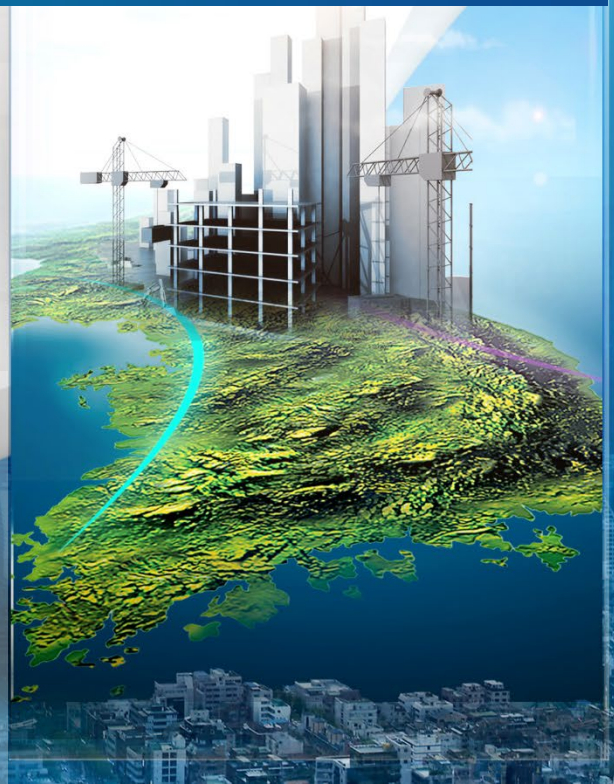
충북대학교
CHUNGBUK NATIONAL UNIVERSITY

빅데이터 분석을 통한 사고사례 분석



2023.07.03

충북대학교 빅데이터협동과정
신승현

Contents

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

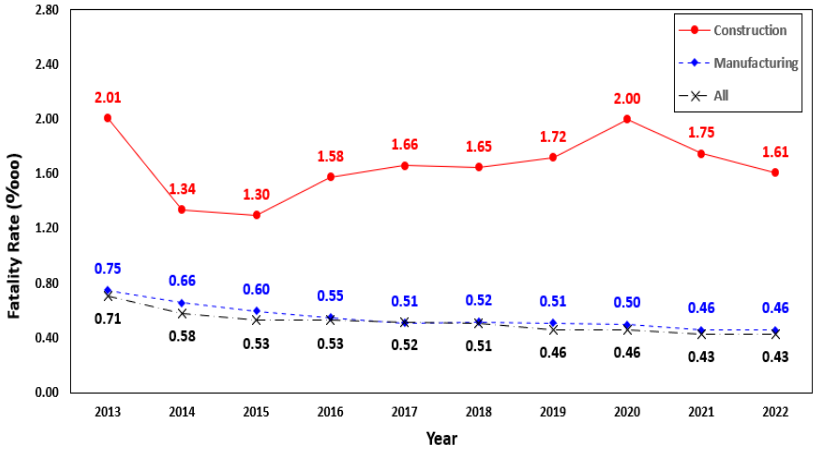
1

연구 배경

01 연구 배경

01 국내·외 건설현장의 산업재해 현황

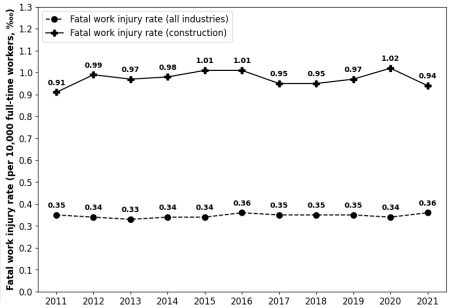
국내 건설업의 높은 사고사망만인율



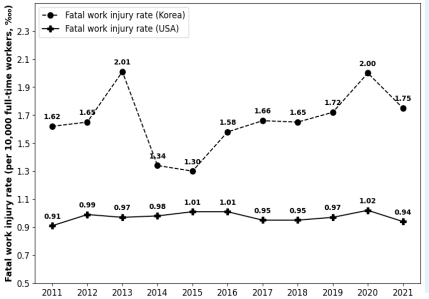
- 건설업 근로자는 위험성이 높은 작업으로 고위험 환경에 노출되어 있으며, 건설업의 산업재해는 타 산업에 비해 높은 실정임
- 제조업·전산업에 비해 건설업의 사고사망만인율은 매우 높음.

※2013년~2022년까지 건설업의 사고사망만인율은 제조업에 비해 평균 3.09배가 높음
※2022년 제조업과 건설업의 사고사망만인율 차이는 약 3.50배

한국의 건설업 사고사망만인율은 미국의 1.7배



〈미국의 전산업·건설업 사고사망만인율〉



〈한국과 미국의 건설업 사고사망만인율〉

- 전세계적으로 건설업은 타산업 대비 사망사고가 다수 발생하고 있으며, 미국 또한 전산업에 비해 건설업 사고사망만인율이 매우 높으므로 건설업의 안전문제는 전세계적인 이슈인 것을 알 수 있음

※2011년~2021년까지 미국 건설업의 사고사망만인율은 전산업에 비해 평균 2.81배가 높음

- 특히 국내 건설업 근로자 1만명당 업무상 사고사망자는 미국에 비해 평균 1.7배가 높으므로 국내 건설업 사망사고는 매우 심각한 수준임을 알 수 있음

※2020년 미국과 한국의 사고사망만인율 차이는 약 2배 이상 발생

국내 건설업에서는 타 업종에 비해 많은 사망자가 발생하고 있으며, 이는 타 국가에 비해서도 매우 높은 수치임
산업재해예방을 위한 건설업 안전보건관리의 근본적인 개선방안 마련이 필요함.

01 연구 배경

02 건설현장 위험 관리와 사고 분석 기술의 한계

건설현장 위험요인과 리스크 감소의 중요성 01

- 건설현장에서 발생하는 위험요인을 정확히 파악하고 이를 통해 리스크를 감소시키는 것은 안전보건관리의 핵심적인 방안 중 하나임
- 위험요인 파악과 리스크 감소를 위해 기존 유사현장에서 발생한 건설사고를 분석하고 현장에서 발생할 사고를 예측하는 것이 중요함
- 건설사고를 제시해주는 시스템이 잘 갖춰져 있으면, 현장 사전안전 관리가 가능하여 미리 안전사고를 예방하고, 위험물 등을 관리할 수 있으며. 궁극적으로 인적 또는 물적 피해를 최소화 할 수 있음

기존의 건설현장 사고 분석 기술의 한계 02

- 지금까지의 건설 현장의 사고예측모델과 관련된 기술은 날씨, 기온, 습도, 근로자수, 작업내용 등 일일 변동 요소들을 반영하지 못하여 복잡성, 불연속성, 비반복성 등 건설업 고유 특성을 고려하지 못함
- 또한 일부 기술들의 작동메커니즘은 사고 및 현장정보 데이터를 모델에 인코딩 하는 과정에서 범주를 고정하거나 데이터를 조작하여 새로운 상황을 반영하는데 한계가 있으며, 데이터의 일부정보가 손실되거나 왜곡되어 신뢰도가 떨어지는 경향이 존재함

01 연구 배경

03 건설현장 사고 예방을 위한 DX와 AI Innovation

DX를 활용한 건설사고 예방

- 디지털 전환(DX)은 기업이 디지털 기술을 활용하여 작업 프로세스를 개선하고 새로운 가치를 창출하는 과정. 즉, 기존의 비효율적인 작업 프로세스를 개선하고, 새로운 가치를 창출하는 데 중요한 역할을 함(Nambisan et al, 2019)
- 건설 현장의 사고 예방을 위한 사고 예측 및 사례 제시는 DX의 중요한 적용 분야가 되고 있음(Salguero-Caparros et al, 2023).
- 과거에는 사고 조사 보고서가 오프라인으로 관리되거나, 산재 관련 정보를 사내 인트라넷망에만 저장하여 데이터의 공유가 제한적이기 때문에 사고 예방을 위한 효과적인 데이터 분석에 한계가 존재함
- 그러나 2019년부터 국토교통부를 중심으로 디지털 전환을 통해 그동안 내부 사고사례 데이터에 대하여 일반인도 접근할 수 있게 인터넷에 공개하면서, 사고 사례 분석이 본격적으로 시작되었으며, 이는 건설현장에서의 사고 예방에 큰 영향을 미침

AI 혁신과 사고사례 분석

- DX를 통해 사고사례가 본격적으로 전산화되고 일반인이 접근 가능하게 되었으나, 사고 예방을 위해 필요한 구체적인 사고 예측은 여전히 어려운 실정이었음
- 연구자들은 사고 보고서의 일부 사례만을 가지고 정성적으로 평가하여 사고에 영향을 미치는 요인들을 도출하는 연구를 진행함(Esmaeili and Hallowell, 2011, Tixer et al, 2016). 이는 사고 예방을 위한 중요한 단계이지만, 제한된 사고사례만을 가지고 분석을 통해 도출된 결과를 가지고 개선방안을 제시한다는 것은 건설현장의 복잡성과 일회성 등을 반영하지 못하므로 사고 예방에 한계가 있음
- 이를 극복하기 위해, 딥러닝을 활용한 자연어 처리로 사고 예측 분석이 시작되었으나 정확도 문제로 현장 적용에 제한이 있었음 (Zhang et al, 2020a; Kim et al, 2022; Qiao et al 2022)
- 최근 트랜스포머 모델이 개발되면서 이 문제가 해결되었고, 대용량 데이터를 효과적으로 활용할 수 있게 되었으며(Vaswani et al, 2017), 이로 인해 실제 현장에서도 사용 가능한 사고 예측 모델이 개발될 수 있게 됨

디지털 전환(DX) 및 인공지능(AI) 혁신에 기반한 빅데이터 분석 방법론을 도입하여
본 연구에서는 고도화된 건설 사고사례 분석모델의 개발함

II

사고사례 분석 모델 개요

02 사고사례 분석 모델 개요

01 제안 모델의 개요

- 건설업 산업재해를 예방하기 위한 기술은 효과적인 사고 예방을 위해 필요한 복잡성, 불연속성, 비반복성 등 건설업의 특성을 충분히 고려하지 못함
- 또한 사고 데이터를 제대로 활용하지 못하는 한계가 있어, 신뢰도가 떨어지는 문제가 있음
- 문제를 해결하기 위해 본 연구에서는 디지털 전환(DX)과 인공지능(AI) 혁신을 결합한 새로운 접근 방식을 통해 사고사례 데이터를 효과적으로 분석하여 사고 예방에 활용하고자 함

- 본 연구에서는 문제점을 개선하기 위하여 건설 사고사례 분석 모델을 제안함
- 본 모델의 목적은, 당일 현장의 주요 현장 조건에 따른 건설사고 정보를 제공하는 것임
- 사고사례모델은 트랜스포머 아키텍쳐와 Large Language Model(LLM)을 기반으로 당일 현장의 주요 현장 조건에 따른 유사현장의 사고 사례와 사고당시 유사현장 정보, 유사도를 제시하는 것임

02 사고사례 분석 모델 개요

02 선행연구와의 연관성 및 차별성

유사현장 사고사례 도출 및 사고 예측 관련 논문

	연구내용(목적, 방법, 결론)	본 연구와의 연관성 및 차별성
Construction site accident analysis using text mining and natural language processing techniques (Zhang et al, 2018)	▪ 해당 연구의 목적은 건설사고분석을 위해 텍스트마이닝과 NLP를 적용하는 것임 ▪ SVM, 선형회귀, KNN, DT, Naive Bayes와 앙상블 모델을 제안하여 사고원인을 분류함. 또한 SQP 알고리즘이 앙상블 모델에 포함된 각 분류기의 가중치를 최적화하는데 사용됨 ▪ 분석결과, 최적화된 앙상블 모델이 다른 모델에 비해 정확도가 높으며, 낮은 사례에 대해서도 비교적 잘 분석 한 것을 밝힘, 또한 사를 일으키는 주요 유해위험요인을 식별하는 데 도움이 되는 것을 밝힘	▪ 본 연구와 해당 연구는 모두 과거 사고 데이터를 분석하고 이를 바탕으로 사고 예측 및 관리방안을 제시한다는 점에서 유사성이 있음 ▪ 또한 기존 사고사례를 바탕으로 사고를 분석하는 것이 효과적인 안전관리 방안임이 본 연구에서 증명됨 ▪ 해당 연구는 주로 과거 데이터를 분석하고 이해하는데 초점을 맞추고 있다면, 본 연구는 과거 데이터를 바탕으로 현재와 미래의 사고를 예방하고 관리하는데 초점을 맞추고 있다는 차이가 있음 ▪ 또한 해당 연구에서 사용한 모델들은 모두 back propagation, forward propagatio의 문제를 해결하지 못하였다는 점과, 건설현장의 고유한 특성을 반영하지 못하였다는 문제가 있음
MetaInjury: Meta-learning framework for reusing the risk knowledge of different construction accidents (Li et al, 2021)	▪ 해당 연구의 주요 목적은 건설 산업에서 작업 관련 부상의 위험을 예측하는 메타 학습 프레임워크인 MetaInjury를 제안하는 것임 ▪ 이 프레임워크는 안전 관리자가 위험 지식을 공유하고 다양한 건설 산업 사고에서 작업 관련 부상의 위험을 예측하는 데 도움이 됨 ▪ 새로운 사고 유형에 대한 소수의 샘플 데이터가 있을 때, 사고 설명의 문서 벡터를 계산하고 Meta-knowledge 데이터베이스 내의 벡터와 비교하여 가장 유사한 사고 유형을 찾은 다음 메타 특징을 데이터 세트에서 가장 좋은 기계 학습 알고리즘과 연결하여 사고 예측 알고리즘의 추천을 구현함	▪ 해당 연구 또한 건설사고예방에 초점을 맞추어, 데이터 기반의 접근방식을 사용하여 예측에 중점을 두었다는 점, 이 때 AI를 사용한다는 점에서 본 연구와 유사성이 존재함. ▪ 그러나, 본 연구는 트랜스포머 모델링을 기반으로 하여 건설 현장 정보를 입력하면 유사 사고 현장의 사례를 제시하고, 이를 바탕으로 현장에 적합한 사고 예측 시나리오를 제시하는 플랫폼을 개발하고 있으며, 이 플랫폼은 제시된 사고 사례 또는 사고 시나리오를 기반으로 건설공사 이해관계자에게 맞춤형 안전보건관리 방안을 제시하는 기능을 가지고 있음 ▪ 반면, 해당 연구는 안전 관리자가 위험 지식을 공유하고 다양한 건설 산업 사고에서 작업 관련 부상의 위험을 예측하고, 다양한 사고 위험 지식을 재사용한다는 점에서 차별성이 존재함

02 사고사례 분석 모델 개요

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

	연구내용(목적, 방법, 결론)	본 연구와의 연관성 및 차별성
Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques (Fan and Li, 2013)	▪ 건설 사고가 발생할 경우, 이로 인해 발생하는 분쟁을 해결하는 데 있어 과거에 발생한 유사한 사고 사례를 찾아내는 것은 중요함. 이 논문의 목적은 이러한 유사한 사례를 효과적으로 찾아내는 방법을 연구하는 것임 이를 위해 텍스트 마이닝 기법을 사용하여 사례를 검색하는 방법을 제안함 ▪ 이 논문에서는 벡터 공간 모델을 사용하여 각 사례를 인덱스 용어의 벡터로 표현함. 이 모델은 문서를 특징짓는 용어들로 구성되며, 이 용어들은 사례의 주요 특징을 반영함. 또한, 각 용어에는 그 중요도를 나타내는 가중치가 부여됨. 이러한 방법을 통해 텍스트 마이닝을 통해 유사한 사례를 검색하는 프로세스를 자동화하고, 이를 통해 분쟁 해결에 필요한 정보를 효과적으로 찾아낼 수 있음 ▪ 이 논문에서 제안하는 텍스트 마이닝 기반의 사례 검색 방법은 구조화된 속성이 각 사례에 대해 미리 정의되어야 하는 전통적인 사례 기반 추론보다 우수하다고 주장함	▪ 두 연구 모두 건설 사고를 예방하는 데 중점을 두고 있음. 이 논문은 과거의 유사한 사고 사례를 찾아내는 방법을 통해 분쟁을 해결하고자 함. 반면 본 연구는 건설 현장 정보를 입력하여 유사한 사고 사례와 예측 시나리오를 제시하는 플랫폼을 개발함 ▪ 또한 두 연구 모두 NLP기반으로 분석하였다는 점에서 유사성이 있음 ▪ 그러나 본 연구와 해당연구는 크게 세가지 측면에서 차이가 존재함 1. 접근 방식: 해당 연구는 과거 사례를 분석하여 분쟁 해결에 도움을 주는 방식을 사용하고 있음. 반면에 본 연구는 현재의 건설 현장 정보를 기반으로 사고 예측 시나리오를 제시하고, 이를 바탕으로 안전 보건 관리 방안을 제시하는 접근 방식을 사용함 2. 기술 사용: 해당연구는 텍스트 마이닝 기법을 사용하여 사례를 분석하고 있으나, 본 연구는 트랜스포머 모델링을 사용하여 사고 예측을 수행함 3. 결과물: 해당연구의 결과물은 텍스트 마이닝 기법을 사용한 사례 검색 방법론임. 반면에 본 연구의 결과물은 건설 현장 정보를 기반으로 사고 예측 시나리오와 안전 보건 관리 방안을 제시함
Retrieving similar cases for construction project risk management using Natural Language Processing techniques (Zou et al, 2017)	▪ 해당 연구의 목적은 건설 프로젝트 위험 관리에서 유사한 사례를 빠르고 정확하게 검색하는 방법을 개선하는 것임. 이를 위해 두 가지 자연어 처리(NLP) 기법, 즉 벡터 공간 모델(VSM)과 의미론적 쿼리 확장을 결합하는 방법을 제안하고 있음 ▪ 제안하는 방법은 두 가지 자연어 처리 기법을 결합하는 것임. 첫째, 벡터 공간 모델(VSM)은 문서와 쿼리 간의 유사성을 측정하는 데 사용되며, 둘째, 의미론적 쿼리 확장은 쿼리의 의미를 확장하여 검색 결과의 정확성을 향상시키는 데 사용됨. 이 두 가지 기법을 결합하여 위험 사례 검색 시스템의 프레임워크를 개발함 ▪ 이 논문에서 제안하는 시스템은 자동으로 유사한 사례를 검색하고 상위 10개의 유사한 사례를 반환할 수 있음을 보여주었습니다. 그러나 이 시스템은 내부 위험 사례 데이터베이스 내에서만 사례 검색이 가능하며, 수집된 위험 사례의 총 수는 상대적으로 작음. 또한, 의미론적 유사성 문제는 여전히 발생하고 있음.	▪ 해당연구와 본 연구 모두 건설현장정보를 입력하면 유사한 사고사례를 제시하는 모델을 개발한다는 점과, AI와 NLP를 사용해 유의미한 정보를 추출한다는 점에서 연관성이 있음 ▪ 그러나 본 연구에서는 트랜스포머 모델링을 활용하고 있습니다. 트랜스포머 모델은 자연어 처리에서 효과적인 결과를 보여주는 최신 기법 중 하나로, 복잡한 문맥을 이해하고 예측하는 데 탁월한 성능을 보임. 반면, 이 논문에서는 벡터 공간 모델과 의미론적 쿼리 확장 기법을 사용하므로 분석결과의 한계가 존재함. ▪ 또한 본 연구에서는 사고 예측 시나리오를 제시하는 기능을 개발하고 있으며, 이는 건설 현장에서 발생할 수 있는 다양한 사고 시나리오를 예측하고 이를 통해 사고를 미리 예방하는 데 도움이 됨 ▪ 또한 본 연구에서는 제시된 사고 사례나 사고 시나리오를 기반으로 건설공사 이해관계자에게 맞춤형 안전 보건 관리 방안을 제시하는 기능을 개발하였다는 점에서 차별성이 존재함

02 사고사례 분석 모델 개요

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

	연구내용(목적, 방법, 결론)	본 연구와의 연관성 및 차별성
Utilization of similar accident cases for safety education (Luo and Hirogane, 2022)	▪ 해당 연구의 목적은 건설 산업에서의 직업 사고의 원인과 대책을 이해하는 것임. 이를 위해, BERT 모델을 사용하여 건강, 노동, 복지부의 직업 사고 사례 데이터베이스에서 유사한 직업 사고 사례를 추출하였으며,이를 바탕으로, 체크리스트 형태로 안전 교육에 적용할 방법을 제안함 ▪ 해당 연구에서는 BERT 딥러닝 모델을 사용하여 건강, 노동, 복지부가 제공하는 직업 사고 사례의 내용을 학습하고, 사고의 유사성을 분석함. 추출된 사고 사례를 기반으로 PFA 방법을 사용하여 분석하고, 산업재해를 예방하기 위한 체크리스트를 생성함	▪ 해당연구와 본 연구 모두 정부의 자료를 바탕으로 BERT를 사용해 유사 사고 사례를 추출하였다는 점에서 매우 유사함이 있음 ▪ 그러나 해당 모델은 사고사례를 입력하면 유사한 사고사례를 제시하는 반면, 본 모델에서는 현장정보를 입력하면 해당 현장과 유사한 현장에서 발생한 사고사례를 제시한다는 점에서 차이가 있음 ▪ 해당 연구의 모델을 사용하려면 사고사례를 본인이 인지하고 있고 이를 입력해야 적합한 사고사례와 체크리스트를 도출할 수 있다는 측면에서 해당 모델은 전문가 외 사용하기가 매우 어려울 수 있다는 단점이 존재하며, 선행 사고사례가 반드시 있어야만 특정 사고사례에 대한 안전체크리스트가 도출된다는 문제점이 존재함 ▪ 반면 본 연구에서는 이러한 문제점이 발생되지 않음

유사사고사례 도출 및 사고시나리오 예측과 관련된 선행연구 분석 결과,

과거사고를 분석하여 유해위험요인과 사고시나리오를 도출하는 것은 핵심 안전관리방안 중 하나임이 증명됨

그러나 여전히 건설현장의 복잡성과 일시성 등 고유한 특성을 고려한 모델은 아직 개발되지 않았으며, 또한 back propagation, forward propagation 문제를 해결하지 못하여 정확도가 떨어져 실제 현장에서는 사용하기 힘든 측면이 있었음

최근 트랜스포머 모델링 기반으로 AI 혁신이 이루어져 사고의 대용량의 데이터를 데이터 손실 없이 사용됨에 따라

현장에서 사용할 수 있는 유사현장의 사고사례 도출 및 사고시나리오 예측 모델 개발이 필요함

02 사고사례 분석 모델 개요

국내 유사 특허 현황

	기술핵심내용	발명기술과의 연관성 및 유사성	발명기술과의 차별성
건설 안전관리 시스템 및 방법 (2018, 등록)	▪ 대상 건설 프로젝트와 유사한 기존 프로젝트를 분석하여 예상되는 안전사고 데이터를 추출하고, 이를 바탕으로 외부 전문가에게 질의 ▪ 받은 답변을 통해 예상되는 안전사고의 발생확률 또는 심각도를 도출하여 안전관리자나 작업자에게 안전관리 항목 및 사고예방 지침에 대한 정보를 제공	▪ 건설 프로젝트와 유사한 프로젝트를 분석하여 예상되는 안전사고를 도출한다는 점에서 본 발명기술과 일부 유사성이 존재함 ▪ 분석 데이터(공종, 공사면적, 비용 등) 는 본 발명기술에서도 활용되었다는 점에서 일부 연관성이 있음	▪ 해당 기술에서 사용된 데이터는 프로젝트 계획 및 설계 단계에서 도출되는 데이터를 활용함에 따라, 본 발명기술과 달리 실제 시공 중 발생되는 데이터(기상정보, 작업자수 등)을 반영하지 못하여 실제 당일 현장에서 발생할 사고예측부분에서 한계가 존재함 ▪ 해당 기술의 현장 유사도 계산 방식은 같은 범주 내에서 정량적인 지표간의 유사성을 계산하는 방식으로 본 발명기술에서 제안한 자연어처리기반의 유사도 도출에 비해 정확도가 떨어짐(데이터의 왜곡 및 데이터 수의 한계)
건설현장 작업환경 개선시스템 및 그 제공방법 (2020, 등록)	▪ 건설현장의 공간정보를 모델링하고, 과거 사고사례를 저장하여 이를 학습하는 시스템 ▪ 이 시스템은 건설현장에서 촬영장치로부터 얻은 정보를 분석하여 참여객체의 행위와 이동경로를 수집하고, 이를 바탕으로 현재 건설현장의 사고 위험도를 판단	▪ 두 기술 모두 건설현장의 안전성을 향상시키는데 초점을 두고 있음 ▪ 과거의 사고사례 데이터를 학습하여 현재의 사고 위험도를 판단하는 방식을 사용하고 있음. 이는 과거의 사고 사례를 분석하고 이를 바탕으로 현재의 위험 요소를 예측하는 방식을 채택함	▪ 해당 기술은 주로 건설현장의 공간정보를 바탕으로 촬영 데이터를 분석하여 건설장비로 인한 사고 예방을 목표로 함. 발명기술은 장비로 인한 사고 뿐만 아닌 모든 사고 예방을 목표로 개발됨 ▪ 해당 기술의 사용을 위해서는 건설현장의 촬영장치를 설치하여야 하며, 상기 과거사례가 존재하여야 위험도 예측이 가능함 ▪ 본 발명기술은 별도의 추가 장비의 설치가 불필요 하며, 과거 사례가 존재하지 아니하여도 LLM 기반의 사고예측시나리오를 제시할 수 있다는 장점이 있음

02 사고사례 분석 모델 개요

	기술핵심내용	발명기술과의 연관성 및 유사성	발명기술과의 차별성
현장조건과 사고사례를 이용한 작업자의 위험도 예측 시스템 및 그 방법 (2021, 등록)	▪ 작업 환경 데이터, 개인조건 데이터, 사고 유형 데이터를 수집하고 정량화하여 작업자의 위험도를 예측하는 시스템 ▪ 이 시스템은 가장 높은 유사도를 가지는 사고 사례를 선택하고, 해당 사고 유형에 대한 가중치를 이용하여 현장조건 위험도와 개인조건 위험도를 연산	▪ 두 기술 모두 과거의 사고 데이터를 분석하고, 이를 바탕으로 현재의 위험 요소를 예측하는 방식을 사용하고 있음	▪ 해당 기술은 작업 환경의 데이터와 저장된 작업 환경 데이터를 이용하여 유사도를 연산하는 방식을 사용하고 있음. 해당 기술은 데이터를 정수형 형태의 범주로 변환하는 등 원시 데이터를 자체적으로 변환 한 뒤 분석함에 따라 데이터의 왜곡 및 데이터 수의 한계가 발생하며, 새로운 건설 상황을 반영하지 못함 ▪ 반면, 발명기술은 사고 예측을 위해 자연어 처리 기반의 유사도 도출 방식을 사용하여 결과에 대한 신뢰도가 훨씬 높음 ▪ 해당기술은 위험도를 예측하는 것에 초점을 두는 반면, 발명기술은 건설 사고 정보를 산출하는 것에 초점을 두고 있음
인공지능 기반의 위험 정보 자동화 추천 장치 및 방법 (2021, 등록)	▪ 유해 요인 관련 데이터를 수집하고, 사전 학습된 딥러닝 기반의 자연어 처리 모델을 이용하여 해당 데이터의 사고 원인 유형을 식별하는 시스템 ▪ 이 시스템은 식별된 사고 원인 유형과 관련된 사고 사례 정보를 출력하며, 이를 통해 위험 요인에 대한 정보를 자동화하여 추천하는 기능을 제공	▪ 두 기술 모두 사고 예측 및 분석을 위해 자연어 처리와 딥러닝 기반의 인공지능 모델을 활용하고 있음 ▪ 이를 통해 사고 원인 유형을 식별하고, 관련된 사고 사례 정보를 출력함	▪ 해당 기술은 위험 정보를 자동으로 추천하는 시스템에 초점을 맞추고 있음. 이는 사고 원인 유형을 식별하고, 이에 대한 정보를 자동으로 추천하는 것이 주요 목표로 함 ▪ 반면, 발명 기술은 건설사고 정보를 산출하는 시스템에 초점을 맞추고 있음. 이는 건설 현장에서 발생할 수 있는 다양한 사고 유형을 예측하고 분석하는 데 중점을 둠 ▪ 즉, 발명 기술은 사고 원인 유형을 식별하고 사고 사례 정보를 출력하는데 초점을 두므로 발명기술은 생성형 AI(LLM)를 활용해 더욱 다양한 시나리오를 고려하여 정확한 사고예측을 제공할 수 있음

02 사고사례 분석 모델 개요

	기술핵심내용	발명기술과의 연관성 및 유사성	발명기술과의 차별성
빅데이터를 활용한 데이터마이닝기반 건설사고 객체정보 추출 방법 (2022, 거절)	▪ 건설사고 데이터를 수집하고, 이 데이터에 대해 자연어 처리 기술과 문서 처리 기술을 적용하여 정보를 추출 및 가공하는 텍스트마이닝 과정을 포함 ▪ 이 과정을 거친 데이터에서 핵심 키워드를 추출하고, 이 핵심 키워드에 대해 객체유형 통합 및 통계적 분석을 통해 위험 시나리오를 구성하는 방법을 제시	▪ 건설사고 데이터로부터 텍스트 마이닝을 통해 사고 위험 시나리오를 구성한다는 점에서 일부 연관성과 유사성이 있음	▪ 해당 기술은 구체적으로 DB를 구축하는 과정과 데이터를 활용해 텍스트마이닝을 수행하는 과정에 대한 구체적인 내용이 전혀 제시되지 않음 ▪ 즉, 해당 기술의 결정적인 특허 거절 이유는 매우 포괄적이며, 기술에 대한 구체적인 내용이 전혀 제시되지 않아 거절된 것으로 판단됨 ▪ 발명기술에서는 DB에 대한 내용을 포함하여 DB로부터 데이터를 전처리, 임베딩, 딥러닝하여 결과를 도출하기까지의 자세한 내용이 제시되었으며, 등록된 다른 유사특허의 구성을 모두 충족함

유사사고사례 도출 및 사고시나리오 예측과 관련된 선행연구 분석 결과,

과거사고를 분석하여 유해위험요인과 사고시나리오를 도출하는 것은 핵심 안전관리방안 중 하나임이 증명됨

그러나 여전히 건설현장의 복잡성과 일시성 등 고유한 특성을 고려한 모델은 아직 개발되지 않았으며, 또한 back propagation, forward propagation 문제를 해결하지 못하여 정확도가 떨어져 실제 현장에서는 사용하기 힘든 측면이 있었음

최근 트랜스포머 모델링 기반으로 AI 혁신이 이루어져 사고의 대용량의 데이터를 데이터 손실 없이 사용됨에 따라

현장에서 사용할 수 있는 유사현장의 사고사례 도출 및 사고시나리오 예측 모델 개발이 필요함

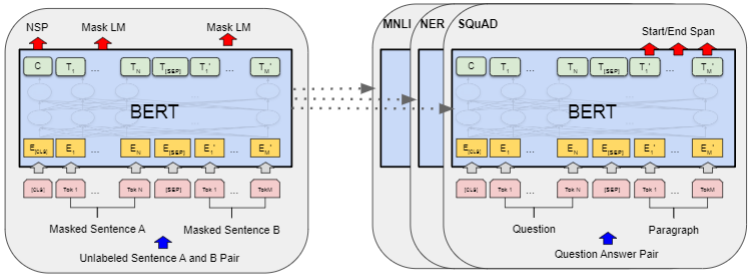
02 사고사례 분석 모델 개요

03 본 모델에 적용된 딥러닝 모델

- 본 사고사례 분석 모델은 입력되는 다수의 현장 정보를 기억하고(back propagation), 입력되는 단어의 순차에 따라 정보가 손실되는 문제 (forward propagation)를 개선하기 위하여 Self-attention 기반의 트랜스포머 아키텍처를 적용하여 개발함
- 본 모델은 BERT(Bidirectional Transformers for Language Understanding)를 기반으로 개발됨

BERT(Bidirectional Transformers for Language Understanding)

- BERT는 기본적으로 대용량의 unlabeled data로 모델을 미리 학습 시킨 후 특정 task를 가지고 있는 labeled data로 transfer learning을 하는 모델임
- 이때 특정 task를 처리하기 위해 새로운 network를 붙일 필요 없이, BERT모델 자체의 fine-tuning을 통해 해당 task를 수행함
- 앞서 언급한데로 BERT는 Transformer의 Encoder 부분을 이용한 Architecture를 가지고 있으며, 마스크 언어 모델은 아래 그림과 같이 토큰들을 랜덤으로 [MASK]토큰으로 대체한 뒤 원래 토큰을 예측하는 문제임
- 이를 통해 마스킹된 양방향 토큰들의 정보를 동시에 이용함과 동시에 컨닝 문제를 피할 수 있었고 그 결과 downstream task에서 그 이전 SoTA모델인 GPT1과도 꽤 큰 차이로 압도하는 성능을 보임



System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

※ Devlin et al(2018) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

02 사고사례 분석 모델 개요

- BERT 사용 이유

1) 양방향 학습 방식

✓ BERT는 양방향 학습 방식을 사용하여 입력 텍스트의 모든 정보를 인코딩합니다. 이를 통해 주어진 텍스트의 중요한 특징과 시맨틱 관계를 더 정확하게 포착할 수 있음.
✓ 따라서 BERT는 텍스트 사이의 유사성을 더 정확하게 판별하는 데 GPT 등 다른 모델에 비해 더욱 유리함.

2) 고차원 인코딩

✓ BERT는 텍스트에 대한 더은 차원의 인코딩을 생성하여 세부 정보를 포착할 수 있음.
✓ 따라서, 사용자가 입력하는 건설 현장 정보와 기 저장된 사고 사례 데이터 사이의 상세한 유사성을 더 잘 이해할 있게 하므로 입력과 저장된 데이터 사이에서 더 적절한 사고 내용을 찾을 수 있음

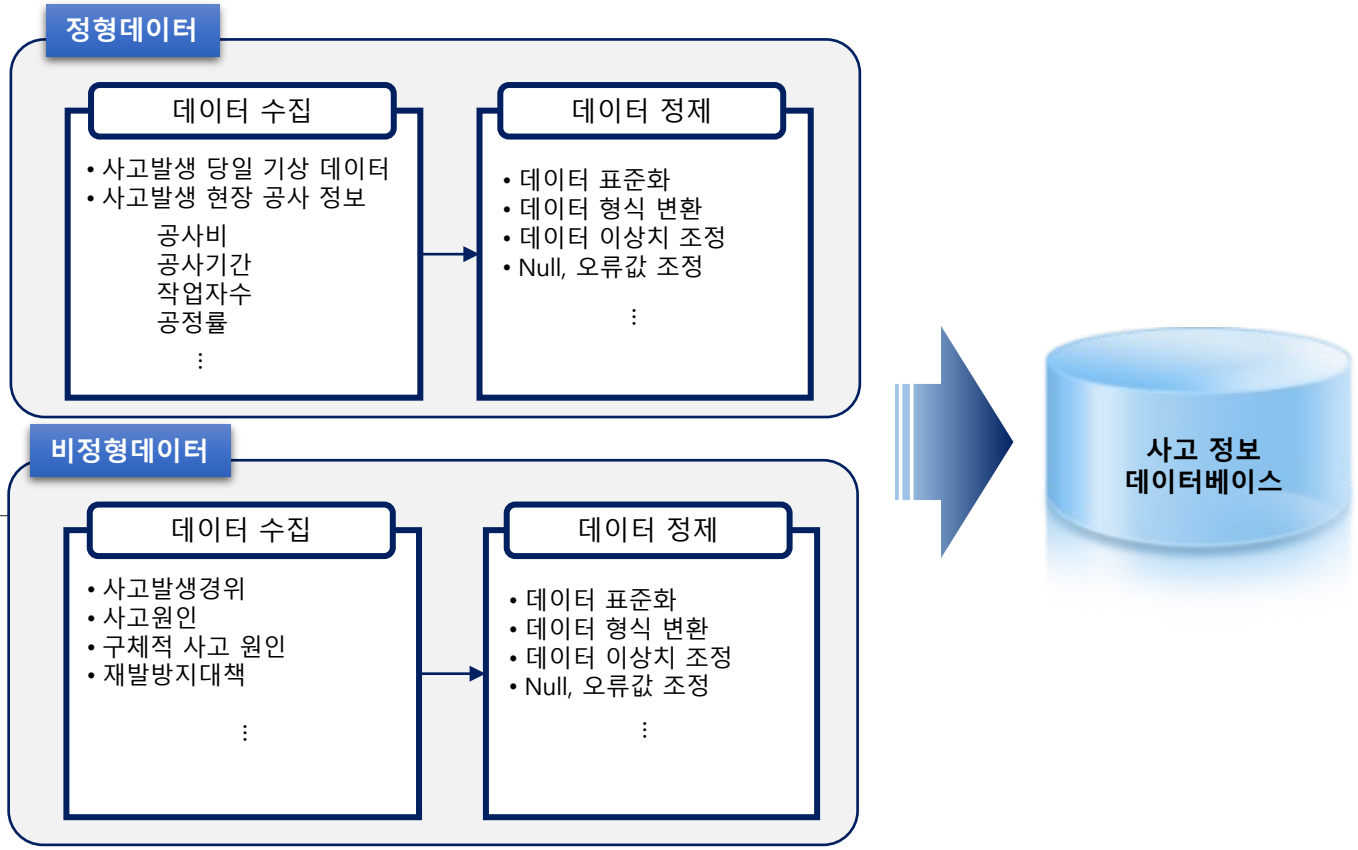
3) 전이학습 향상

✓ BERT는 마스크된 언어 모델링을 사용하여 전이학습이 더 효과적으로 이루어짐.
✓ 이를 통해 주어진 건설 현장 정보에 대한 이해도가 더욱 향상되고, 기 저장된 사고 사례 데이터와의 관련성을 더 높일 수 있음

02 사고사례 분석 모델 개요

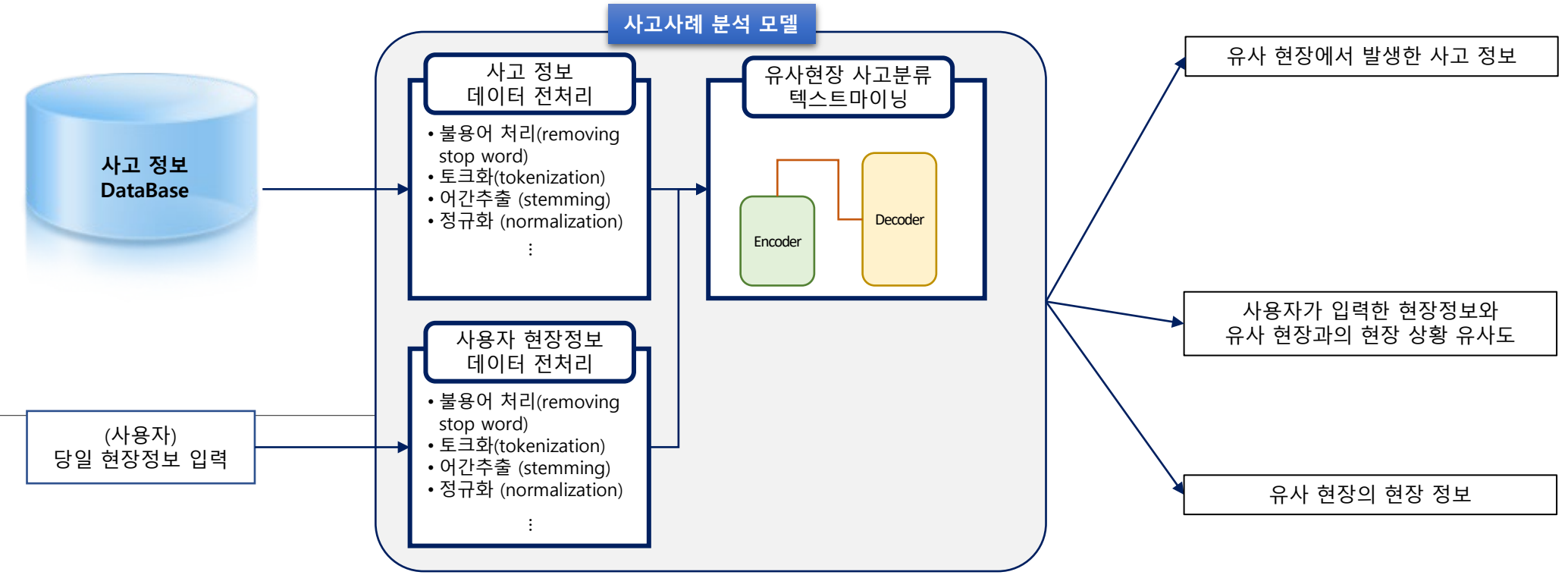
04 제안 모델의 구성 및 기능

- 사고 정보 데이터베이스는 건설 사고 사례를 포함하는 데이터를 저장함
- 데이터는 정형 데이터와 비정형 데이터로 구성됨
- 사고 정보 데이터 베이스는 국토안전관리원 및 안전보건공단 등 외부 서버로부터 소정 횟수로 자동 업데이트 될 수 있음



02 사고사례 분석 모델 개요

- 현재 모델에서는 '유사 현장에서 발생한 사고 정보', '사용자가 입력한 현장정보와 유사 현장과의 현장 상황 유사도', '유사 현장의 현장 정보'를 제공하고 있음



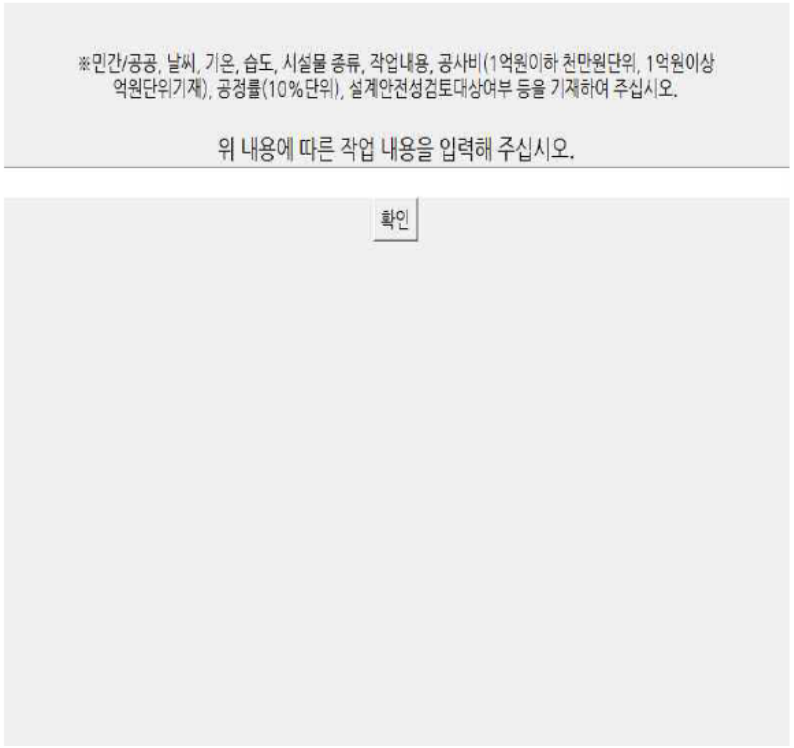
III

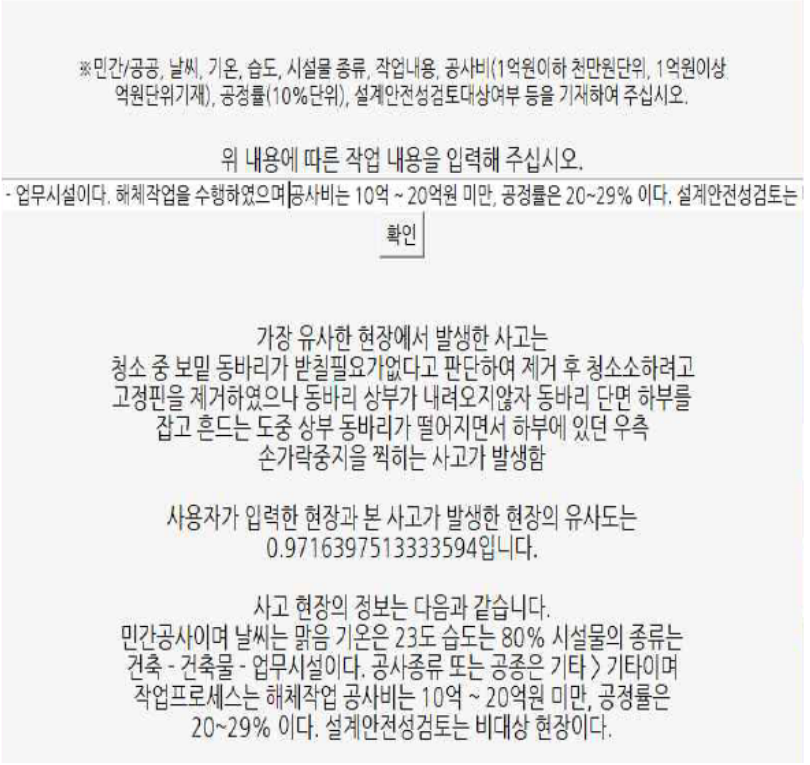
추진경과

03 추진 경과

01 연구 진행 사항

- 시연 모델은 현장정보를 입력하면 국토안전관리원에 등록된 사고사례를 기반으로 하여, 유사현장과 사고사례, 유사도를 도출하도록 함
- 시연 모델은 유사도가 가장 높은 사례 하나만을 도출되도록 코딩하였으나, 현재 개발중인 모델은 다수의 사례를 나타내도록 수정중임
- 시연 모델은 exe 파일로 실행할 수 있도록 개발되었으며, 현재 프리 알파 테스트(Pre-Alpha Test)를 진행중



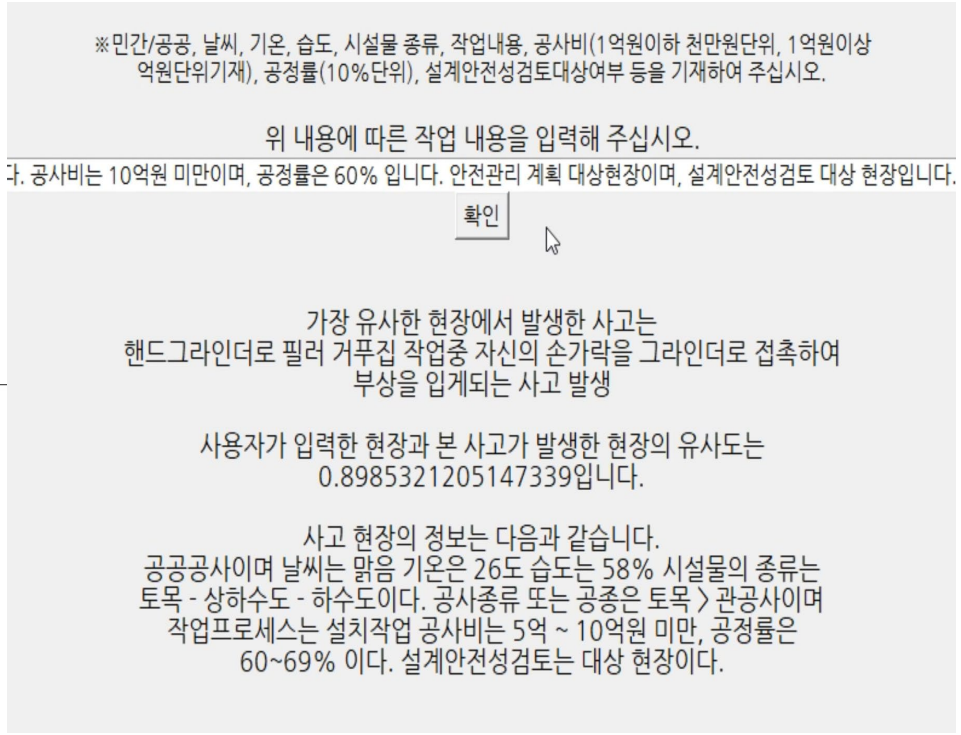


03 추진 경과

02 실제 시연 결과

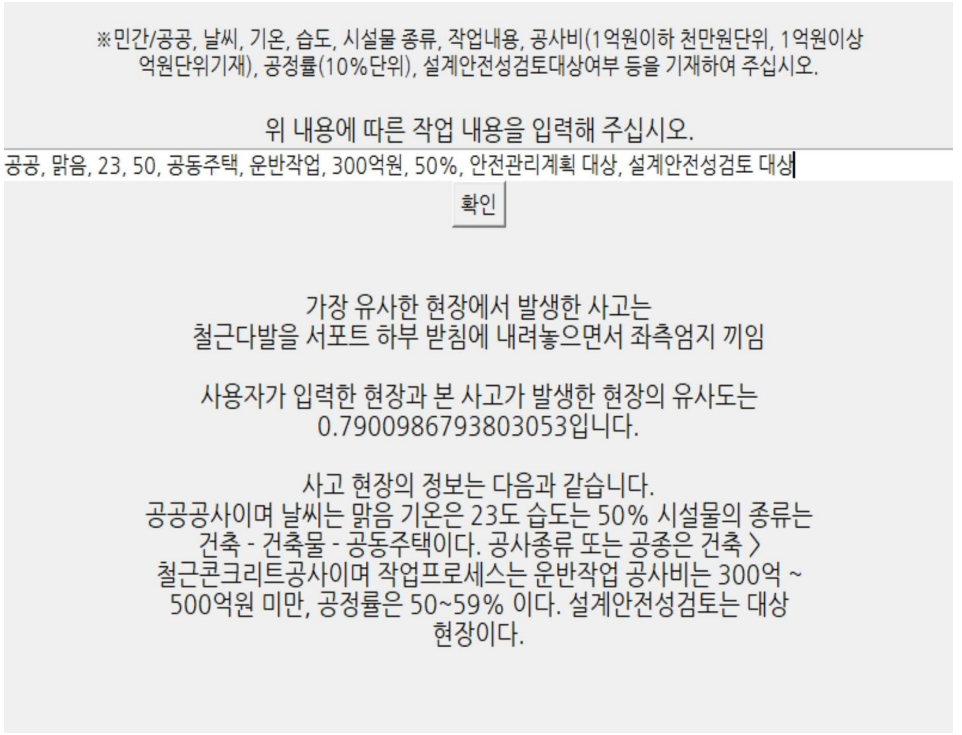
본 모델에서 권장하는 순서에 맞추어 문장으로 입력한 경우

- 입력 문장 : “저희는 공공공사현장입니다. 금일 날씨는 맑으며, 기온은 26도, 습도는 58% 입니다. 시설물의 종류는 하수도이며, 금일 작업 내용은 설치작업 입니다. 공사비는 10억원 미만이며, 공정률은 60% 입니다. 안전관리 계획서 대상현장이며, 설계안전성검토 대상 현장입니다.”



본 모델에서 권장하는 순서에 맞추었으나 문장으로 입력하지 않은 경우

- 입력 문장 : “공공, 맑음, 23, 50, 공동주택, 운반작업, 300억원, 50%, 안전관리계획 대상, 설계안전성검토 대상”

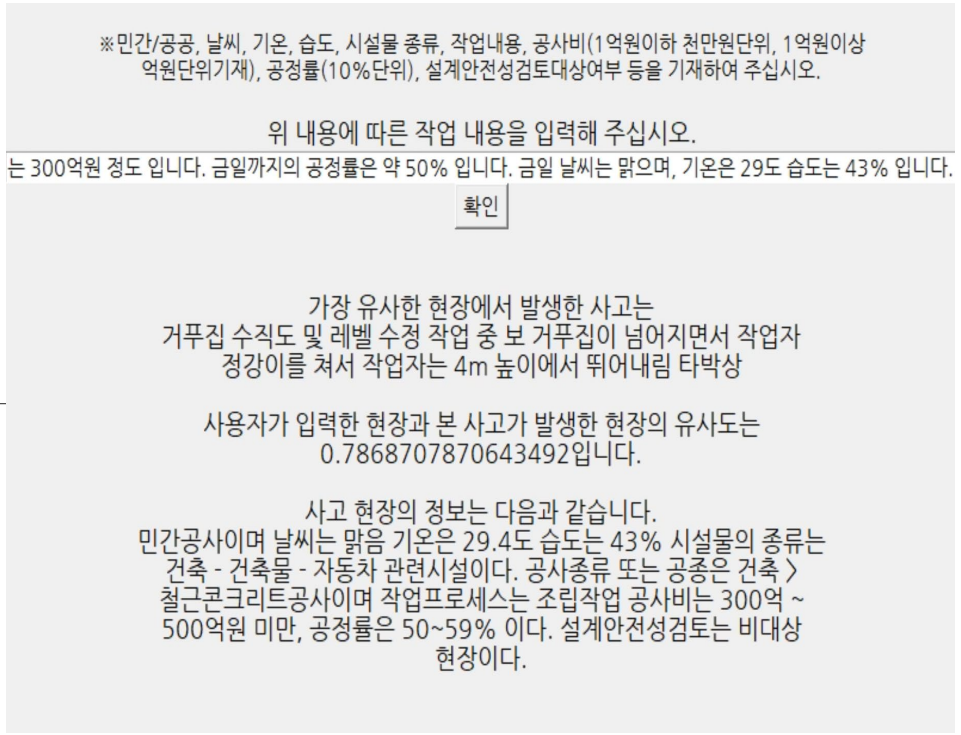


03 추진 경과

02 실제 시연 결과

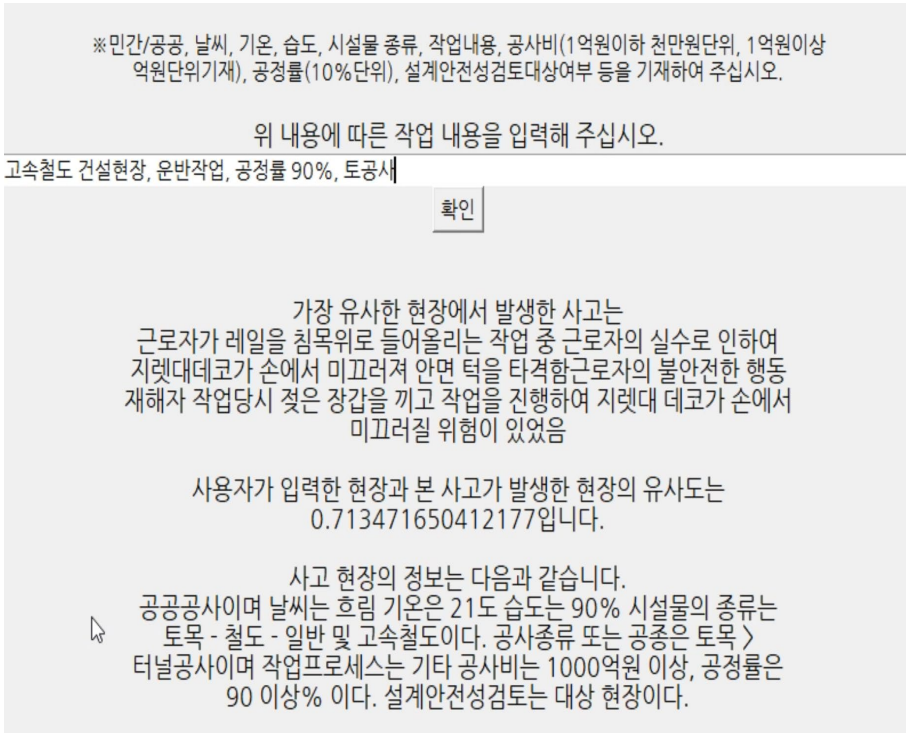
본 모델에서 권장하는 순서에 맞추지 않고 문장으로 입력한 경우

- 입력 문장 : “저희 현장은 민간현장이며, 자동차 관련 시설을 건설하고 있습니다. 공종은 철근콘크리트 공사, 공사비는 300억원 정도 입니다. 금일까지의 공정률은 약 50% 입니다. 금일 날씨는 맑으며, 기온은 29도 습도는 43% 입니다.”



현장정보를 적게 입력한 경우

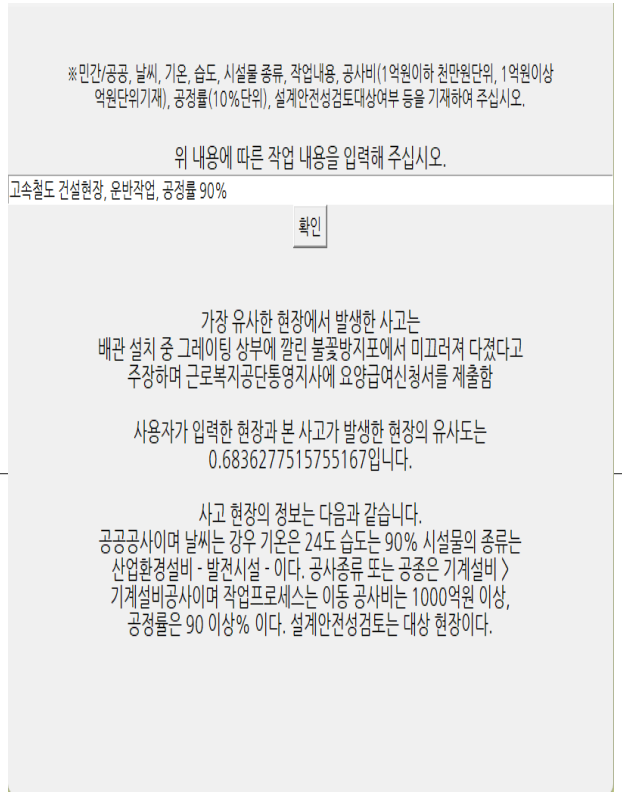
- 입력 문장 : “고속철도 건설현장, 운반작업, 공정률 90%, 토공사”



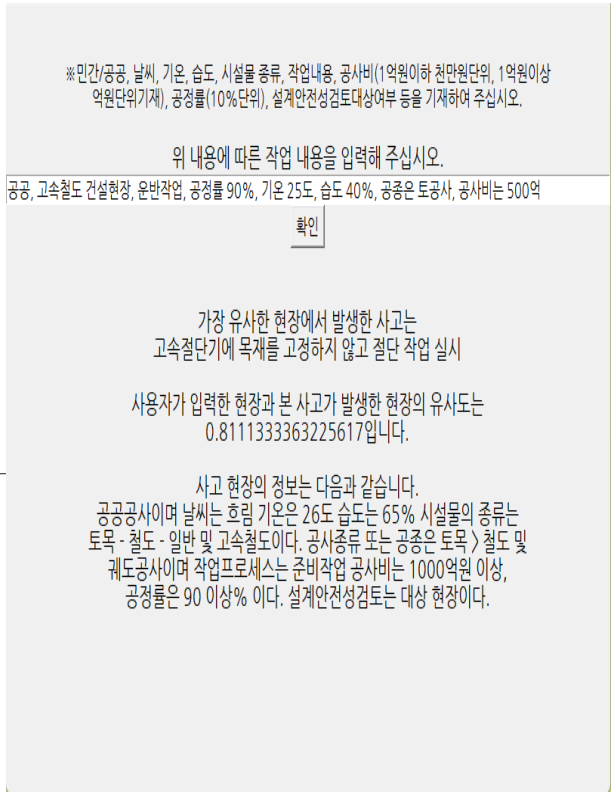
03 추진경과

한계

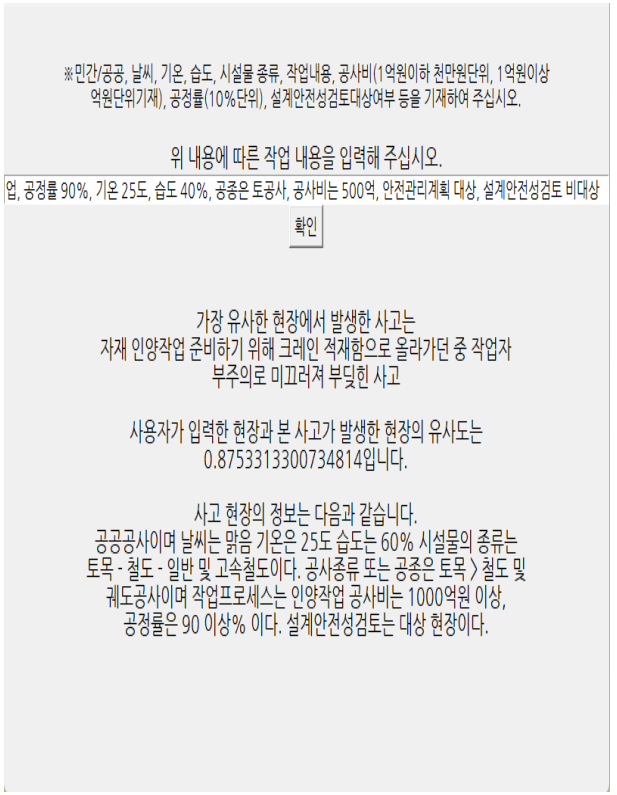
- 현장정보를 적게 입력할 경우, 현장과 전혀 맞지 않는 사고사례가 도출되는 경우가 다수 발생
- 10개 이상의 정보를 입력하여야 실질적인 유사현장 도출 가능
- 문제를 개선하기 위하여 현재 파인튜닝을 지속적으로 진행중



〈현장 정보 3개〉



〈현장 정보 8개〉



〈현장 정보 10개〉

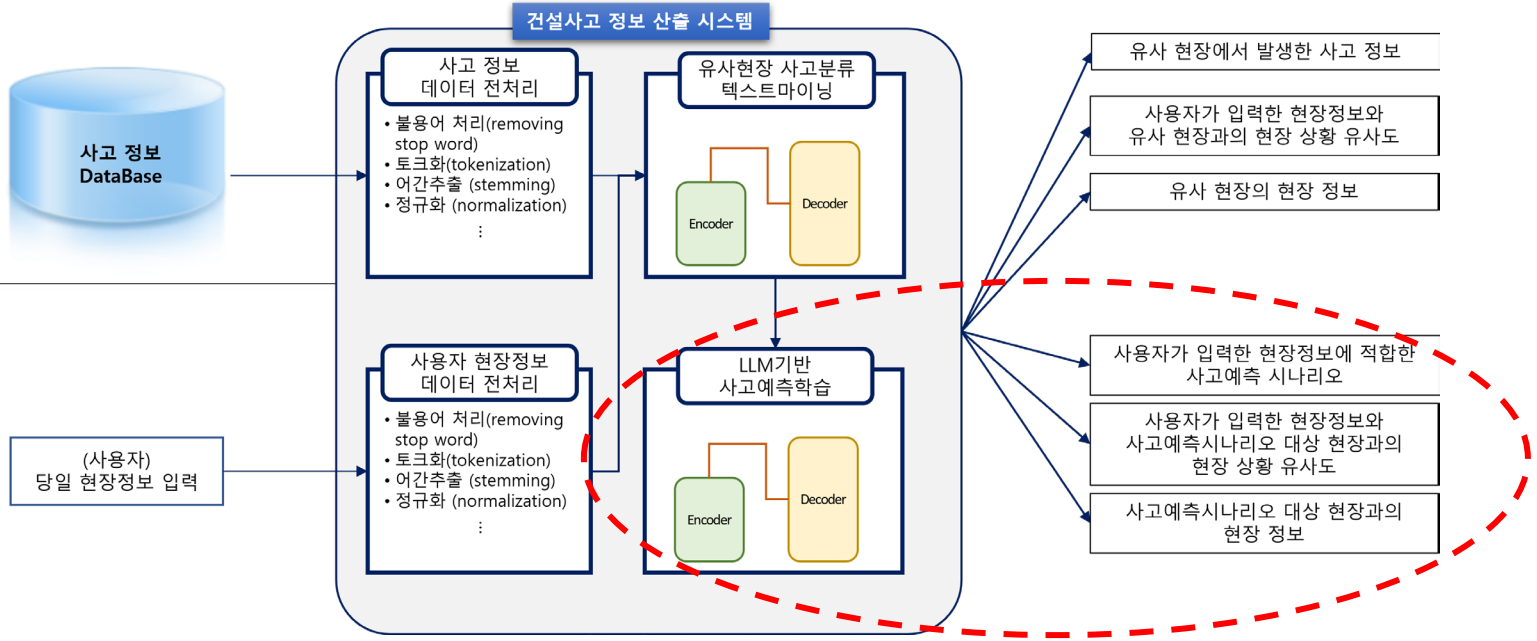
IV

향후 계획

04 향후 계획

01 향후 개발 및 상용화 계획

- 현재 LLM 기반의 사고예측학습을 수행하고 있으며, 구체적으로 Meta의 LLaMA 기반의 파인튜닝 통해 사고예측시나리오와 관련 정보를 제공하도록 개발중임
- 현재 프리 알파 테스트(Pre-Alpha Test)를 진행중이며, 사고예측기능을 추가하기 전 연구실 내 직장인을 대상으로 알파테스트를 진행할 예정임
- 이후 데이터 사용에 대한 공식적인 승인을 받은 뒤 안전보건공단 또는 국토안전관리원 등 공공기관과 건설안전커뮤니티를 통해 베타테스트를 수행한 뒤 최종 개발 후 상용화를 실시할 예정임



Thank You.
감사합니다 :)

일터안전에서 국민안심으로

산업재해 발생개요 분류 모델 시범 개발 현황 및 분석 계획

안전보건공단 빅데이터부 유현철

산업재해예방
안전보건공단

일터안전에서 국민안심으로

CONTENTS

산업재해예방 안전보건공단

일터안전에서 국민안심으로

1

추진 목적 및 개요

산업재해예방
안전보건공단

1 추진 목적 및 개요

추진 목적

▶▶ 업무상 재해 승인자료, 사업주의 산업재해조사표 자료, 재해조사 자료 등 다양한 경로를 통해 수집되는 **재해발생개요(이하 '재해개요')**에서

▶▶ **'발생형태', '기인물', '작업내용'** 등을 자동으로 생산/분류할 수 있는 인공지능 모델 개발로

① **산재예방정책 수립에 필요한 기초자료 생산**

② **분석자료 생산의 효율성 향상**

③ **생산 자료의 교차 검증을 통한 신뢰성 제고**

1 추진 목적 및 개요

재해개요 수집 경로

▶▶ 업무상 재해에 대한 근로자 등의 **산재보상 승인자료**,
사업주의 **산업재해조사표** 등 **2종**이 **주요 수집 경로**

구분	산재보상 승인자료 승인통계	산업재해조사표
기준	4일 이상 요양이 필요한 재해	3일 이상 휴업이 필요한 재해
서식	요양급여 및 휴업급여 신청서	산업재해 조사표
신청/보고	근로복지공단 지사	관할 노동지청
재해개요	행위자, 장소, 작업내용, 목적, 경위, 동작, 원인 등을 기재	발생일시, 발생장소, 재해관련 작업유형, 재해발생 당시 상황, 재해발생 원인 구분 기재

공단에서 **발생형태, 기인물** 분류

노동지청에서 **발생형태, 기인물, 작업지역공정, 작업내용** 분류

1 추진 목적 및 개요

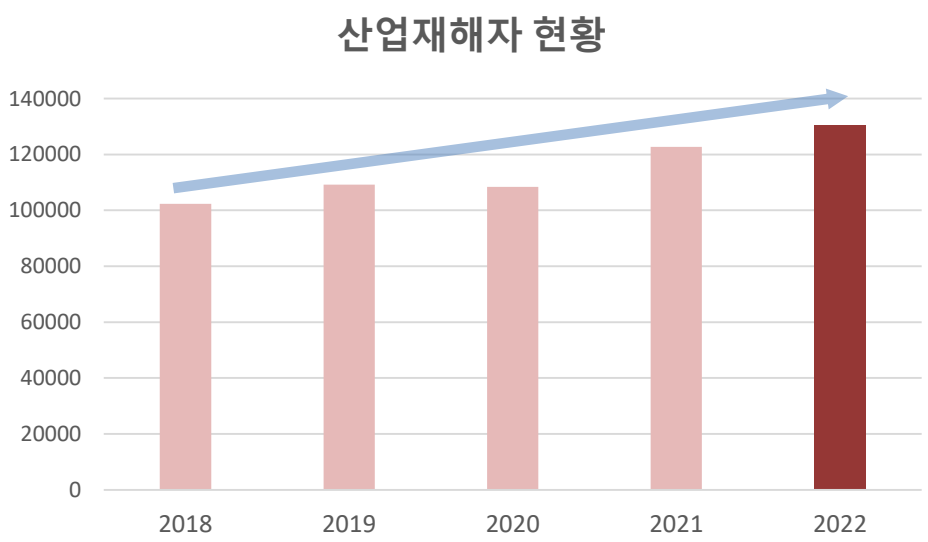
산재통계 분류 방법 및 물량

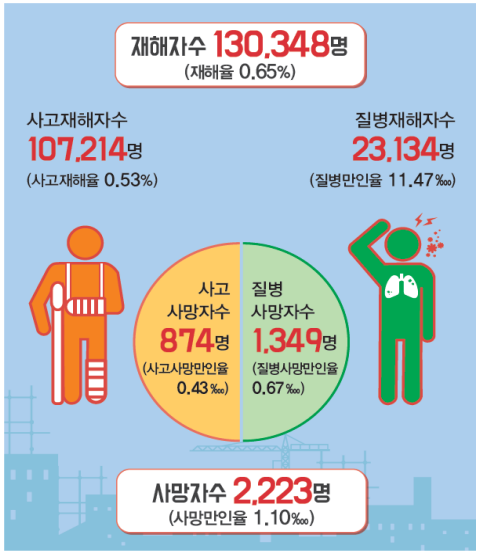
▶▶ **(방법)** 근로복지공단으로부터 승인된 재해개요를 받아

분류자가 텍스트를 읽고, 수작업 분류

※ 발생형태는 120여개, 기인물은 840여개 코드

▶▶ **(물량)** 2022년 기준, **연간 13만건**(일평균 4~500여건), 매년 증가 추세





1 추진 목적 및 개요

자체 분석 개요

▶▶ **(분석 기간)** 2023년 5월 중 ~ 9월 말

▶▶ **(분석 내용)**

- **(텍스트 분석)** 재해개요의 문장 길이, 주요 키워드 등 분석 및 텍스트 전처리 방안 탐색
- **(인공지능 모델링)** 언어 모델 및 분류 방안 모색, 모델링 수행 및 성능 평가, 후보 모델 간 성능 비교
- **(결과 분석)** 최적 모델의 분류 매커니즘 및 적정성 분석

일터안전에서 국민안심으로

2

산업재해 발생개요 주요 특징

산업재해예방
안전보건공단

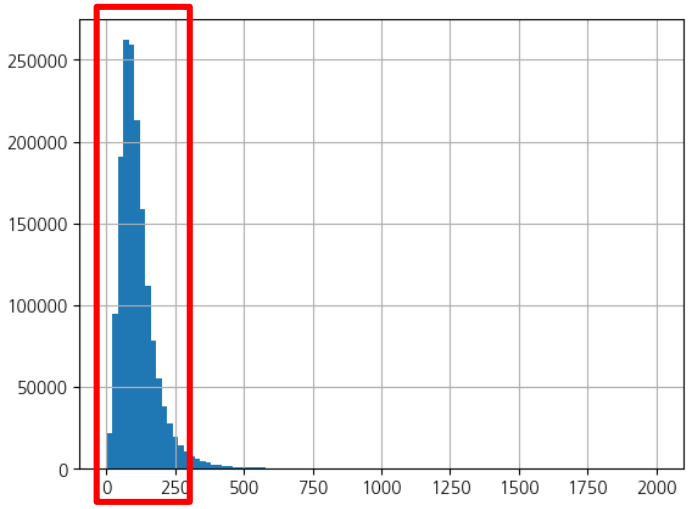
2 산업재해 발생개요 주요 특징

문장 형태

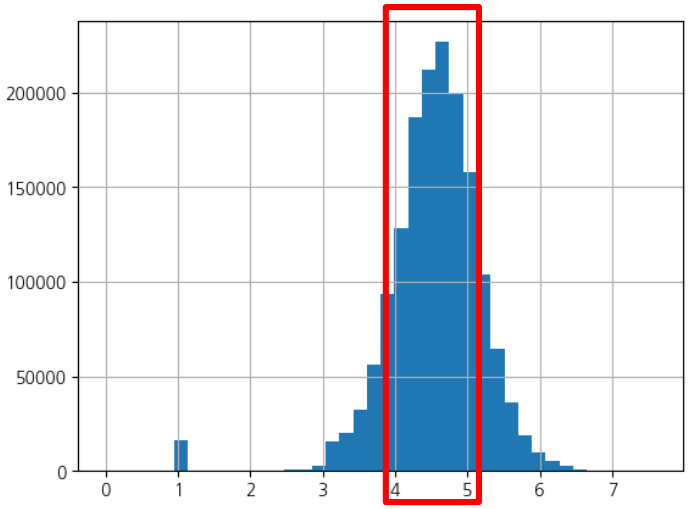
▶▶ (대상) 2007년부터 2022년까지 159만건 / 문장길이는 주로 60~140자 내외로 작성

Mean	Std	Min	25%	50%	75%	Max
102.6	73.96	1*	68	98	140	2,000

* 아무 글자가 없는 경우 공백문자(1)로 카운트(정제 전 재해개요 기준)



문장길이 빈도 분석(bins=100)



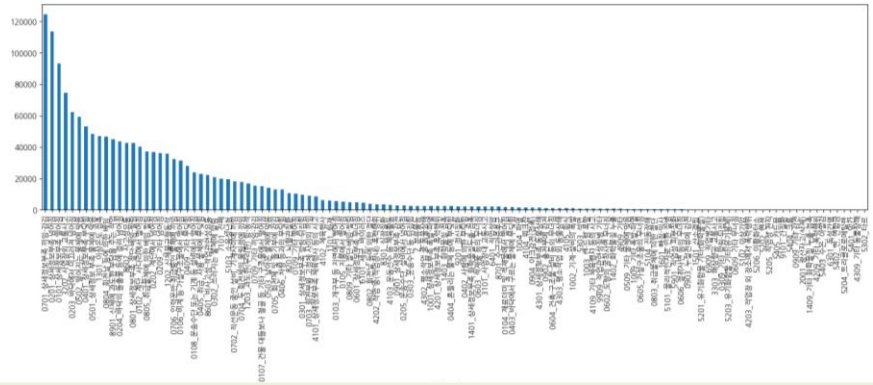
문장길이(Log) 빈도 분석(bins=40)

2 산업재해 발생개요 주요 특징

분류 형태

▶▶ **(발생형태)** 26개 대분류, 125개 중분류로 구분

→ **실제 분류된 데이터는 122개 중분류 항목으로 분류**



▶▶ **(기인물)** 10개 대분류, 840개 세분류로 구분

→ 실제 분류된 데이터는 769개 세분류 항목으로 분류

▶▶ **(조합)** 각각 독립 가정 시 이론적으로는 105,000가지 조합 발생(125 x 840)

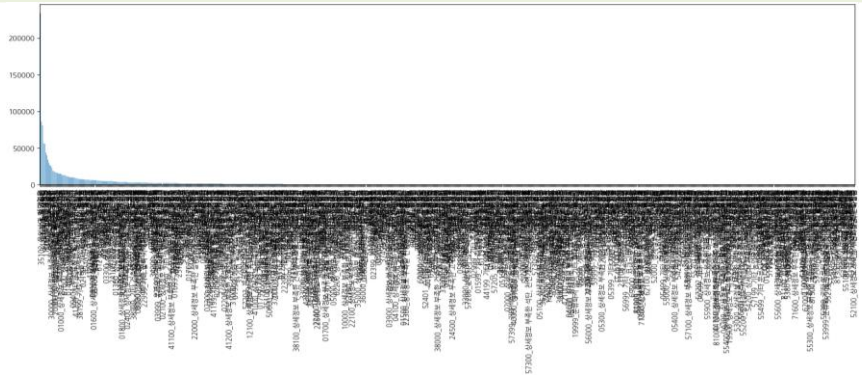
→ 최근 10년 동안 17,202가지의 조합 발생

2 산업재해 발생개요 주요 특징

분류 형태

- **(발생형태)** 26개 대분류, 125개 중분류로 구분
  → 실제 분류된 데이터는 122개 중분류 항목으로 분류

- **(기인물)** 10개 대분류, 840개 세분류로 구분
  → 실제 분류된 데이터는 769개 세분류 항목으로 분류



- **(조합)** 각각 독립 가정 시 이론적으로는 105,000가지 조합 발생(125 x 840)
  → 최근 10년 동안 17,202가지의 조합 발생

2 산업재해 발생개요 주요 특징

분류 형태

- **(발생형태)** 26개 대분류, 125개 중분류로 구분
  → 실제 분류된 데이터는 122개 중분류 항목으로 분류

- **(기인물)** 10개 대분류, 840개 세분류로 구분
  → 실제 분류된 데이터는 769개 세분류 항목으로 분류

- **(조합)** 각각 독립 가정 시 **이론적으로는 105,000가지 조합** 발생(125 x 840)
  → 최근 10년 동안 17,202가지의 조합 발생

분류 항목별 레이블 종류가 많고, 각각의 조합도 고려할 필요

2 산업재해 발생개요 주요 특징

문장의 형태

▶▶ **(어휘 관점)** 문맥에 따라 동음이의어나 다의어

(예) 중량물을 들고 이동하다 **다리**에 걸려 넘어짐 **(사람 다리? / 물건 다리? / 교량?)**

공사 중 **종이** 떨어져 머리에 부딪힘 **(종(Bell) / 종이(Paper))**

▶▶ **(구조 관점)** 주어, 목적어, 수식 관계 등이 불분명

(예) **목격자는 피재자와 트럭기사가** 상차 작업하는 것을 구경하던 중에 적재물이 쏟아짐

(피재자와 트럭기사 작업하는것을 구경? / 피재자와 트럭기사가 함께 작업한 것을 구경?)

피재자는 고장 난 장비를 고치려고 **상부에서 조립하던 중 미끄러져** 손가락을 다침

(피재자가 바닥이 미끄러워 넘어진 것인지, 도구나 자재가 미끄러져 다친 것인지 모호함)

▶▶ **(맞춤법)** 오타나 부정확한 철자, 띄어쓰기 등 다수(tokenization / tagging에 영향)

(예) 하수구 **덮게**을 잘못 짚어서 **덮게**가 빠지면서 바닥에 **부디쳐** 얼굴이 **환몰(맞춤법)**

MCC반 부하설비 고장 점검 중 **불꼬치** 튀어 착용한 옷에 불이 붙음 **(불꽃 / 음식)**

2 산업재해 발생개요 주요 특징

정보의 품질

▶▶ **(내용 관점)** 기술된 내용이 완전하지 않아 해석의 차이

(예) 현장에서 내려오다가 **1m 높이**에서 떨어져 다침 **(단부? / 계단? / 개구부?)**

현장으로 **차**를 끌고 가다가 **사고**가 남 **(교통 사고? / 자가용? / 화물차?)**

▶▶ **(용어 관점)** 해당 현장 또는 관련 업계에서만 사용

(예) **야기리** 작업 중 상하 이동 중에 멍에가 떨어져 손등을 다침

M-15 작업 중 단부를 보지 못해 바닥으로 추락함

단어 수준을 넘어 문맥을 고려한 분석 필요

일터안전에서 국민안심으로

3

모델 설계 방향

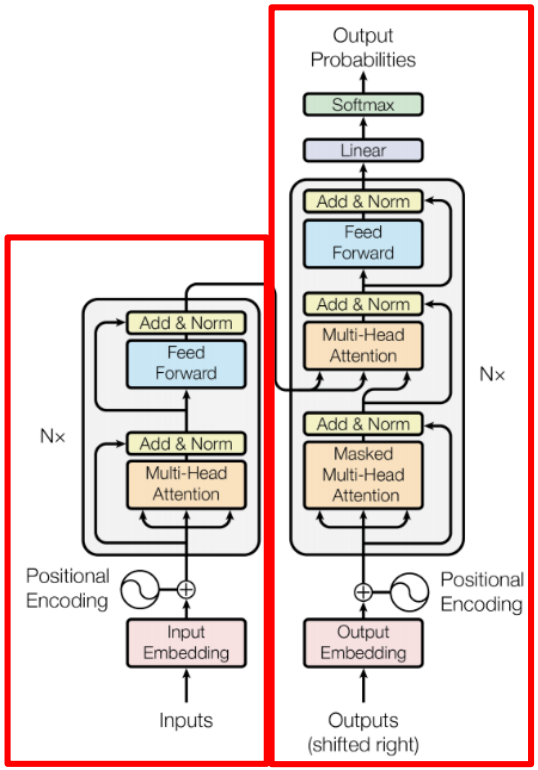
산업재해예방
안전보건공단

딥러닝 기반 언어 모델

- 구글 연구원들이 2017년 논문에서 시퀀스 모델링을 위해 제안한 새로운 신경망 아키텍처(Transformer)
  - 인코더-디코더 프레임워크
  - 어텐션 매커니즘
  - 전이 학습
- 기계번역작업의 품질과 훈련 비용 면에서 기존의 순환신경망(RNN) 능가
- 효율적인 전이 학습이 가능, 적은 양의 데이터로도 최고 수준의 텍스트 분류 모델임을 입증
- 자연어 처리 모델 발전의 촉매
  대표적인 것이 BERT*와 GPT**

*BERT(Bidirectional Encoder Representations from Transformers), **GPT(Generative Pre-trained Transformer)

< Transformer 구조 >



BERT(인코더) GPT(디코더)

딥러닝 기반 언어 모델

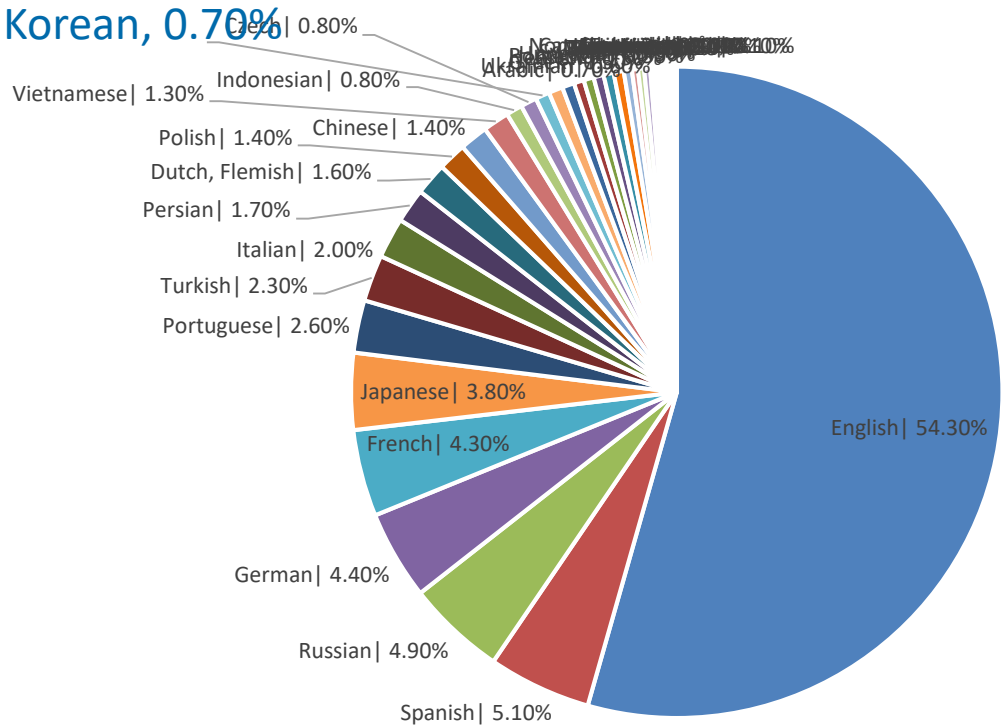
▶▶ 훈련 시간이 오래 걸리고, 하드웨어 자원도 한계

	BERT	RoBERT	DistilBERT
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base**: 66
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base:** 8 x V100 x 3.5 days; 4 times less than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	5% degradation from BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation

V100 그래픽카드
- GV100 GPU
- 16GB RAM
- 5120 CUDA cores
- 640 Tensor Cores

< Comparison of BERT(Bidirectional Encoder Representation from Transformers): Kdnuggets(2019.09) >

언어: 한국어



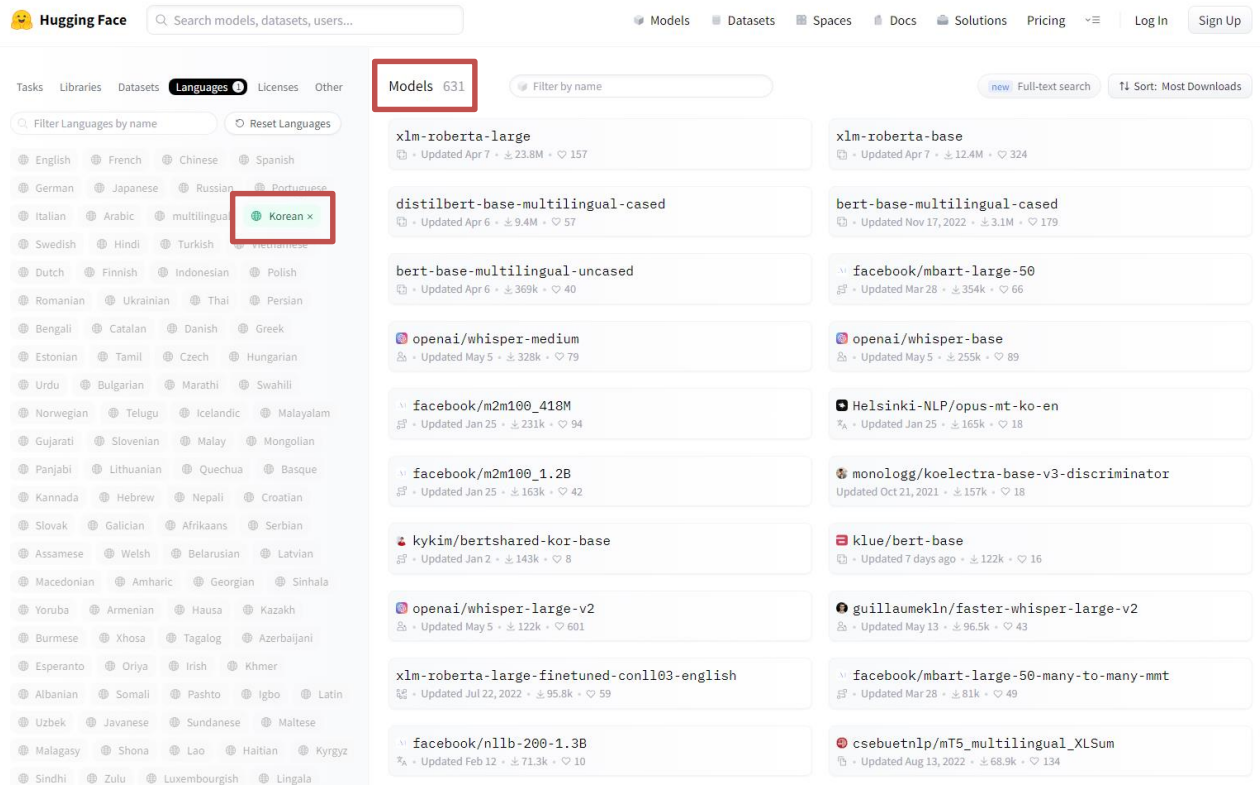
순위	언어	비율
1	English	54.30%
2	Spanish	5.10%
3	Russian	4.90%
4	German	4.40%
5	French	4.30%
6	Japanese	3.80%
7	Portuguese	2.60%
8	Turkish	2.30%
9	Italian	2.00%
10	Persian	1.70%
11	Dutch, Flemish	1.60%
12	Polish	1.40%
13	Chinese	1.40%
14	Vietnamese	1.30%
15	Indonesian	0.80%
16	Czech	0.80%
17	**Korean**	**0.70%**
18	Arabic	0.70%
19	Ukrainian	0.60%
20	Greek	0.50%
21	Hebrew	0.50%
22	Romanian	0.50%
23	Hungarian	0.50%
24	Swedish	0.50%
25	Thai	0.40%
26	Danish	0.30%
27	Slovak	0.30%
28	Finnish	0.30%
29	Bulgarian	0.20%
30	Serbian	0.20%
31	Croatian	0.10%
32	Norwegian Bokmål	0.10%
33	Lithuanian	0.10%
34	Slovenian	0.10%
35	Catalan, Valencian	0.10%
36	Estonian	0.10%
37	Norwegian	0.10%
38	Latvian	0.10%
39	Hindi	0.10%

< Usage statistics of content languages for websites(w3techs.com) >

3 모델 설계 방향

언어: 한국어

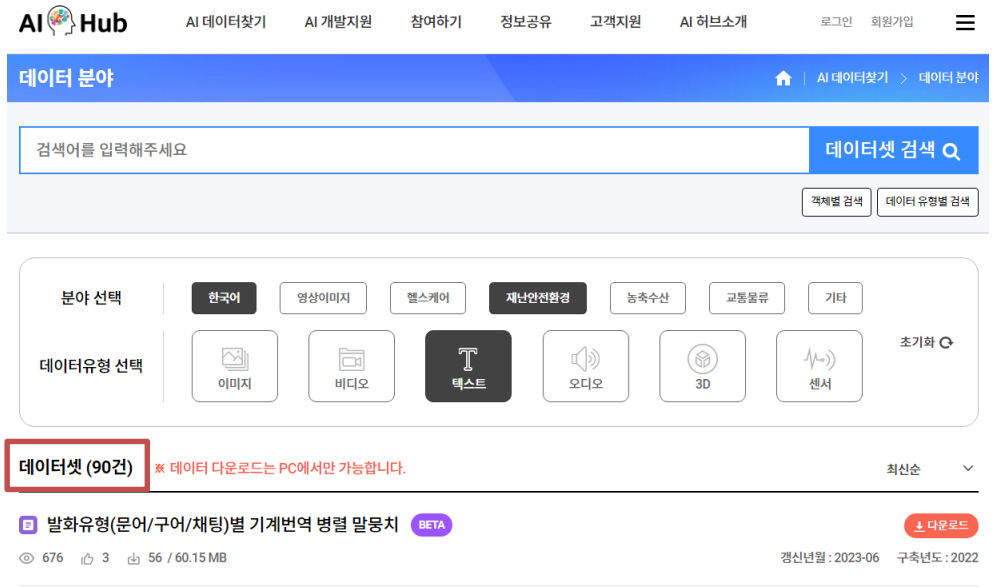
▶▶ 등록된 모델 전체 23만여개(Hugging Face기준) 중 한국어 모델은 631개(약 0.2%)



3 모델 설계 방향

언어: 산업안전분야

▶▶ 산업안전분야에 특화된 학습용 말뭉치(corpus) 추가 수집 한계



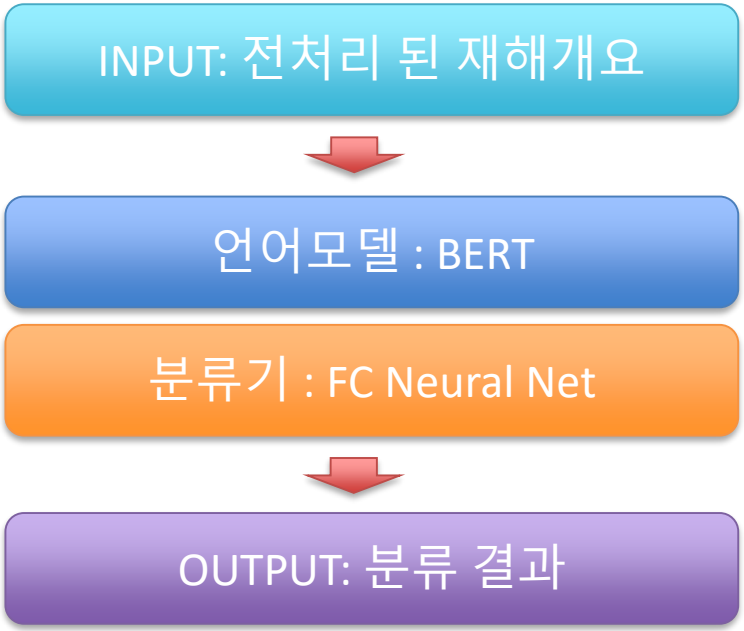
새로 언어 모델을 만들기 보다는 기존 모델을 변형, 전이학습이 효과적

모델 설계 방향

▶▶ (언어 모델) 한국어 또는 다국어 BERT 모델을 종류별로 선택적 활용

[BERT Base(기본) / BERT Multilingual / BERT Distil(경량화 모델)]

▶▶ (분류기) 언어 모델을 통과한 결과에 분류층을 쌓아 분류(Text Classification)



3 모델 설계 방향

분류기 설계

언어모델(BERT / BERT Multilingual / DistilBERT) [재활용 가능]

Fully Connected Layer(1024~4096) x 4 Layer
Dropout(0.1~0.3), RELU/GELU [가변적 구성]

Softmax, CrossEntropyLoss

재해자구분(5)*, 발생형태(122), 기인물(769) = 896노드

*사고부상,질병이환,사고사망,질병사망,그외사망

OUTPUT: 분류 결과

HW

Ryzen 5600G
RAM 32G
RTX 3060 12G

SW

Python 3.10.6
Torch 2.0.0
Transformers 4.28.1

PARA

LearningRate = 1e-5
BatchSize = 24
MaxLength = 300
EarlyStopping = False

일터안전에서 국민안심으로

4

분류 모델 시범 개발 현황

산업재해예방
안전보건공단

4 분류 모델 시범 개발 현황

텍스트 전처리

- (띄어쓰기) Soyspacing 휴리스틱 알고리즘 적용(github.com/lovit/soyspacing)
  - 산업안전보건분야에서 띄어쓰기가 잘 되어있는 데이터를 구하기 쉽지 않고, 재해개요 데이터는 대부분 띄어쓰기가 맞지 않음
  - 대다수의 단어 분포(Vocabulary Distribution)을 반영하여 빠르게 교정 컨셉
    예) '~하던중' → '~하던 중'으로 교정해야 하나,
    대부분 '~하던중'으로 틀리게 기술한 경우, '~하던 중'을 '~하던중'으로 교정
- (맞춤법) 아직 적용하지 못함 → 추후 검토(폐쇄망 환경 제약)
- (불필요 문자 제거) 한글, 영문, 띄어쓰기를 제외하고 모두 제거(정규식 활용)
- (토크나이저) BERT 모델별 기본 토크나이저 사용
- (학습데이터) 최근 10년치 자료로 한정

4 분류 모델 시범 개발 현황

분석 현황

▶▶ KoBERT / DistilBERT 모델 <추후 재검토>

- 한국어 특화 언어 모델, Base 모델보다는 다소 적은 파라메터(92M < 110M), 단어사전이 8,002개, 상대적으로 모델이 가볍고 학습속도 비교적 빠름
- 발생형태 단일 분류학습에서 학습 정확도(Accuracy)가 0.4미만에서 정체 ('22년도 모델은 학습 시 발생형태 0.7 / 기인물 0.6의 성능을 보였음)
- Tokenizer encode 과정에서 재해개요 대부분의 단어가 <unk>으로 처리 확인

원형기둥 해체 작업과정에서 상부거푸집이 떨어지는 것을 꼭 잡아서 일어난 일임

<CLS> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <SEP>

▶▶ Tokenizer를 공유하는 KoBERT모델도 동일 현상 → 단어사전 문제라 판단 제외

※ Transformer API 문제, KoBERT 라이브러리 활용 시 정상 작동 → 추후 재분석 예정

4 분류 모델 시범 개발 현황

분석 현황

▶▶ BERT Multilingual MSMarco(MS Machine Reading Comprehension) 모델

- 한국어 포함 100개 이상의 언어를 지원, BERT Multilingual을 Base로 MSMarco 학습
- 단어사전이 105,879개로 많으며, DistilBERT 대비 모델이 크고, 학습 속도 느림
- 발생형태/기인물 분류학습에서 검증 정확도(Accuracy)가 각 0.8, 0.6 수준

▶▶ KPF-BERT CrossEncoder 모델 => 사전 학습 시도

- 한국언론진흥재단에서 20년치 기사 학습한 모델을 베이스로 문장쌍 유사도 학습
- 안전보건자료 활용 단어사전 36,440개 => 47,003개로 확장, MLM(15% 적용) 학습

분류기의 문제 보다는 언어모델의 문맥 해석결과가 더 중요하다고 판단

4 분류 모델 시범 개발 현황

분석 현황

INPUT: 전 처리된 재해개요

언어모델 : BERT

분류기 : FC Neural Net

OUTPUT: 분류 결과

Masked Language Model(MLM) 학습

예) 나는 내일 <MASK>에 가고 싶다 → 카페, 극장, 운동장, 학교

재해개요 150만 건
산업안전보건법
산업안전보건법 시행령
산업안전보건법 시행규칙
기준에관한 지침, 고시/지침
KOSHA GUIDE(약 1,400건)

1Epoch 약 55시간 소요(학습 loss 1.13 → 0.18)

테스트 정확도 기준: 발생형태 0.8 / 기인물 0.7

적은 양이라도 산업안전보건 텍스트 MLM 학습을 통해 문맥 해석 능력이 향상

4 분류 모델 시범 개발 현황

검증('23년 4월 통계 자료 기준)

▶▶ 학습한 방법이 다르기 때문에 단순 비교는 한계가 있으나,

- DistilKoBERT(Tokenizer 문제) < MARCO BERT < KoBERT('22) < KPF BERT(MLM학습) 순

※ 모델 간 성능 차 보다는 한국어 및 안전보건분야의 언어 학습 여부 차이라 판단

- 언어모델 재활용을 통해 분류기 성능 고도화 및 다른 작업 활용 가능

구분	MARCO BERT	KoBERT('22)	KPF BERT(MLM)		RELU	GELU
발생형태	0.177	**0.750**	0.748		0.743	0.740
기인물	0.547	0.507	**0.656**		0.644	0.644
산술평균	0.362	0.629	**0.702**	신경망 재구성	0.694	0.692
조화평균	0.267	0.605	**0.699**		0.690	0.689

발생형태, 기인물 개별 모델

단일 모델
정확도 +15% 속도 +67%

2Epoch만으로도 성능의
98%

공표 가능한 수준의 정확도(95% 이상)는 달성하지 못해 공식화 한계

일터안전에서 국민안심으로

5

향후 분석 계획

산업재해예방
안전보건공단

5 향후 분석 계획

언어 모델 개선

- 안전보건 말뭉치(Corpus) 추가 학습
  - 조사보고서, 안전보건용어 설명집, 민원자료, 미디어 자료 등
- 사람(정답)과 인공지능 분류의 오분류 사례 및 loss 검토를 통해
  - 노이즈에 강건한 모델, 효과적인 처리 방안(전처리 등), 재학습 방안 모색

구분		사람																								
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	31	32	33	41	42	43	49	합계	정확도	정확도 순위	빈도 순위
		떨어짐	넘어짐	깔림 뒤집힘	부딪힘	물체에 맞음	무너짐	끼임	절단 베임찔림	감전	폭발파열	화재	불균형및무리한동작	이상온도 접촉	화학물질 누출접촉	사업장내 교통사고	사업장외 교통사고	체육행사	폭력행위	동물상해	기타					
23 인공지능	01 떨어짐	1,008	55	5	7	9	11	3					16										1,114	90.48	6	2
	02 넘어짐	63	1,984	14	38	4		4	3			1	90	10			3	1	9	2			2,226	89.12	8	1
	03 깔림뒤집힘	3	6	174	7	10	2	4	2					3		1	9						221	78.73	17	10
	04 부딪힘	6	25	4	564	26	1	40	7				14	6		1	5		3	1			703	80.23	14	5
	05 물체에 맞음	12	3	8	37	543	5	6	5				2	3	1							1	626	86.74	9	6
	06 무너짐	2	2	3	2	2	23	1															35	65.71	20	14
	07 끼임	4	5	5	44	15	1	978	31	1			7	2									1,093	89.48	7	3
	08 절단베임찔림	1	5	3	75	20		23	722				4										853	84.64	13	4
	09 감전		1							19													20	95.00	3	18
	10 폭발파열				1						13		1	3									18	72.22	18	19
	11 화재									2	4	23		4									33	69.70	19	15
	12 불균형및무리한동작	14	28	3	12	1	1	7	1				445	2									514	86.58	10	8
	13 이상온도 접촉		1		2	1		1						349	2								356	98.03	2	9
	14 화학물질누출접촉							1			1			4	23								29	79.31	16	16
	31 사업장내교통사고															2							2	100.00	1	23
	32 사업장외교통사고	1	19	34	13			2					4			2	475		1		1		552	86.05	11	7
	33 특성장공교통사고							1										4					5	80.00	15	21
	41 체육행사	1	3		2								3						133		2		144	92.36	4	11
	42 폭력행위		1	1	5															42			49	85.71	12	12
	43 동물상해				1	1													1		33		36	91.67	5	13
	51 물리적인자										1												1		21	25
	82 뇌혈관질환		3										1										4		21	22
	86 요통	1	1		3			1					21						1				28		21	17
	89 기타 근골	1	1		1								12										15		21	20
	92 스트레스성질환			1									1										2		21	23
	Z 분류불능									1													1		21	25
	합계	1,117	2,143	255	813	633	44	1,072	771	23	19	24	621	386	26	6	492	5	148	45	36	1	8,680			
	정확도	90.24	92.58	68.24	69.37	85.78	52.27	91.23	93.64	82.61	68.42	95.83	71.66	90.41	88.46	33.33	96.54	80.00	89.86	93.33	91.67	-				
	정확도 순위	9	5	18	16	12	19	7	3	13	17	2	15	8	11	20	1	14	10	4	6	21				
	빈도 순위	2	1	10	4	6	13	3	5	17	18	16	7	9	15	19	8	20	11	12	14	21				

구분		사람													
		0	1	2	3	4	5	6	7	8	Z	합계	정확도	정확도 순위	빈도순위
		설비기계	인력기계	부품재료	건축표면	용기용품	화학물질	교통수단	사람 동식물	자연현상	분류불능				
23 인공지능	# 해당없음	5	5	12	14	22		3	1			62	-	11	8
	0 설비기계	1,158	57	82	63	23	2	9	1		1	1,396	82.95	4	2
	1 인력기계	60	779	12	12	11		3	2			879	88.62	3	6
	2 부품재료	137	37	627	127	55		8	3			994	63.08	8	3
	3 건축표면	58	26	92	2,976	61	2	33	25		1	3,274	90.90	1	1
	4 용기용품	68	10	15	76	728		7	6	2		912	79.82	5	4
	5 화학물질	4		1		3	8					16	50.00	9	9
	6 교통수단	14	2	10	53	6		819	6			910	90.00	2	5
	7 사람동식물	1	12	2	24	8		3	172			222	77.48	6	7
	8 자연현상			1		2				1		4	25.00	10	11
	Z 분류불능		1	1						1	8	11	72.73	7	10
	합계	1,505	929	855	3,345	919	12	885	216	4	10	8,680			
	정확도	76.94	83.85	73.33	88.97	79.22	66.67	92.54	79.63	25.00	80.00				
	정확도 순위	7	3	8	2	6	9	1	5	10	4				
	빈도순위	2	3	6	1	4	8	5	7	10	9				

< 사람 – 인공지능의 발생형태 및 기인물 분류 비교표 >

5 향후 분석 계획

분류기 개선

▶▶ 분류 방식을 변경하여 분류기의 성능을 최적화

(現) 신경망 → (改) 불균형 데이터 처리를 위한 학습 가중치 조정 고려

▶▶ 작업지역공정, 작업 내용 분류 방안 모색

- 현재 공단은 작업지역공정 / 작업 내용 미분류(데이터 없음)
- 조사표 분류는 있으나, 정확도가 낮아 학습 데이터로 활용 한계

→ 발생형태, 기인물 분류 모델에서 Task만 변경하여

1) 정확도가 낮더라도 조사표 분류를 학습 시키는 방안

2) 일부를 재분류하여 Few Shot 또는 Zero Shot Classification을 적용하는 방안

3) Q&A나 MLM 등 그 외 다른 방안(생성형 전환) 모색

※ 정확한 데이터가 없어 3가지 경우 모두 높은 성능은 장담할 수 없음

일터안전에서 국민안심으로

감사합니다

산업재해예방
안전보건공단

재해개요 분류모델 개발 필요성 및 정책적 활용방안

안전보건감독기획과 김동현

1. 중대재해감축 로드맵과 산재통계 분석

2022년 우리나라의 업무상사고사망자는 874명으로 아직도 하루 평균 2.4명의 근로자가 산업재해로 사망하고 있다. 그 동안 정부는 산업재해를 줄이기 위한 여러가지 방법을 사용해왔고 어느 정도 성과를 거두기도 했지만 비슷한 경제규모의 다른 나라와 비교할 때 산업재해가 빈번하게 발생하고 있다.

따라서 산업재해를 줄이기 위한 새로운 접근방법이 시도되고 있으며, 2022년 11월 30일 발표된「중대재해감축 로드맵」은 처벌·감독을 통한 타율적 규제가 아닌 안전주체들의 책임에 기반한 '자기규율 예방체계'를 중심으로 산업재해를 예방한다는 것이 주된 내용이다.

중대재해감축 로드맵에는 산업재해 통계와 관련된 다음 내용도 포함되어 있다.

○ **(대상 선정) 산재통계**(보상) **분석** 등을 통해 **재해 발생 경향성**을 **사전**에 확인 후 **감독 방향 설정** → 고위험 기업 **자동 선정***(23)

로드맵에 포함된 이 내용은 어찌보면 당연한 내용이라고 볼 수 있지만, 실무적으로는 큰 변화를 체감할 수 있는 내용이다. 재해 발생 경향성을 사전에 확인한 후 감독 방향을 설정하고, 고위험 기업을 자동 선정한다는 것은 지방관서가 아닌 고용노동부 본부에서 사전에 감독 대상을 선정한다는 의미를 갖기 때문이다.

지금까지의 감독 대상 선정 방법이 본부에서 주제를 정해주면 지방관서에서 그 주제에 맞게 자체적으로 대상을 선정하는 방법이었다면, 앞으로는 주제에 맞는 감독 대상을 일괄적으로 선정하고 이를 지방관서에 시달하는 방법이 주로 사용될 것이다.

산업안전보건 분야에서 감독은 가장 주요한 정책 수단 중의 하나이기 때문에 감독 대상 선정 방법을 바꾼다는 것은 정책 효과에 큰 변화를 줄 수 있는 사안이다. 따라서 정부는 보다 정확한 산재통계 분석을 위해 여러 가지 노력을 하고 있으며 최근 발표한 '산업재해 고위험요인 분석'도 그러한 시도 중의 하나다.

산업재해 고위험요인 분석은 고용노동부와 한국산업안전보건공단이 최근 6년간의 사고사망사례 4,432건을 분석한 결과인데, 그간 일부 재해사례의 사고개요와 재해 원인 등을 제공해오던 것을 전체 사고사망사례를 대상으로 사고개요, 기인물, 고위험작업/상황, 재해유발요인, 위험성 감소대책을 체계적으로 정리하여 발표하였다.

제공되는 분석자료는 다음과 같은 형태이다.

< 제조업· 기타사업 분석자료 중 일부 >

산재 업종			재해개요	기인물	고위험 작업/상황 (35개)	재해유발요인 (101개)	위험성감소대책
대분류	중분류	소분류					
제조업	기계기구 · 금속 · 비금속광물제품 제조업	제강 압연업	2016년 10월경 ○○사업장 지붕위에서 재해자가 방청도료 칠 작업 중 밟고 있던 선라이트(채광창)이 깨지면서 5.4m 아래 바닥으로 추락해 사망	선라이트(채광창)	비정형 작업(정비 · 보수)	작업을 위해 밟고 있던 구조물이 파손되거나 균형을 잃기 쉬운 불안한 자세로 작업하여 떨어짐 등 재해발생 위험	▶안전난간이 설치된 작업발판을 설치·사용 ▶고소작업대 사용 또는 3.5m 미만은 A자형 사다리 사용 (2인1조) 관리 ▶안전대 및 안전모 지급·착용 등 관리감독자 확인
운수창고통신업	철도·항공·창고·운수관련서비스업	철도·궤도 운수업	2016년 3월 ○○역에서 재해자가 이동식 사다리(A형)에 올라가 에스컬레이터 벽체 청소작업 중 사다리가 쓰러지면서 2.5m 아래로 떨어져 사망	이동식 사다리(A형)	사다리를 이용한 작업 또는 통행	아웃트리거를 사용하지 않는 등 사다리를 불안정하게 걸쳐놓고 이동(작업)하던 중 사다리가 넘어지면서 떨어지는 등 재해발생 위험	▶3.5m 미만 작업 시 A자형 사다리만 사용 (2인1조) ▶아웃트리거 설치 및 안전대, 안전모 등 개인보호구 착용
기타의 사업	시설관리및사업지원서비스업	건물등의종합관리사업	2016년 10월경 ○○아파트 화단에서 재해자가 이동식 사다리에 올라가 가지치기 작업을 하던 중 중심을 잃고 떨어져 사망	이동식 사다리(조경용)	사다리를 이용한 작업 또는 통행	사다리 위에 올라가 작업하던 중 균형을 잃고 머리부터 바닥으로 떨어지는 등 재해발생 위험	▶3.5m 미만 작업 시 A자형 사다리만 사용 (2인1조) ▶아웃트리거 설치 및 안전대, 안전모 등 개인보호구 착용

< 건설업 분석자료 중 일부 >

업종	고위험 작업/상황			재해개요	기인물	재해유발요인	위험성감소대책
	공종	작업명	단위 작업명				
건설업	1. 토공사	1.1 굴착 작업	1.1.1 굴착 장비반입	2017년 9월경 OO공사에서 화물차에 적재된 공업용수관(L=9.1m)을 지상으로 내리기 위해 굴착기를 이용하여 인양·하역 중 달기체인 부재(커넥트링)이 하중을 버티지 못하고 파손되어 관이 낙하하면서 하부에 있던 재해재가 머리에 맞아 사망	굴삭기 (백호우)	굴삭기 붐에 달기체인을 연결한 후 공업용수관의 양단에 훅을 걸어 인양 및 하역하던 중 커넥트링이 하중에 버티지 못하고 파손되어 관이 낙하하여 하부에 있던 재해자가 맞음	▶중량물 인양시 Sling rope 각도에 따른 안전하중 값을 고려해 줄걸이 안전수칙 준수 ▶양중용 달기구의 변형 여부 사전점검 실시 ▶자재 인양에 적합한 기계장비 사용
건설업	2. 철근콘크리트공사	2.1 거푸집 작업	2.1.4 거푸집 동바리 인양	2016년 5월경 고속도로 건설공사 현장에서 협력업체 재해자 두 명이 접속교 교각4번 부근에 설치 된 교량 상부슬라브 거푸집 동바리용 멍에브라켓 등을 운반작업 중, 재해자들이 서있던 STEEL BOX GIRDER 사이의 방호선반(강관파이프+합판)이 무너지며, 지상바닥(H≒16m)으로 추락	방망	교각에 설치된 방호선반 위에서 교량 상부슬라브 거푸집동바리용 멍에브라켓 등을 운반작업 중 방호선반이 무너지며 근로자 추락	▶중량물 취급하는 작업할 시 작업계획서를 작성하고 작업지휘자를 지정한 후, 계획에 따라 작업 준수 ▶추락위험이 있는 장소에서 작업 시 안전대 부착설비 설치 및 안전대 착용

즉, 해당 재해의 재해개요, 기인물, 고위험 작업/상황, 재해유발요인, 위험성감소대책까지 일목요연하게 정리한 자료로 이해할 수 있으며, 고용노동부와 산업안전보건공단의 직원들이 몇 차례의 검토를 거쳐 확인한 신뢰할 수 있는 자료라 할 수 있다.

2. 데이터 분석을 통한 재해개요 분류모델의 필요성

산업재해 고위험요인 분석은 그 동안 제공되지 않았던 4,432건의 사망사고에 대한 세부 정보를 정리하여 제공했다는 점에 산업재해 통계가 개선되고 있다는 증거가 될 수 있을 것이다.

하지만, 이 분석자료 외에도 산업재해 예방을 위해서는 더 많은 자료들이 필요하다. 재해 예측이라는 관점에서 볼 때 사업주 또는 산업안전감독관은 재해개요나 고위험 작업/상황, 기인물, 재해유발요인 등을 파악하고, 그와 같거나 유사한

고위험 작업을 할 때, 사고가 많이 발생하는 기인물로 일을 하는 경우 다양한 재해유발요인을 최대한 줄이려는 노력을 해야 한다.

실제로 이런 노력은 산업재해를 예방하는 효과가 있지만 조금 더 생각해보면 고위험 작업/상황, 기인물, 재해유발요인 외에도 아주 많은 요인들이 산업재해 발생에 영향을 미칠 수 있다.

예를 들어 고위험 작업/상황이 '사다리를 이용한 작업 또는 통행'으로 동일하다고 하더라도 사다리의 형태, 재질, 무게, 사용방법 등이 다르고, 작업자의 경험이나 숙련도도 다르다. 또한 사다리가 쓰러진 것인지 아니면 사다리는 서 있는데 작업자가 떨어진 것인지 작업을 하다 사고가 난 것인지 통행을 하다 사고가 난 것인지도 구분하기 어렵다.

물론 이러한 세부적인 내용들이 모두 사고발생에 유의미한 영향을 주었다고 확신할 수는 없지만, 반대로 작업/상황, 기인물, 발생형태 등으로 사고를 분류하는 현재의 분류모델의 설명력이 충분하다는 확신도 없다. 따라서 산업재해 예방을 위해 더 나은 재해모델 분류방법이 존재하는지에 대한 검토는 지속적으로 이뤄져야 한다.

3. 데이터 분석 관점에서 본 더 나은 재해 분류모델 개발 방법론

그렇다면 현재보다 더 나은 재해 분류모델은 어떻게 만들어야 할까? 답이 있는지 없는지 그리고 답이 있다면 그것이 하나인지 여러 개인지 알 수 없지만, 적어도 시작을 어디서부터 해야 하는지에 대해서는 대부분 의견이 일치할 것이라 생각한다.

데이터 분석이라는 방법을 사용해서 재해를 분류한다면 가장 먼저 해야 하는 일은 신뢰할 수 있는 데이터를 확보하는 것이다. 그렇다면 산업재해가 발생했을 때 어떤 데이터들이 생산되는지, 그리고 그 중에서 어떤 데이터가 의미있는지 살펴볼 필요가 있다.

산업재해가 발생하면 대부분 사업장에서 가장 먼저 사고 발생을 알게 되고, 사고를 처리하는 과정에서 고용노동부의 산업안전감독관과 공단의 직원이 사고발생 개요를 파악해서 조사보고서(또는 의견서)를 작성한다. 이 보고서에는 사업장에 대한 내용, 재해자에 대한 내용, 재해발생 경위 및 원인에 대한 내용, 법령 위반 사항에 대한 내용 등이 포함되어 있다.

그 중에서도 가장 중요하게 생각하는 것이 재해발생 경위 및 원인에 대한 내용인데, 실제 사고가 왜, 어떻게 발생했는지를 이해하는데 필수적인 내용이기 때문

이다. 사고에 따라 다르지만 재해발생 경위는 육하원칙에 따라 작성한 몇 개의 문장으로 작성되어 있으며, 재해발생 원인은 감독관(또는 공단 직원)이 판단하는 원인을 열거하는 방식으로 작성된다.

산업재해 통계를 작성하기 위해 통계 담당자가 재해발생 경위와 원인을 읽어보고, 고위험 작업/상황, 기인물, 발생유형 등을 각각의 분류체계에 따라 분류하는 과정을 거치는데, 실제로 하나의 항목으로 분류하기가 어려운 경우가 많아 토론이 필요한 경우가 자주 발생한다.

즉, 재해를 분류하기 위해서 가장 중요하게 활용되어야 하는 자료는 산업안전감독관(또는 공단직원)이 작성한 보고서 중 재해발생 경위와 원인에 대해 작성된 부분이라고 할 수 있다.

또한, 조사보고서에 있는 내용 중에서 사업장, 재해자, 위반 법령에 대한 내용들도 분류모델 개발에 활용하는 방안 역시 생각해볼 수 있다. 두 사업장에서 똑같이 크레인에 의한 사고가 났다고 하더라도 사업장이 언제부터 운영되어 왔는지, 규모·업종은 어떤지, 최근에 인력이 다수 채용되거나 다수의 직원이 퇴사하지는 않았는지, 교대제가 어떻게 운영되고 있었는지 등에 따라 재해예측 결과는 달라질 수 있다. 또한 근로자 측면에서도 해당 근로자의 연령이나 성별, 장애가 있는지 여부, 경력 또는 숙련도, 건강 상태 등 다양한 요인이 재해발생에 영향을 줄 수 있다.

이러한 관점으로 볼 때 산재통계 분석을 통해 재해발생 경향성을 예측하려면 현재보다 훨씬 더 많은 시간과 인력을 동원해서 다차원적인 분석을 실시하거나 새로운 통계분석 방법을 도입해야 할 필요가 있다. 정교한 분석을 위해 인력을 더 많이 투입하는 것보다는 데이터 분석 방법을 활용하는 것이 더 효율적이기 때문에 정부는 데이터 분석 방법으로 재해 분류모델을 개발할 계획이다.

다만, 데이터 분석 방법의 신뢰성이 검증되지 않은 상황에서 가지고 있는 전체 데이터를 활용해서 완전히 새로운 분류모델을 만드는 것은 적절하지 않기 때문에 우선 사람이 하는 일을 알고리즘을 통해 수행해보고 그 결과를 비교해보는 연구를 시범적으로 수행할 필요가 있다.

시범연구는 재해와 관련해서 존재하는 많은 자료 중에서 '재해개요' 자료를 활용해서 재해의 고위험 작업/상황, 기인물, 발생유형을 분류해보고, 사람이 분류한

결과와 비교하여 신뢰성을 검증하는 것이 주된 내용이 될 것이다. 연구를 위해서는 자연어 형태의 재해개요를 분석할 수 있는 자연어처리 알고리즘과 분류체계에 결과를 도출할 수 있는 학습모델을 활용해야 할 것으로 예상된다.

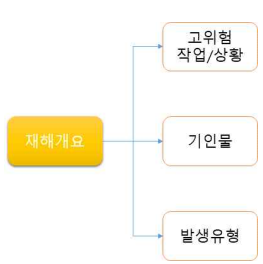
시범연구 이후에는 보다 많은 정보들을 활용하여 기존의 분류 결과에 대한 정확도를 높이는 후속 연구와 우리가 아직까지 알지 못하는 새로운 분류체계가 필요한 것은 아닌지에 대한 연구가 필요할 것으로 예상된다. 물론 이러한 연구는 짧은 기간에 성과를 내지 못할 수도 있기 때문에 중장기 적인 접근이 필요할 수 있다.

4. 재해 분류모델의 연구 목표 및 활용 방안

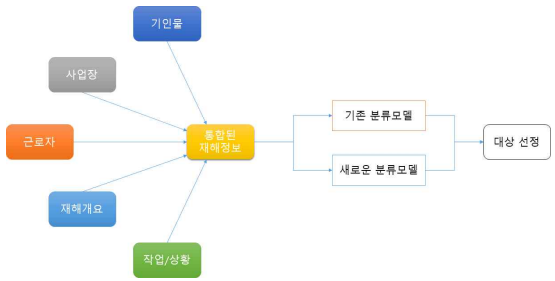
재해개요 분류모델 시범연구로 시작하는 재해 분류모델의 연구의 목표는 실제 현장에서의 활용도가 높은 분류모델을 개발하는 것이다. 재해 분류모델의 활용도가 높다는 것은 개별 사업장의 정보를 바탕으로 발생 가능성이 높은 산업재해를 예측하여 사전에 위험성을 줄일 수 있다는 것을 의미한다.

개념적으로는 재해에 대한 1차원적인 분석을 뛰어넘어 다양한 차원의 정보를 활용하여 재해를 분류하는 것과 기존의 분류체계에서 알 수 없는 새로운 분류체계를 탐색하는 것이 주요한 목표가 될 것이다.

< 현재의 분류모델 >



< 향후 분류모델 >



새로운 재해 분류모델의 예측력이 검증되면 점차 감독대상 선정 등에 활용될 것이고, 이 과정에서 발생하는 피드백 정보를 다시 분류모델에 반영하는 과정은 지속되어야 한다. 즉, 분류모델을 특정한 시기에 개발하는 것이 아니라 지속적인 보완·유지 과정을 이어나가야 할 것이다.

이러한 과정은 정부 단독으로는 수행하기 어렵기 때문에 데이터 공개 및 공동 연구 등의 방법을 활용하여 산업재해 통계 개선을 위한 연구환경을 조성해 나가는 노력 역시 지속할 계획이다.