가지 알고리즘의 성능 분석을 통해 Decision forest 알고리즘을 사용하였고 비정형데이터의 학습은 BERT 언어모델을 사용하였다. 정형 및 비정형데이터를 동시에 활용한 건설재해 예측 모델의 성능 비교 결과, 정형데이터만을 활용한 경우보다 약 20 % 향상된 95.41 %의 예측정확도가 도출되었다. 본 연구 결과, 비정형데이터를 동시에 활용함으로써 예측 모델의 효과적인 성능 향상을 확인하였으며, 보다 정확한 예측을 통한 건설재해 저감을 기대할 수 있다.

검색어 : 건설재해, 예측모델, 기계학습, BERT, Decision Forest

* 정회원 · 성균관대학교 미래도시융합공학과 석사과정 (Sungkyunkwan University · raonik6713@naver.com)
** 종신회원 · 성균관대학교 미래도시융합공학과 연구교수, 공학박사 (Sungkyunkwan University · ycleedh@gmail.com)
*** 성균관대학교 건설환경시스템공학과 박사과정, 공학석사 (Sungkyunkwan University · mitjy@gmail.com)
**** 종신회원 · 교신저자 · 성균관대학교 건설환경공학부 교수, 공학박사 (**Corresponding Author** · Sungkyunkwan University · shparkpc@skku.edu)

Received November 5, 2021/ **revised** December 10, 2021/ **accepted** December 16, 2021

1. 서 론

고용노동부의 산업재해 현황분석에 따르면 국내 건설업의 경우, 근로자 1,000명당 발생하는 재해자수의 비율인 재해천인율이 2010년 7.03 ‰에서 2019년 10.94 ‰로 10년간 꾸준히 증가하는 것으로 나타났다. 더불어, 재해로 인한 강도율 또한 함께 증가하여 2019년을 기준으로 전체산업 대비 2배 이상의 높은 수치를 기록하였다(Ministry of Employment and Labor, 2020). 이러한 높은 강도율은 대형 및 사망 사고의 비율이 높아 큰 피해가 발생할 가능성이 크다는 것을 의미한다(Kim, 2008). 특히, 최근 2010년부터 2019년까지 10년간 건설업의 대형사고 발생건수 및 사망자수는 전체산업 중 50 %가 넘는 것으로 나타났으며(Korea Occupational Safety and Health Agency, 2019), OECD 국가의 건설업 산재 사망사고 실태 비교·분석 보고서에 따르면 국내 건설업 근로자수는 OECD 35개 회원국 중 7번째인 반면에 사고사망자수는 2번째로 근로자 대비 많은 사망사고가 발생하는 것으로 나타났다(Choi, 2020). 전 산업에서 가장 많은 중대 재해를 발생시키는 건설재해는 직적접인 경제적 손실뿐 아니라 국가신인도의 저하 및 사회적 불안감을 고조시킬 수 있어 국가적인 부담 요인으로 적용한다(Cho et al., 2017).

이에 따라, 정부 차원에선 건설재해 관리를 목적으로 중·장기 건설업 산재예방 정책과제를 체계적으로 수립하여 대규모 프로그램들을 실행해왔다(Korea Labor Institute, 2013). 특히, 현 정부는 건설재해 감소를 주요 과제 중 하나로 설정하고 “사전예방형”, “발주자”라는 두 개의 핵심키워드를 바탕으로 안전관리체계의 전환을 도모하고 있다(Lim et al., 2019).

건설재해 저감을 위한 다양한 연구 또한 진행되어왔다. Lim et al.(2019), Yu et al.(2016) 그리고 Cho(2012)는 건설재해 원인분석을 통해 재해위험요인 도출하였고 재해예방대책을 제안하였다. Park and Kim(2021), Zhang et al.(2019) 그리고 Lee(2018)는 건설재해에 관련된 비정형 텍스트 데이터를 분석하여 재해위험요인 도출하였고 건설재해 예방을 위한 비정형 텍스트 데이터 활용의 중요성을 강조하였다. 마지막으로, Choi et al.(2021), Kim et al.(2017) 그리고 Cho et al.(2017)는 정형데이터의 지도학습을 통해 건설재해를 사전에 예측할 수 있는 모델을 제안하였다. 그러나 주로 건설재해의 원인을 분석한 연구의 경우, 사후분석적 연구결과를 도출하였기에, 사전적 건설재해 예방에는 한계가 있다. 이러한 한계를 극복하기 위한 기계학습 기반의 예측 모델 개발 연구의 경우, 건설현장의 다양한 현장 특성을 포함하는 비정형데이터를 활용하지 않고 정형데이터만을 활용한 예측 모델을 제안하였기에, 다양하고 동적인 건설현장 조건을 고려하기에 충분하지 않다는 한계가 존재한다.

따라서, 본 연구에서는 다양한 건설현장 조건을 고려하여 건설재해를 예방하기 위해 사전에 파악할 수 있는 정형 및 비정형 텍스트 데이터를 동시에 활용한 기계학습 기반의 건설재해 예측 모델을 개발하였다. 모델 개발에 사용한 학습데이터는 건설공사 안전관리 종합정보망(CSI)에서 최근 3년간 발생한 건설재해 6,837건을 수집하여 사용하였다. 또한, 건설재해 예측을 위한 최적의 모델을 선별하기 위하여 정형데이터만을 활용한 모델, 비정형데이터만을 활용한 모델 그리고 정형 및 비정형데이터를 모두 고려한 모델 총 3가지 유형의 모델을 개발하고 성능을 정량적 지표를 통해 비교하였다. 구체적으로, 정형데이터는 다중클래스 분류(Multi-class classification)에 대표적으로 적용되는 기계학습 알고리즘인 Logistic Regression (LOGIT), Artificial Neural Network (ANN), Support Vector Machine (SVM), Naive Bayes (NB), Decision Forest (DF) 총 5가지 알고리즘을 적용하였고, 비정형데이터는 자연어처리 분야에서 우수한 성능을 보이는 Bidirectional Encoder Representations from Transformers (BERT) 언어모델을 적용하였다.

2. 건설재해 예측 모델 개발

본 연구에서 제안하는 정형 및 비정형데이터를 모두 고려할 수 있는 건설재해 사전 예측 모델 개발은 데이터 수집(Data collection), 데이터 전처리(Pre-processing), 예측 모델 개발(Model development), 모델 분석(Anaysis) 4단계로 진행되며, 전체 프로세스는 Fig. 1과 같다.

데이터 수집단계(Fig. 1(a))에서는 건설공사 안전관리 종합정보망(CSI)에 공개된 건설업 재해사례를 웹 크롤링(Web crawling)을 통해 수집하였고, 데이터 전처리단계(Fig. 1(b))에서는 수집한 데이터를 정형 및 비정형데이터로 구분하여 기계학습 및 BERT 알고리즘에 적용할 수 있도록 전처리를 수행하였다. 예측 모델 개발단계(Fig. 1(c))에서는 건설재해를 예측할 수 있는 최적의 모델을 선별하기 위해 첫 번째로 정형데이터만을 사용한 기계학습 모델, 두 번째로 비정형데이터만을 사용한 BERT 모델, 마지막으로 정형 및 비정형데이터를 모두 고려한 기계학습 모델을 개발하였다. 이때, 기계학습의 경우 다양한 방법(LOGIT, ANN, SVM, NB, DF)을 적용하였다. 모델 분석단계(Fig. 1(d))에서는 개발한 모델들의 성능을 평가하여 건설재해 사전 예측을 위한 최적의 모델을 선별하고 분석하였다.

2.1 데이터 수집 및 전처리

본 연구는 건설업에서 가장 빈번히 일어나 집중적인 관리가 필요한 주요 안전사고 유형인(Kim et al., 2017) 6개의 재해유형(떨어짐, 넘어짐, 물체에 맞음, 끼임, 절단 및 베임, 부딪힘)을 대상으로 2019년부터 2021년까지 최근 3년간 발생한 재해사례 데이터 6,826

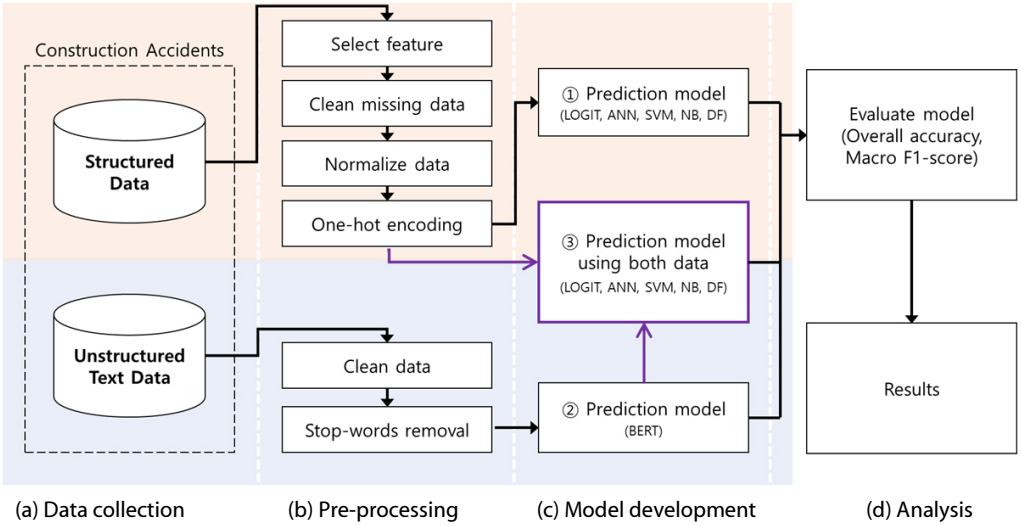

Fig. 1. Research Process to Develop Prediciton Model

건을 수집하여 활용하였다. 데이터는 Python의 BeautifulSoup 패키지(Beautiful Soup, 2020)를 사용한 웹 크롤링(Web crawling)을 통해 수집되었으며, 수집된 데이터는 6개의 재해유형에 대한 46개라는 다양한 변수로 구성되어 있었다. 다만, 기계학습에 사용될 데이터의 변수가 다양하거나 예측 목적과 관련이 적은 변수를 사용할 경우, 예측 모델의 성능이 매우 낮아질 수 있다(Woo et al., 2019). 따라서, 본 연구에서는 데이터 표본이 매우 적거나 사전에 파악 불가능한 변수를 제거하였고, 변수의 중요성을 평가할 수 있는 Permutation feature importance (Fisher et al., 2019) 기법을 적용하여 관련성이 높은 변수들을 선정하였다. 본 연구에서 최종적으로 고려한 입력변수는 14개의 정형데이터와 1개의 비정형 데이터로 구성되어있다. 또한, 출력변수는 떨어짐(Fall, 2035개(29.8 %)), 넘어짐(Wipeout, 1630개(23.8 %)), 물체에 맞음(Hit, 1301개(19.1 %)), 끼임(Narrowness, 891개(13.1 %)), 절단 및 베임(Cut, 481개(8.5 %)), 부딪힘(Crash, 387개(5.7 %))으로 구성된 6가지 건설재해 유형(Type of accident)으로 구성하였으며, Table 1과 같다.

본 연구에서 사용된 정형 및 비정형데이터는 각각 기계학습 모델과 BERT 모델에 적용하기 위한 전처리 작업이 수행되었다. 수치형 정형데이터의 경우, 모델의 학습 속도를 개선하는 효과가 있는 정규화(Normalization) 작업을 수행하였다(Shanker et al., 1996). 또한, 범주형 정형데이터의 경우, 기계학습 모델이 이해할 수 있는 데이터형식으로 변경해주는 원핫 인코딩(One-hot encoding)을 수행하였다. 반면, 비정형 텍스트 데이터인 사고경위(Details of the accident)의 경우엔 특수문자와 같은 문장에 불필요한 요인을 제거하는 데이터 정제(Data cleaning) 작업과 접미사와 같이 무의미하거나 분석에 큰 의미를 갖지 않는 단어를 제거하는 불용어 처리(Stop-words removal) 작업을 수행하였다.

2.2 기계학습 기반 건설재해 예측 모델

기계학습 기반의 예측 모델의 경우, 사용된 데이터와 예측하고자 하는 목적에 따라서 어떠한 기계학습 알고리즘이 최적의 성능을 발휘하는지 사전에 파악하기 힘든 한계가 있다(Ha and Ahn, 2019). 따라서, 본 연구에서는 다중클래스 분류에 우수한 5가지의 기계학습 알고리즘(LOGIT, ANN, SVM, NB, DF)을 적용하고 예측 모델을 개발하고 정량적 성능 평가를 통해 건설재해 예측을 위한 최적의 모델을 도출하였다.

그 중, LOGIT은 데이터가 어떤 범주에 속할 확률을 예측하고 그 확률에 따라 범주를 분류해주는 알고리즘이다(Sperandei, 2014). ANN은 생물의 신경망에서 영감을 얻은 통계학적 학습 알고리즘으로 입력과 출력 사이의 복잡한 학습에 유리하다(Hoskins and Himmelblau, 1992). SVM은 데이터 집합을 바탕으로 예측하고자 하는 데이터가 어느 범주에 속하는지 분류해주는 비확률적 이진클래스 분류(Two-class classification) 알고리즘이다(Cortes and Vapnik, 1995). NB는 간단한 디자인과 가정에도 불구하고, 복잡한 실제 상황에서 높은 성능을 보이는 알고리즘이다(Zhang, 2004). 마지막으로 DF는 분류, 회귀 등을 포함한 다양한 학습 작업에서 높은 성능을 보이는 의사결정 트리(Decision tree)를 반복하여 훈련하고 예측결과를 결합하여 의사결정 트리의 성능을 더욱 고도화한 알고리즘이다(Rokach, 2016).

하지만, 본 연구에서 사용되는 기계학습 알고리즘 중 SVM을 포함한 몇몇 알고리즘은 두 가지 유형(class)만 분류할 수 있는 이진클래스 분류 모델이기에, 6개의 유형으로 구성된 건설재해를 예측할 수 없다. 따라서, 2가지 이상의 유형을 분류할 수 있도록

Table 1. Variables and Feature Considered in This Study

Variable		Type	Feature
Output	Type of accident	Categorical (6 categories)	Fall (2035*), Wipeout (1630), Hit (1301), Narrowness (891), Cut (481), Crash (387)
Input (Structured data)	Temperature	Numerical	-18~37 °C
	Humidity	Numerical	0~100 %
	Season	Categorical (4 categories)	Spring, Summer, Autumn, Winter
	Time	Categorical (2 categories)	0~12 hour, 12~24 hour
	Weather	Categorical (6 categories)	Sunny, Cloudy, Rainfall, etc.
	Type of construction (Main category)	Categorical (7 categories)	Civil engineering, Architecture, etc.
	Type of construction (Sub-category)	Categorical (39 categories)	Temporary work, Excavation, etc.
	Type of facility (Main category)	Categorical (4 categories)	Civil engineering, Architecture, etc.
	Type of facility (Sub-category)	Categorical (20 categories)	Building, Bridge, Road, etc.
	Workplace	Categorical (61 categories)	House, Factory, Road, etc.
	Work object	Categorical (9 categories)	Temporary facilities, Materials, etc.
	Work location (Main category)	Categorical (5 categories)	Inside, Outside, Roof, etc.
	Work location (Sub-category)	Categorical (8 categories)	Floor, Top, bottom, etc.
	Work process	Categorical (58 categories)	Installation, Dismantling, Assembly, etc.
Input (Unstructured data)	Details of the accident	Text	Text-type unstructured data

* Number of data by type of accident

만들어주는 OVA (One-vs-All) 접근 방식을 적용하여 다중클래스 분류를 위한 지도학습을 수행하였다.

또한, 기계학습 기반 예측 모델의 성능은 편향(bias)과 분산(variance) 사이의 균형을 맞출 때 사용되는 하이퍼파라미터(Hyper-parameter)를 적절하게 설정하여 최적화하는 것이 중요한 요소로 작용한다(Raschka, 2018). 따라서, 본 연구에서는 기계학습 기반의 예측 모델의 성능을 최적화하기 위해 k-겹 교차 검증(k-fold cross-validation) 방식을 사용하여 최적의 하이퍼파라미터를 설정하였다.

마지막으로, 5가지 기계학습 알고리즘을 적용하여 개발된 건설재해 예측 모델에 검증용 데이터를 사용하여 Overall accuracy와 Macro F1-score를 척도로 정량적 평가를 진행하였고, 최고의 예측 성능을 보이는 모델을 분석하였다. 이때, 평가하고자 하는 모델의 클래스 수가 l개일 때, Eq. (1)은 정밀도(Macro-precision), Eq. (2)는 재현율(Macro-recall) 그리고 Eq. (3)은 Macro F1-score를 보여주며(Sokolova and Lapalme, 2009), Overall accuracy는 전체 데이터에 대한 예측률을 의미한다.

$$Precision_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l} \tag{1}$$

$$Recall_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l} \tag{2}$$

$$F1-score_M = 2 \times \frac{Precision_M \times Recall_M}{Precision_M + Recall_M} \tag{3}$$

2.3 BERT 기반 건설재해 예측 모델

본 연구에서 비정형 텍스트 데이터를 활용한 건설재해 예측 모델을 개발하기 위해 우수한 텍스트 분류 성능을 보이는 BERT를 사용하였다. BERT는 attention 기법을 활용하는 기계 번역 모델인 Transformer (Vaswani et al., 2017) 기반의 언어모델로 사전학습(Pre-training)과 파인튜닝(Fine-Tuning) 과정을 통해 자연어처리에 우수한 성능을 보이는 모델이며(Devlin et al., 2019), 그 절차는 Fig. 2와 같다. 사전학습은 개발하고자 하는 모델의 성능을 높이기 위해 레이블이 없는 대량의 텍스트 데이터를 사전에 학습하는 과정이며, 학습할 문장의 일부분을 마스킹(Masking)하여 모델이 마스킹한 부분을 예측하도록 하는 Masked Language Model (MLM) 방식과 문장별 관계를 학습하는 Next Sentence Predicion (NSP) 방식이 적용되었다. 파인튜닝은 사전에 학습된 BERT 모델을 다른 예측 모델로 활용하기 위해 새로운 데이터를 학습하고 하이퍼파라미터를 조정하는 과정이다. 따라서, 사전학습된 모델에 파인튜닝을 통해 개발된 BERT 기반의 예측 모델은 적은 양의 텍스트 데이터만으로도 우수한 성능을 보여 최근 많은 연구에서 활용되고 있다(Lee et al., 2020).

본 연구에서는 SK T-Brain에서 개발한 대량의 한국어 텍스트로 사전학습된 한국어 BERT 모델을 사용하였으며, 건설재해 예측 모델 개발을 위해 사고경위 데이터를 추가로 학습시키고 최적의 하이퍼파라미터를 설정하는 파인튜닝 과정을 수행하였다. 특히, 최적화된 모델을 개발하기 위한 파인튜닝 과정은 성능 개선을 위해 Epoch 수를 조정하거나, Batch size, Learning rate, Maximum sequence length 값을 조정하는 하이퍼파라미터 튜닝 과정이 여러 연구에서 대표적인 방법으로 사용된다(Lee et al., 2020). 본 연구에서도 사전학습된 한국어 BERT 모델을 통한 최적의 건설재해 예측 모델을 개발하기 위해 하이퍼파라미터 튜닝 과정을 수행하였다. 하이퍼파라미터의 미세 조정을 통해 약 30 Case의 조합을 실험하고 검증용 데이터를 사용하여 Overall accuracy와 Macro F1-socre를 척도로 정량적 평가를 진행하였고, 건설재해 예측을 위한 최적의 BERT 모델을 개발하였다.

3. 건설재해 예측 모델 평가

본 연구에서는 건설재해를 사전에 예측할 수 있는 최적의 모델을 제안하기 위해 데이터형식(정형 및 비정형)에 따라서 첫 번째, 정형데이터만 활용한 기계학습 기반의 예측 모델, 두 번째, 비정형데이터만 활용한 BERT 기반의 예측 모델, 세 번째, 정형 및 비정형데이터를 모두 활용한 기계학습 기반 예측 모델로 총 3가지 유형의 예측 모델을 개발하고 정량적 평가를 진행하였다. 평가는 Overall Accuracy와 Macro F1-score를 모두 고려하여 진행하였다.

3.1 예측 모델의 성능 분석

Table 2에 나타낸 것과 같이, 정형데이터만을 고려한 기계학습(LOGIT, ANN, SVM, NB, DF) 기반의 건설재해 예측 모델은 Overall accuracy가 69.19~79.98 %, Macro F1-score가 0.6247~0.7401의 분포로 비교적 낮은 예측성능을 보였다. 그 중, DF 알고리즘을 적용한 기계학습 기반의 건설재해 예측 모델이 가장 높은 Overall accuracy (79.98 %)와 Macro F1-score (0.7401)를 보이는 것으로 분석되었다.

Table 3에 나타낸 것과 같이, 비정형데이터만을 고려한 BERT 기반의 건설재해 예측 모델은 5가지의 하이퍼파라미터 설정에 따라, Overall accuracy가 88.45~89.35 %, Macro F1-score가 0.8183~0.8312의 분포로 정형데이터만을 고려했을 때보다 높은 예측성능을 보였다. 그 중, 학습을 위한 Optimizer는 AdamW를 사용하고 Batch size는 32, Learning rate는 0.00005, Maximum sequence length는 64, Epoch는 8로 설정한 BERT 기반의 건설재해 예측 모델이 가장 높은 Overall accuracy (89.35 %)와 Macro

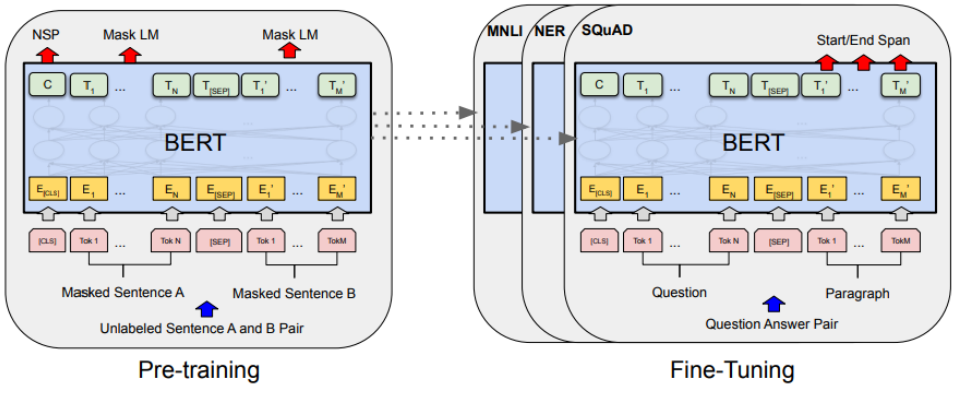

Fig. 2. Overall Pre-Training and Fine-Tuning Procedures for BERT

Table 2. Predictive Performance of Machine Learning Algorithms Using Only Structrured Data

Algorithm	Overall accuracy (%)	Macro F1-score
LOGIT	77.43	0.7022
ANN	69.19	0.6247
SVM	75.21	0.6878
NB	79.64	0.7062
DF*	79.98*	0.7401*

*The best algorithm with the highest predictive performance.

Table 3. Predictive Performance by Hyperparameter of BERT Using Only Unstructured Data

	Optimizer	Batch size	Learning rate	Max_seq_length	Epoch	Overall accuracy (%)	Macro F1-score
Case 1*	AdamW	32	5E-05	64	8	89.35*	0.8312*
Case 2	AdamW	32	5E-05	128	8	89.13	0.8289
Case 3	AdamW	16	5E-05	64	8	89.01	0.8235
Case 4	AdamW	32	3E-05	64	10	88.79	0.8188
Case 5	AdamW	16	3E-05	64	10	88.45	0.8183

*The best hyperparameter of BERT with the highest predictive performance.

Table 4. Predictive Performance of Machine Learning Algorithms Using Both Types of Data

Algorithm	Overall accuracy (%)	Macro F1-score
LOGIT (+BERT)	93.93	0.8709
ANN (+BERT)	91.63	0.8436
SVM (+BERT)	92.34	0.8477
NB (+BERT)	94.32	0.8762
DF (+BERT)*	95.41*	0.8912*

*The best algorithm with the highest predictive performance.

F1-score (0.8312)를 보이는 것으로 분석되었다.

마지막으로, 본 연구에서 최종적으로 제안하는 정형 및 비정형데이터를 동시에 고려한 기계학습 기반의 건설재해 예측 모델은 앞서 개발한 BERT 모델에 비정형 입력변수를 사용하여 건설재해를 예측하고 14개의 정형 입력변수에 추가하여 총 15개의 입력변수를 사용하여 개발되었다.

Table 4에 나타낸 것과 같이, 정형 및 비정형데이터를 모두 고려한 기계학습(LOGIT, ANN, SVM, NB, DF) 기반의 건설재해 예측 모델은 Overall accuracy가 91.63~95.41 %, Macro F1-score가 0.8436~0.8912의 분포로 우수한 예측성능을 보였다. 그 중, DF (+BERT) 알고리즘을 적용한 기계학습 기반의 건설재해 예측 모델이 가장 높은 Overall accuracy (95.41 %)와 Macro F1-score (0.8912)를 보이는 것으로 분석되었다.

3.2 예측 모델의 성능 평가

Table 5에 나타낸 것과 같이, 본 연구에서 데이터형식에 따라서 개발한 3가지 유형의 예측 모델 가운데, 정형 및 비정형데이터를 동시에 고려한 DF (+BERT) 기반의 예측 모델이 가장 높은 예측성능을 보이는 것으로 평가되었다. 특히, 정형데이터만을 활용한 예측 모델의 Overall accuracy (79.98 %)와 Macro F1-score (0.7401)보다 약 20 %의 예측성능이 향상된 Overall accuracy (95.41 %)와 Macro F1-score (0.8912)를 보여, 비정형데이터를 함께 고려함으로써 효과적인 성능 향상을 보이는 것으로 평가되었다.

따라서, 본 연구에서는 정형 및 비정형데이터를 동시에 고려한 DF (+BERT) 기반의 예측 모델을 건설재해 예측을 위한 최적의 모델로 제안하였다. 마지막으로, 본 연구에서 제안한 예측 모델의 출력변수로 사용된 6종류의 재해유형별 예측정확도를 평가하기 위해 혼동 행렬(Confusion Matrix)을 도출하였으며, Fig. 3과 같다.

재해유형의 분포는 떨어짐(29.8 %), 넘어짐(23.8 %), 물체에 맞음(19.1 %), 끼임(13.1 %), 절단 및 베임(8.5 %), 부딪힘(5.7 %) 순이며, 데이터가 부족한 부딪힘에 대한 예측은 비교적 낮은 81.3 %의 예측정확도를 보이는 것으로 평가되었다. 반면, 나머지

Table 5. Predictive Performance Evaluation of the Model Developed in This Study

Algorithm	Overall accuracy (%)	Macro F1-score	Data type
DF	79.98	0.7401	Use only structured data
BERT	89.35	0.8312	Use only unstructured data
DF (+BERT)*	95.41*	0.8912*	Use both structured and unstructured data

*The best algorithm with the highest predictive performance in this study.

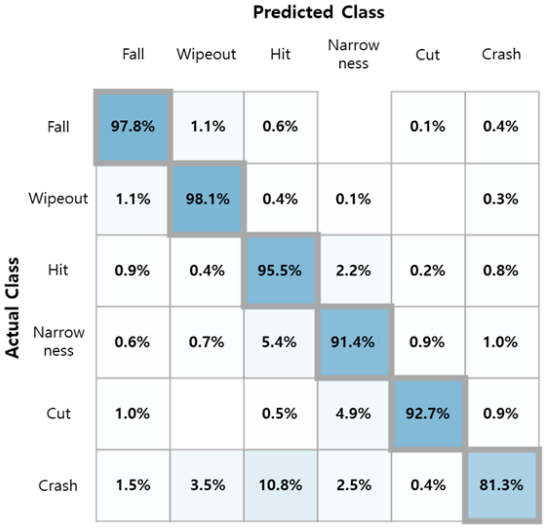

Fig. 3. Confusion Matrix of Proposed Model

재해유형에 대해선 90 % 이상의 높은 예측정확도를 보이며, 데이터 수가 많은 상위 3개의 재해유형(떨어짐, 넘어짐, 물체에 맞음)은 제안모델의 Overall accuracy인 95.41 %보다 높은 예측정확도를 보이는 것으로 평가되었다.

4. 결 론

건설현장에서 발생하는 건설재해는 다양한 정책적 노력과 연구에도 불구하고 지속적으로 증가하는 추세이며 타 산업 대비 대형사고와 사망 사고의 비율이 높다. 이러한 건설재해는 막대한 경제적 손실뿐만 아니라 국가신인도를 낮추고 사회적 불안감을 고조시키기 때문에 선제적 예측이 매우 중요하다. 따라서, 본 연구에서는 주로 건설재해의 원인을 분석해온 기존 연구들과 달리, 건설현장에서 취득할 수 있는 정보를 활용하여 건설재해를 사전에 예측할 수 있는 기계학습 기반의 모델을 개발하였다. 이를 위해 2019년부터 2021년까지의 발생한 건설재해 6,826건의 사전적 정보를 수집하여 활용하였다. 특히, 본 연구에서는 정형데이터뿐 아니라 비정형데이터인 사고경위 정보까지 동시에 고려하여 제안모델의 예측 성능을 제고하였다. 이때 비정형데이터에 대한 학습은 최근 주목받고 있는 BERT를 이용하였으며, 정형데이터의 경우에는 다양한 기계학습 알고리즘 중 가장 우수한 성능을 보이는 Decision forest 알고리즘을 이용하였다. 실제 검증용 데이터를 제안모델에 적용해 본 결과, 정형데이터만을 고려하여 건설재해를 예측했을 때보다 비정형데이터를 함께 고려했을 때 Overall accuracy 및 Macro F1-score 지표가 약 20 %까지 향상됨을 확인할 수 있었다. 최종적으로, 본 연구의 제안모델은 Overall accuracy가 95.41 %, Macro F1-score가 0.8912로 높은 건설재해 예측 성능을 보이는 것을 확인할 수 있었다.

본 연구에서 개발된 기계학습 기반 건설재해 예측 모델은 2가지 측면에서 의의가 있다.

첫째, 본 연구는 비정형 텍스트 데이터를 활용하여 정량적인 성능 향상을 확인하였다. 이는 비정형 텍스트 데이터 활용의 중요성을 시사하며 특히, 건설업의 경우에 타 산업 대비 많은 양의 비정형 텍스트 데이터가 생성되기 때문에 다양한 목적의 데이터 활용성 측면에서 그 의미가 더욱 깊다. 둘째, 본 연구의 제안모델은 건설현장에서 작업 사전에 파악할 수 있는 정보를 활용해 건설재해를 예측하였다. 이를 통해 건설현장 안전관리자는 경험적 근거가 아닌 과학적 근거에 의한 의사결정 정보를 지원받으며, 사전에 건설재해에 대한 대책을 마련하는 근거로 활용할 수 있다. 즉, 본 연구에서 개발된 모델은 Table 1에서 사용된 14개의 정형 인풋 데이터와 1개의 비정형 인풋 데이터(작업내용)를 작업 사전에 입력하여 건설재해를 예측하고 건설 재해율 감소를 위해 활용할 수 있다.

하지만 본 연구에서는 발생확률이 낮은 건설재해의 복합적인 원인은 데이터의 불균형으로 인해 예측 모델에 반영되지 못했을 가능성이 있다. 따라서, 향후 보다 많은 건설재해 데이터를 확보하여 본 연구의 예측 모델을 검증하고 보완하는 연구가 필요하다. 또한, 더욱 효과적인 안전관리를 위해선 건설재해를 예측함과 동시에 해당 건설재해로 인한 위험성 또한 예측하고, 그 결과를 근거로 안전관리 대책을 마련할 수 있는 후속연구가 필요하다.

감사의 글

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

본 논문은 2021 CONVENTION 논문을 수정·보완하여 작성되었습니다.

References

Beautiful Soup (2020). *Beautiful soup documentation*, Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (Accessed: June 25, 2020).

Cho, J. H. (2012). "A study on the causes analysis and preventive measures by disaster types in construction fields." *Journal of the Korea Safety Management & Science*, Vol. 14, No. 1, pp. 7-13.

Cho, Y. R., Kim, Y. C. and Shin, Y. S. (2017). "Prediction model of construction safety accidents using decision tree technique." *Journal of the Korea Institute of Building Construction*, Vol 17, No. 3, pp. 295-303 (in Korean).

Choi, S. J., Kim, J. H. and Jung, K. H. (2021). "Development of prediction models for fatal accidents using proactive information in construction sites." *Journal of the Korean Society of Safety*, Vol. 36, No. 3, pp. 31-39 (in Korean).

Choi, S. Y. (2020). *Comparison analysis of deaths in construction industry in OECD countries*, *Construction & Economy Research Institute of Korea*, pp. 13 (in Korean).

Cortes, C. and Vapnik, V. (1995). "Support-vector networks." *Machine Learning*, Vol. 20, pp. 273-297.

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv:1810.04805v2, pp. 1-16.

Fisher, A., Rudin, C. and Dominici, F. (2019). "All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously." arXiv:1801.01489v5, pp. 1-81.

Ha, M. S. and Ahn, H. C. (2019). "A machine learning-based vocational training dropout prediction model considering structured and unstructured data." *Journal of the Korea Contents Association*, Vol. 19, No. 1, pp. 1-15.

Hoskins, J. C. and Himmelblau, D. M. (1992). "Process control via artificial neural networks and reinforcement learning." *Computers & Chemical Engineering*, Vol. 16, No. 4, pp. 241-251.

Kim, B. S. (2008). "The appropriation and the use scheme of safety control cost for reducing severity rate of injury on construction." *Journal of the Korean Society of Civil Engineers*, KSCE, Vol. 28, No. 3D, pp. 383-390 (in Korean).

Kim, Y. C., Yoo, W. S. and Shin, Y. S. (2017). "Application of artificial neural networks to prediction of construction safety accidents." *Journal of the Korean Society of Hazard Mitigation*, Vol. 17, No. 1, pp. 7-14 (in Korean).

Korea Labor Institute (KLI) (2013). *Construction industry accident status analysis and policy direction*, pp. 31 (in Korean).

Korea Occupational Safety and Health Agency (KOSHA) (2019). *2019 Large accident report book*, pp. 9 (in Korean).

Lee, C. H., Lee, Y. J. and Lee, D. H. (2020). "A study of fine tuning pre-trained korean BERT for question answering performance development." *Journal of Information Technology Services*, Vol. 19, No. 5, pp. 83-91 (in Korean).

Lee, S. G. (2018). "A study on the trends of construction safety accident in unstructured text using topic modeling." *Journal of the Korea Academia-Industrial Cooperation Society*, Vol. 19, No. 10, pp. 176-182 (in Korean).

Lim, W. J., Kee, J. H., Seong, J. H. and Park, J. Y. (2019). "Development of accident cause analysis model for construction site." *Journal of the Korean Society of Safety*, Vol. 34, No. 1, pp. 45-52 (in Korean).

Ministry of Employment and Labor (MOEL) (2020). *2019 Industrial accident analysis of current situation*, pp. 32 (in Korean).

Park, K. C. and Kim, H. K. (2021). "Analysis of seasonal importance of construction hazards using text mining." *KSCE Journal of Civil and Environmental Engineering Research*, KSCE, Vol. 41, No. 3, pp. 305-316 (in Korean).

Raschka, S. (2018). "Model evaluation, model selection, and algorithm selection in machine learning." arXiv:1801.01489v5, pp. 1-45.

Rokach, L. (2016). "Decision forest: Twenty years of research." *Information Fusion*, Vol. 27, pp. 111-125.

Shanker, M., Hu, M. Y. and Hung, M. S. (1996). "Effect of data standardization on neural network training." *The International Journal of Management Science*, Vol. 24, No. 4, pp. 385-397.

Sokolova, M. and Lapalme, G. (2009). "A systematic analysis of performance measures for classification tasks." *Information Processing and Management*, Vol. 45, No. 4, pp. 427-437.

Sperandei, S. (2014). "Understanding logistic regression analysis." *Biochemia Medica*, Vol. 24, No. 1, pp. 12-18.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). "Attention is all you need." arXiv:1706.03762v5, pp. 1-15.

Woo, D. C., Moon, H. S., Kwon, S. B. and Cho, Y. H. (2019). "A deep learning application for automated feature extraction in transaction-based machine learning." *Journal of Information Technology Service*, Vol. 18, No. 2, pp. 143-159.

Yu, Y. J., Kim, T. H., Son, K. Y., Lee, K. H. and Kim, J. M. (2016). "Analysis of primary internal and external risk factors according to the accident causes in construction site." *Journal of the Korea Institute of Building Construction*, Vol. 16, No. 6, pp. 519-527 (in Korean).

Zhang, F., Fleyeh, H., Wang, X. and Lu, M. (2019). "Construction site accident analysis using text mining and natural language processing techniques." *Automation in Construction*, Vol. 99, pp. 238-248.

Zhang, H. (2004). *The optimality of naive bayes*, American Association for Artificial Intelligence, USA, pp. 1-6.