Journal of KIIT. Vol. 22, No. 7, pp. 33-44, Jul. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.7.33

사전학습 언어모델 기반 도로 건설공사 민원 유형 분류 모델 개발 및 활용 방안

신재영*, 원지선**

Development of the Road Construction Complaint Classification Model based on a Pre-trained Language Model and its Application Strategy

Jaeyoung Shin*, Jisun Won**

본 연구는 본 연구는 과학기술정보통신부 한국건설기술연구원 연구운영비지원(주요사업)사업으로 수행되었습니다 (20240143-001, 미래 건설산업 견인 및 신시장 창출을 위한 스마트 건설기술 연구)

요 약

최근 건설공사 리스크관리 분야에서 인공지능과 자연어 처리를 활용해 비정형 텍스트로부터 위험 요인을 식별하는 사례가 증가하고 있다. 선행연구는 주로 사고, 재난 등 내부 리스크에 초점을 맞추고 있으며, 건설 현장의 민원 등 외부 리스크에 대한 연구는 제한적이다. 본 논문에서는 민원 문서의 유형 분류 모델을 구현하여 도로 건설사업에서 축적된 민원 데이터를 체계적으로 분석하는 접근 방법을 제안한다. 이를 위해 도로 건설 현장의 민원 공문을 대상으로 토픽 모델링을 수행하여 도로 건설공사 민원 유형 분류 기준을 도출하고, 분류 기준에 따라 3만여 개의 학습 데이터를 구축하였다. KoELECTRA 모델을 전이학습하여 민원 유형 분류 모델을 개발한 결과, 모델의 성능은 정확도 0.94, 정밀도 0.93, 재현율 0.94, F1-score 0.93로 나타났다. 아울러, 실무 적용을 위해 5개 지방국토관리청에서 수집된 37,926건의 민원 공문을 유형별로 분류하고 주요 민원 이슈를 분석하였다.

Abstract

Recently, construction risk management is increasingly using AI and natural language processing to analyze unstructured texts. While previous studies focused on internal risks like accidents, research on external risks such as civil complaints has been limited. This paper proposes a method to classify and analyze complaint data from road construction projects using a language model. Topic modeling on civil complaint documents derived classification criteria, creating a dataset of over 30,000 records. The classification model, developed with the KoELECTRA model, achieved an accuracy 0.94, precision 0.93, recall 0.94, and F1-score of 0.93. For practical application, 37,926 complaint documents from five Regional Offices of Construction Management were classified by type, and key issues were analyzed.

Keywords

civil complaint documents, language model, road construction complaint type classification, data-based analysis

* 한국건설기술연구원 전임연구원
- ORCID: https://orcid.org/0000-0002-5917-0472
** 한국건설기술연구원 수석연구원(교신저자)
- ORCID: https://orcid.org/0000-0002-3690-8470

· Received: Apr. 29, 2024, Revised: Jun. 04, 2024, Accepted: Jun. 07, 2024
· Corresponding Author: Jisun Won
Dept. of Future & Smart Construction Research
Korea Institute of Civil Engineering and Building Technology
Tel.: +82-31-910-0083, Email: wonjisun@kict.re.kr

Ⅰ. 서 론

건설공사 현장에서 발생하는 민원은 공사 지연 및 공사비 증가에 영향을 주는 리스크(Risk) 요소이다[1]. 민원을 예방하고 선제적으로 대응하기 위해서는 과거의 민원 사례에 대한 체계적인 관리와 통찰이 필요하다. 현재 국도 건설공사의 민원 데이터는 공문을 통해 수집되고 있으나 민원 업무에 필요한 민원 정보의 접근성과 활용성이 낮은 실정이다. 최근 건설공사 리스크 관리 분야에서는 인공지능, 자연어 처리(NLP, Natural Language Processing) 등 텍스트 분석 기술을 통해 실무자의 경험 기반 의사결정에서 데이터 기반 의사결정 방식으로 전환이 시도되고 있다. 건설공사 리스크 관리를 위해 수행된 텍스트 분석 선행연구는 사 고, 재난 재해 등 내부 위험 요소에 중점을 두고 있다. 반면, 건설공사 현장에서 도출되는 민원과 같은 외부 위험 요소에 관한 연구는 아직 미미한 실정이다.

본 연구는 도로 건설 현장에서 축적되는 민원 데이터의 체계적 관리와 분석 자동화를 위한 초기 연구로, 도로 건설 민원 문서의 유형을 자동으로 분류하기 위한 민원 유형 분류 모델을 개발한다. 또한, 분류 모델을 활용하여 도로 건설 현장의 민원 정보를 시범 분석하여 모델의 활용성을 검증한다. 본 연구의 구성은 다음과 같다. 첫째, 건설공사 리스크 관리 분야에서 딥러닝 기반 NLP 기술을 적용한 텍스트 분석 연구의 동향을 살펴본다. 또한, 도로 건설공사의 민원 유형 분류 기준을 마련하기 위해 건설사업 민원 유형 분류체계에 관한 선행연구를 검토한다. 둘째, 선행연구에서 제시한 건설사업 민원 유형 분류체계를 토대로 도로 건설공사 특성을 반영하여 도로 건설공사 민원 유형 분류 기준을 정립한다. 국도 건설공사 공문을 분류 기준에 따라 레이블링하여 학습 데이터셋을 구축한다. 셋째, 구축한 학습 데이터셋을 활용하여 KoELECTRA 모델을 전이학습(Transfer learning)한다. 이를 통해 도로 건설공사 민원 문서의 유형을 분류하는 모델(이하, ‘도로 건설공사 민원 유형 분류 모델’로 명명)을 구현한다. 이후, 다양한 성능지표를 활용하여 구현한 모델의 성능 평가를 진행한다. 마지막으로, 모델을 활용하여 3만 7천여개의 국도 건설공사 공문을 민원 유형에 따라 분류하고, 이를 기반으로 도로 건설 현장 민원 이슈를 시범으로 분석한다.

Ⅱ. 선행연구 고찰

2.1 건설공사 리스크 관리 분야 NLP 연구 동향

인공지능 기술이 활성화되면서 건설공사 리스크 관리 분야에서는 딥러닝 기반 자연어 처리 기술을 활용하여 비정형 텍스트 데이터로부터 리스크 요인을 추출하고 분석하는 연구가 시도되고 있다[2]-[15]. 안전관리, 계약관리, 품질관리, 현장관리, 민원관리 등 다양한 세부 분야에서 시도되었으며, 특히 안전관리 분야에서 활발히 진행되고 있다(표 1). 이를테면, 안전관리 분야에서는 건설 사고 유형 분류, 위험 요소 등 사고 정보 추출 및 특징 분석, 안전 지식 베이스 구축을 목적으로 NLP 기술이 적용되었으며, 주로 한국산업안전보건공단에서 제공하는 건설 재해사례 데이터가 활용되었다.

데이터 분석 목적 관점에서 보면, 대부분의 연구는 공사일지, 감독일지, 건설사고 신고 사례, 계약문서 등 사업 문서를 활용하여 프로젝트 내부에서 발생하는 리스크 요소 분석에 초점을 두고 있다. 반면, 민원과 같이 공사 현장에 직·간접적인 영향을 주지만 프로젝트 외부에서 기인하는 리스크 요소를 대상으로 한 연구는 부족하다. 관련 연구가 일부 진행된 바 있으나 주로 공개되어 있는 공공데이터 또는 뉴스데이터를 바탕으로, 시설물의 운영 단계에서 발생하는 재해, 하자 등 이슈 혹은 건설사업의 공공갈등과 해외건설 시장 이슈 등 전체 건설산업과 연관된 사회적 이슈를 분석하는데 한정되어 있다[4][10]-[12][15]. 즉, 도로 건설공사단계에서 발생하는 민원 이슈 분석에 관한 연구는 미흡한 상황이다.

건설공사에서 생산되는 사업 공문은 공사 과정에서 발생하는 현장의 내부 리스크 외에도 주민이 제기하는 민원 등 외부 리스크 요소를 포함하고 있다. 특히 민원 공문에는 민원의 요지, 민원에 대한 건설사업관리자의 검토 의견 및 조치 결과 등이 기술된다.

표 1. 건설공사 리스크 관리 분야 NLP 기반 텍스트 분석 연구 사례
Table 1. Research on text analysis based on NLP in the field of construction risk management

Research case	Risk management area	Text type	Purpose of text analysis
S. J. Lim(2020) [2]	Safety	Journal paper	Deriving the role of safety and health management for construction project clients
H. Y. Kim et al.(2021) [3]		Construction accident case report	Classification of accident types in construction accident cases
Y. Kim et al.(2022) [4]		Disaster damage case report	Information retrieval on disaster damage in infrastructure facilities
J. Liu et al.(2023) [5]		Safety standards & accident reports	Extraction of objects and relations related to construction safety
P. Jafari et al.(2021) [6]	Contract	Contract documents	Extraction of reporting requirements and time/cost prediction for each requirement
T. Ko et al.(2021) [7]		Design change documents	Recognition of design change information
H. Lee et al.(2023) [8]		Bidding guide	Classification of contract clauses
D. Zhang et al.(2022) [9]	Quality	Quality report	Recognition and classification of quality control texts
K. Jeon et al.(2024) [10]		Building defect complaint information	Classification of building defect complaint information
S. Baek et al.(2021) [11]	Issue	News articles	Analysis of conflict factors in public construction projects
J. Baik et al.(2023) [12]		News articles	Analysis of issues in the overseas construction market
J. Park et al.(2023) [13]	Schedule	Construction and supervision log	Classification of schedule delay risks based on WBS
S. H. Eom et al.(2023) [14]	Site	Construction project documents	Derivation of risk types at construction sites
T. Chang et al.(2020) [15]	Civil complaint	Complaint data of safety report	Derivation of user inconvenience factors related to facilities

본 연구에서는 NLP 기술을 활용하여 국도 건설공사 공문 텍스트에 나타난 도로 공사 현장의 민원 유형을 도출하고, 민원 유형을 추가 학습한 언어모델을 활용하여 공문을 민원 유형에 따라 자동 분류하고 분석한다.

2.2 건설사업 민원 유형 분류체계 연구 현황

행정안전부 민원처리법[16]에 따르면 민원은 민원인이 행정기관에 대하여 처분 등 특정한 행위를 요구하는 것을 의미한다. 일반적인 민원은 크게 일반민원(법정민원, 질의민원, 건의민원, 기타민원)과 고충민원으로 나뉘며, 요구하는 행위의 목적에 따라서 분류된다. 건설공사에서 민원은 공기 지연과 공사비 증가를 초래할 수 있는 요인으로 작용될 수 있으므로, 공사 과정에서 이를 최소화하는 것이 중요하다[1]. 건설사업의 효과적인 민원 관리를 위해 일부 연구에서는 건설공사 민원을 민원 원인에 따라 유형화하고 있다.

[17]에서는 건축 진정민원을 크게 인적, 물적, 환경적 진정민원으로 정의하고, 법규 항목과 연계하여 민원 발생 요인에 따라 17개로 분류하였다. [18]에서는 건축 진정 민원을 피해유발 요인과 시정조치 요인으로 분류하였다. 피해유발 요인은 주거 및 생활환경 피해, 환경공해, 재산상 피해, 위법사항과 시정조치 요인은 중지 및 조정요구와 기타로 재구분하고, 소분류 기준으로 총 23개 요소(일조권 침해, 소음, 분진, 진동, 균열, 누수, 이격거리 위반, 공사중지 요구, 허가취소요구, 질의 등)를 제시하였다. 공공건설사업 추진시 민원처리 매뉴얼[19]에서는 공공 건설사업의 민원 유형을 계획단계, 설계단계, 공사단계, 준공후 단계로 구분하여 정의하고 있다.

특히 공사단계에서는 민원 유형을 노선변경, 소음/진동피해 예방 및 보상, 환경피해 예방, 교차로 변경/추가 설치, 농/배수로 설치, 건물 조망권 확보, 잔여지 매수 등 13개로 정의하고, 각 유형별 대응 사례를 제시하였다. [1]에서는 선행연구 분석을 통해 건설사업의 민원과 갈등 요인을 포괄하는 민원 항목을 5가지(주민 요구사항, 안전 피해, 인적·물적 피해, 환경 피해, 건설 프로젝트에 대한 홍보 및 안내 부족)로 범주화하고, 그 하위에 세부 요인에 따라 35개 항목을 도출하였다. [20]에서는 문헌에서 도출한 민원 분류체계를 전문가 대상으로 델파이 조사를 실시하여 공사 중 민원 31개, 공사 후 민원 6개를 정의하였다. 공사 중 민원은 5가지 중분류(홍보 및 안내부족, 인적·물적 피해, 환경피해, 안전피해, 생활불편)와 31개 세분류로 구성되었다.

III. 도로 건설공사 민원 유형 분류 모델 개발

본 장에서는 도로 건설공사 민원 유형 분류 기준을 정의하기 위해 선행연구[20]에서 제시한 공사 중 민원분류체계를 도로 건설공사 특성에 맞추어 재정립하였다[21]. 도로 건설공사 특성을 도출하고자 국도 건설사업관리 현장 공문 중 '민원' 키워드로 검색된 민원 관련 공문 데이터 3만 6천여개를 수집 및 전처리하고, TF-IDF(Term Frequency-Inverse Document Frequency) 및 LDA(Latent Dirichlet Allocation) 토픽모델링 기법을 적용하여 키워드 빈도와 토픽 분석을 수행하였다.

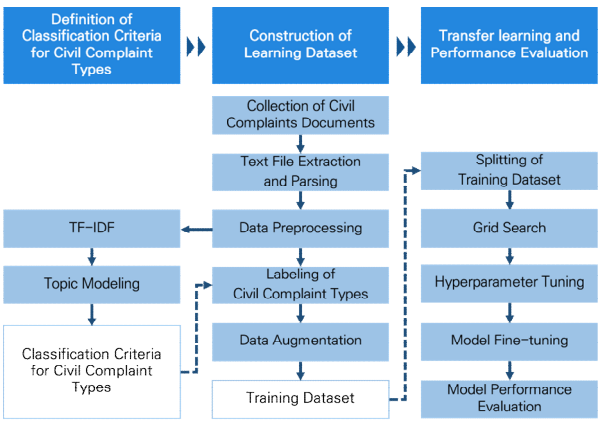


그림 1. 도로 건설공사 민원 유형 분류 모델의 설계 절차
Fig. 1. Design procedure of the road construction complaint classification model

토픽 분석 결과를 토대로 도로 건설공사 민원 유형 분류 기준을 도출하고, 분류 기준에 따라 레이블링 및 데이터 증강을 통해 학습 데이터셋을 구축하였다. 사전 학습된 KoELECTRA 모델을 전이학습하여 도로 건설공사 민원 유형 분류 모델을 개발하였다. 도로 건설공사 민원 유형 모델의 설계 절차는 그림 1과 같다.

3.1 도로 건설공사 민원 유형 분류 기준 정의

민원 공문 데이터 전처리에는 형태소 분석과 결손값 처리, 분석에 불필요한 문자 삭제 등 불용어 처리가 진행되었다. 이외에도 민원 공문 데이터 특성에 기인한 불용어를 5가지 주요 유형(단어 빈도수가 10,000개 이상 또는 2개 이하인 단어, 수발신처, 별첨 등 공문 양식에 사용되는 단어, 지역명/도로명, 사람명 및 기업명 등 개인정보)으로 구분하고 불용어를 제거하였다.

먼저 기존 민원분류체계와 연관된 핵심 키워드별 빈도수(TF-IDF)를 분석한 결과, 실제 사례에서 다뤄지는 민원 유형 간 중요도 차이가 있음을 파악할 수 있었다. 가령 환경피해의 하위 유형인 '대기오염 발생', '수질오염 발생', '지하수위 저하 및 지하수 고갈' 등을 표현하는 주요 키워드(오염, 대기오염, 수질오염, 지하수 등)는 다른 유형에 비해 빈도수가 낮으나 인적·물적 피해의 하위 유형 중 '소음', '진동', '지반침하 및 건물균열'과 관련된 주요 키워드(소음, 진동, 균열 등)는 빈도수가 높았다. 이처럼 실제 사례의 빈도를 고려하여 기존 민원분류체계의 일부 유형을 통합하였다(그림 2 좌측).

LDA는 텍스트 기반 문서의 핵심 토픽을 찾는 데이터 분석 방법론 중 대표적인 알고리즘으로, 확률에 따라 토픽별로 구성된 키워드 분포를 통해 문서에 내재된 주제를 도출하는데 유용하다. LDA 하이퍼파라미터인 토픽 개수를 1부터 15까지 변경하면서 실험한 결과, 토픽 개수를 15개(A~O)로 설정하여 추출된 상위 100개의 키워드 분포에서 유의미한 주제를 도출할 수 있었다(그림 2 우측).

기존 민원분류체계는 15개 토픽 중 8개 토픽을 수용할 수 있지만, '도로 점용(L)', '토지/용지 보상(F, L)', '체불 및 하도급(M)' 관련 3개 토픽을 포괄하는데 한계가 있었다.

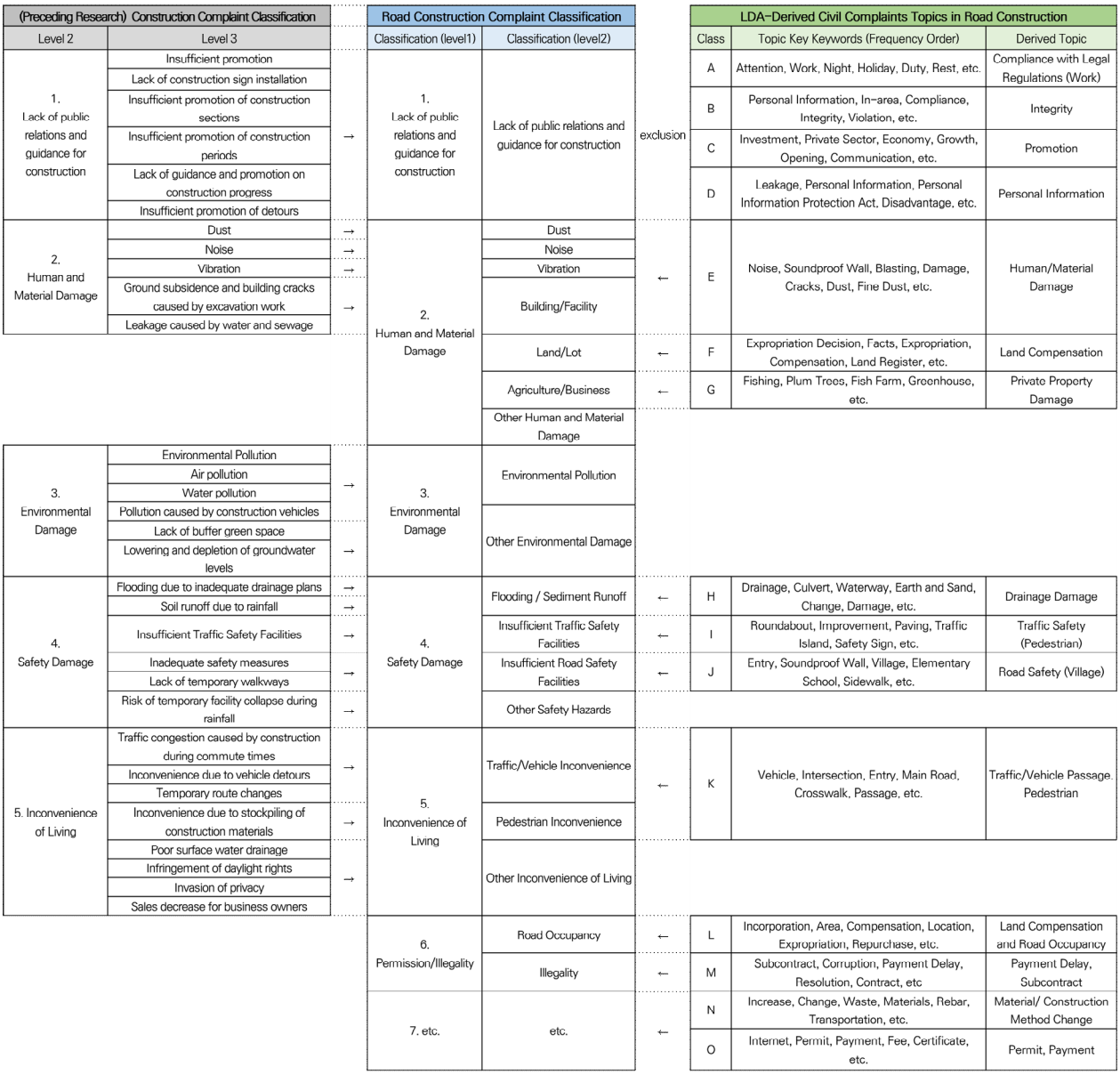


그림 2. 현장 데이터의 토픽 분석 결과를 반영한 도로 건설공사 민원 유형 분류 기준

Fig. 2. Classification criteria for types of civil complaints in road construction based on topic analysis of field data

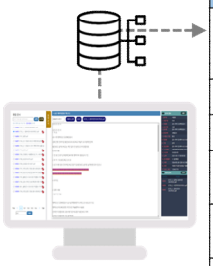
이를 보완하기 위해 기존 분류체계 대분류에 허가/불법 유형을 추가하고, 그 하위에 '도로점용', '불법' 유형을 정의하였다. 또한, 기존 인적·물적 피해의 하위 유형에 '토지/용지'유형을 추가하였다. '자재 공법 변경(N)', '허가, 납부(O)'와 같이 민원의 원인 보다는 조치 사항과 관련된 토픽은 '기타' 유형으로 분류하였다. 한편, 공공 기관의 공문은 홍보, 사례 전파, 법규 준수 등 업무 협조 성격의 내용을 다수 포함하고 있다. 도출된 토픽 중 '법규정 준수(A)', '청렴(B)', '홍보(C)', '개인정보(D)'와 같이 협조 성격의 공문 특성이 반영된 4개 토픽은 민원 이슈와 무관하므로 대상에서 제외하였다. 결과적으로 최종 정의한 도로 건설공사 민원분류체계는 그림 2의 가운데 목록과 같다.

3.2 학습 데이터셋 구축

국토교통부 건설사업관리시스템은 5개 지방국토관리청(이하, 지방청)의 국도 건설공사 사업에서 생산되는 공문, 성과품, 보고서 등 다양한 사업관리 문서를 축적하고 있다.

본 연구에서는 도로 건설공사 민원 유형 분류 목적의 학습 데이터셋을 구축하기 위해 건설사업관리 시스템에 축적된 공문 중 2015년부터 2021년까지 7년 동안 건설사업관리자(감리단)과 시공사 또는 발주기관 간 수발신된 민원 공문 파일 72,163건을 수집하였다. 이 중 텍스트 파일을 대상으로 공문의 텍스트 추출 및 통합 후, 전처리하여 총 37,926개 원천 데이터셋을 확보하였다.

전체 원천 데이터셋 중 8,828개를 대상으로 자체 개발한 어노테이션 툴을 이용하여 그림 2의 도로 건설공사 민원 유형 분류의 중분류를 기준으로 레이블링하였고, 그 결과 9,583개 학습 데이터셋을 구축하였다(그림 3). 구축된 학습 데이터셋은 실제 사례의 분포에 따라 분류된 데이터이기 때문에 클래스별로 데이터 개수의 차이가 존재하였다. 이를 테면, '교통/차량통행 불편' 클래스는 2,099개의 데이터셋이 구축된 반면 '도로점용'과 '진동' 클래스는 각각 77개, 86개의 데이터셋이 구축되었다.



Labeling of Civil Complaint Types Using Annotation Tool

id	Civil Complaint Sentence	Complaint Type
1	Request for installation of noise reduction facilities due to damage caused by vehicle noise adjacent to the open road	Noise
2	Request for purchase of items on remaining land	Land/Lot
3	Request for compensation for crop damage due to road construction	Agriculture/ Business
4	Collapse and soil runoff of adjacent land due to cutting and soil filling of farmland	Flooding / Sediment Runoff
5	Request for improvement of warning signs due to reduction of lanes from three to two	Insufficient Traffic Safety Facilities
6	Accident caused by protruding stones in the middle of the road	Insufficient Road Safety Facilities
7	The existing bus stop in the section was demolished and has not been reinstalled yet ···	Traffic/Vehicle Inconvenience
···	···	···

그림 3. 구축된 학습 데이터 예시
Fig. 3. Example of built training data

분류 모델의 정확도 향상을 위해서는 클래스별로 균형 잡힌 데이터셋을 구축하는 것이 중요하다. 데이터 불균형을 해소하기 위해 데이터 증강 기법 중 RD(Random Deletion), RS(Random Swap) 방식을 적용하여 학습 데이터셋을 증강하였다. RD 방식은 문장에서 임의의 단어를 삭제하는 방식이고, RS 방식은 임의의 두 단어의 위치를 바꾸는 방식으로, 소규모 데이터셋에 적용 시 모델의 성능을 향상시키고 오버피팅을 방지하는데 장점이 있다[22]. 본 연구에서는 20개 민원 유형 분류 중 데이터가 가장 많은 '교통/차량통행 불편' 클래스를 기준으로 증강 비율을 결정하여 증강하였다. 그 결과 민원 유형 분류별 학습 데이터셋은 총 30,439개가 확보되었다(표 2).

표 2. 학습 데이터셋 구축 결과
Table 2. Results of building a training data set

Road construction complaint classification (Level2)	Data count		
	Labeled dataset	Augmentation ratio	Augmented dataset
Lack of public relations and guidance for construction	123	8.1	1,000
Dust	165	6.9	1,145
Noise	867	2.5	2,181
Vibration	86	9.5	819
Building/facility	152	7.0	1,058
Land/lot	1,172	2.5	2,974
Agriculture/business	464	4.1	1,918
Other human and material damage	487	4.0	1,942
Environmental pollution	111	7.7	858
Other environmental damage	129	6.7	862
Flooding/sediment runoff	484	4.0	1,944
Insufficient traffic safety facilities	340	3.9	1,335
Insufficient road safety facilities	539	3.9	2,121
Other safety hazards	188	5.1	962
Traffic/vehicle inconvenience	2,099	1.0	2,099
Pedestrian inconvenience	684	2.5	1,711
Other inconvenience of living	414	4.0	1,650
Road occupancy	77	9.6	736
Illegality	113	7.9	895
etc.	889	2.5	2,229
Total	9,583	(Around) 3.2	30,439

3.3 사전학습 언어모델 기반 모델의 전이학습 및 성능 평가

본 연구에서는 도로 건설공사 민원 유형 분류 모델을 개발하기 위해 표 2의 증강 데이터를 활용하여 KoELECTRA-Base-v3 모델[23]의 전이학습을 진행하였다.

ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)는 기존 MLM(Masked Language Model) 방식을 개선하여 대규모 언어모델 학습을 위한 사전학습 방법으로, 생성자(Generator) 및 판별자(Discriminator) 구조를 활용하여 마스킹된 단어의 위치에 임의의 단어를 생성하고 이를 참, 거짓으로 판별하는 방법으로 학습을 진행한다[24]. KoELECTRA(Korean ELECTRA)모델은 ELECTRA 모델을 기반으로 뉴스, 위키피디아, 모두의 말뭉치[25]의 신문, 문어 등 대량의 한국어 텍스트를 학습한 한국어 사전학습 언어모델이다.

전이학습을 위한 하이퍼파라미터 조건을 도출하고자 전체 학습 데이터셋의 10%를 이용하여 표 3의 범위에서 그리드 탐색을 실시하였다. 표 4와 같이 도출된 하이퍼파라미터 최적값의 조건으로 베이스 모델을 전이학습하였다. 전이학습에는 전체 학습 데이터셋의 80%를 사용하였다. 학습된 모델은 입력 데이터에 대해 각 민원 유형의 예측값을 산출한 후, 소프트맥스 함수를 통해 확률값으로 변환하고 최대 확률값을 갖는 민원 유형으로 분류한다.

표 3. Grid search에 적용한 하이퍼파라미터 범위
Table 3. Hyperparameters range applied to grid search

Hyperparameters	Range of values
learning_rate	1e-5, 2e-5, 3e-5, 4e-5, 5e-5,
batch_size	8, 16, 32
dropout_rate	0.1, 0.2, 0.3, 0.4, 0.5
num_epochs	7, 8, 9, 10, 11, 12

표 4. 모델 학습을 위한 하이퍼파라미터 설정값
Table 4. Hyperparameters settings for model training

Hyperparameters	Setting value
batch size	8
max_position_embeddings	512
embedding_size	768
hidden_size	768
dropout_rate	0.2
optimizer	AdamW
learning_rate	2e-5
epochs	11

다중 분류 모델은 클래스 간 데이터 불균형으로 인한 성능이 왜곡될 수 있으므로, 성능 평가 시 다양한 성능지표를 활용하여 검증하는 것이 필요하다[8]. 본 연구에서는 정확도, 정밀도, 재현율, F1-Score를 성능지표로 활용하고 가중 평균 방식과 Macro 평균 방식을 함께 적용하였다. 모델 성능 평가에는 모델의 학습 시 사용되지 않은 전체 데이터셋의 20%를 사용하였다.

성능평가 결과, 데이터 증강된 학습 데이터셋을 학습한 KoELECTRA 모델은 모든 성능지표에서 0.93 이상의 우수한 성능을 보였다(표 5).

표 5. 학습 모델의 성능
Table 5. Performance of the trained model

Performance metrics	Performance value
Accuracy	0.936
Weighted avg. precision	0.938
Weighted avg. recall	0.936
Weighted avg. F1-Score	0.936
Macro avg. precision	0.930
Macro avg. recall	0.936
Macro avg. F1-score	0.932

표 6. 학습 모델의 클래스별 성능
Table 6. Performance of the trained model

Classes (Complaint classification)	Class-wise performance metrics		
	Precision	Recall	F1-Score
Lack of public relations and guidance for construction	0.98	0.98	0.98
Dust	0.8	0.78	0.79
Noise	0.95	0.82	0.88
Vibration	0.73	0.84	0.78
Building/facility	0.97	0.97	0.97
Land/lot	0.96	0.98	0.97
Agriculture/business	0.97	0.95	0.96
Other human and material damage	0.97	0.95	0.96
Environmental pollution	0.87	0.99	0.93
Other environmental damage	0.94	0.93	0.94
Flooding/sediment runoff	0.97	0.99	0.98
Insufficient traffic safety facilities	0.96	0.95	0.95
Insufficient road safety facilities	0.96	0.97	0.96
Other safety hazards	0.93	0.96	0.94
Traffic/vehicle inconvenience	0.94	0.86	0.9
Pedestrian inconvenience	0.94	0.96	0.95
Other inconvenience of living	0.9	0.94	0.92
Road occupancy	0.97	0.99	0.98
Illegality	0.96	0.95	0.96
etc.	0.92	0.96	0.94

다만 Confusion matrix와 클래스별 F1-score 지표를 통해 세부적으로 분석해 보면, ‘진동’(0.79)과 ‘분진/먼지’(0.79)에 대한 분류 정확도는 다른 클래스에 비해 상대적으로 낮게 나타났으며(표 6), ‘소음’을 ‘분진/먼지’로 분류하는 경우가 있었다. 이는 실제 민원 공문 데이터에서 ‘진동’, ‘분진/먼지’가 대부분 ‘소음’과 함께 복합적인 민원 원인으로 기술되고 있는 점에서 비롯된 것으로 보인다. “소음 및 진동 등으로 학습권이 침해”, “공사로 인한 소음 진동 먼지 피해보상” 등이 그 예시이다.

Ⅳ. 모델의 활용성 검증을 위한 공문의 민원 이슈 시범 분석

도로 건설공사 민원 유형 분류 모델이 현장 실무에 효과적으로 활용되기 위해서는 모델을 통해 민원 데이터를 체계적으로 분석하고, 민원 관리 업무에 유용한 정보를 생성하는 것이 필요하다. 건설사업관리자 관점에서 건설 현장의 민원 업무 수행에 필요한 주요 정보는 민원 이슈별 현황, 지방청별/현장별 민원 추이, 사전예방시설에 대한 민원 현황, 공사중지 및 제3자피해 관련 민원 현황, 유사 민원 및 유사 프로젝트 민원 공문 사례 등이 있다[27].

본 장에서는 제안한 모델의 활용성을 검증하기 위해 모델을 활용하여 앞서 수집한 민원 공문 원천 데이터셋 37,926개(7개년 5개청 공문)를 대상으로 민원 유형에 따라 분류하고, 다양한 관점에서 도로 건설공사의 민원 동향을 시범 분석하였다.

4.1 민원 유형 분류 및 통계 분석

학습 모델을 활용하여 전체 민원 공문 데이터셋 37,926개를 민원 유형별로 자동 분류한 결과(<그림 4>), 대분류 민원 유형별 비율은 ‘생활불편’(12,149건, 32%), ‘기타’(8,836건, 23.3%), ‘인적·물적피해’(7,365건, 19.4%), ‘안전피해’(6,643건, 17.5%) 순으로 나타났다. 중분류 민원 유형의 비율은 ‘교통/차량 통행 불편’(9,214건, 24.3%), ‘기타’(8,836건, 23.3%), ‘토지/용지’(4,498건, 11.9%), ‘침수/토사유출’(2,924건, 7.7%), ‘도로안전시설 미흡’(1,833건, 4.8%), ‘소음’(1,787건, 4.7%) 등 순으로 분석되었다.

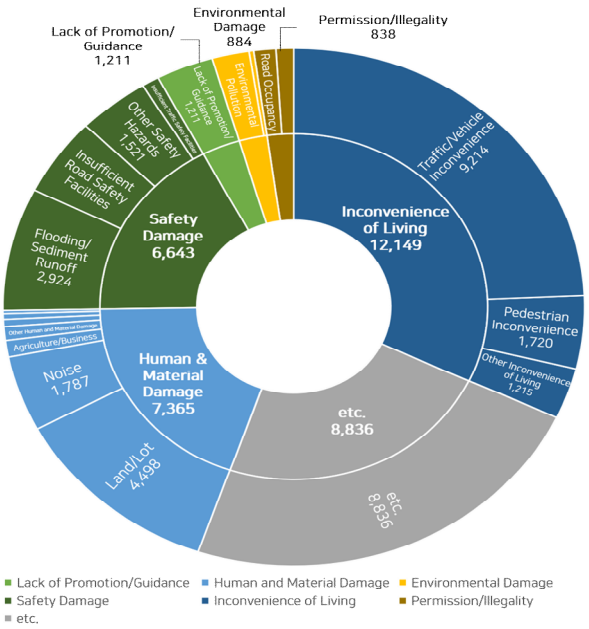


그림 4. 민원 공문의 민원 유형 분류 및 비율
Fig. 4. Classification and portion of civil complaint types in documents

지방청(서울청, 원주청, 대전청, 익산청, 부산청)별로 살펴보면(그림 5), 서울청을 제외한 4개 지방청의 민원 유형별 데이터 분포는 전체 민원 유형별 데이터 분포와 유사하게 나타났다.

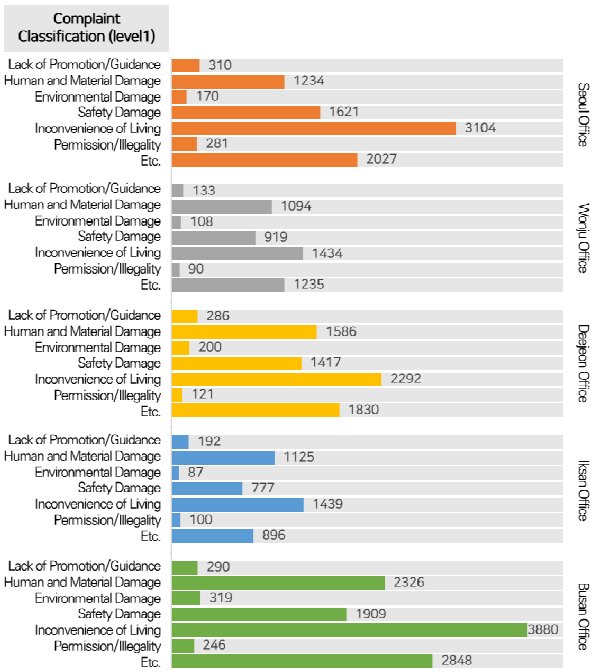


그림 5. 민원 공문의 지방청별 민원 유형 분류 및 빈도
Fig. 5. Classification and frequency of civil complaint types by local office in documents

가령, 서울청 외 4개 지방청의 중분류 민원 유형 비율은 '교통/차량통행 불편'(20.9~27%), '토지/용지'(11.5~16.8%), '침수/토사유출'(7.4~8.8%) 등 순이다. 반면 서울청의 민원 유형 비율은 '교통/차량통행 불편'(23.5%), '보행자 불편(8.5%)', '도로안전시설 미흡(7.9%)' 등 순으로, 도로 통행 및 안전과 관련된 이슈가 주를 이루었다. 일반적으로 유동 인구가 많고 교통의 흐름이 복잡한 도심지의 특성상 도심 내 도로 건설사업은 교통의 혼잡을 최소화하고 기존 도로 시설의 효율적 활용을 위한 세심한 관리가 필수이다. 서울청의 민원 데이터 분포는 수도권 내 도로 건설사업의 특성에서 비롯된 것으로 유추될 수 있다.

4.2 주요 민원 이슈와 연관 시설물 현황 분석

공사 현장에서는 공기 지연 또는 중지를 초래할 수 있는 민원에 대한 선제적 관리와 대응이 중요하다. 특히 제3자피해를 야기할 수 있는 주요 시설물에 대한 예방 관리가 요구된다. 본 장에서는 앞서 민원 유형에 따라 분류된 도로 건설공사 민원 공문 중 '공사 중지'가 언급된 문서와 제3자피해 유형으로 분류된 문서의 텍스트 분석을 통해 주요 이슈와 연관 시설물에 대한 현황을 파악하였다.

'공사 중지', '공사 중단', '공사 정지'가 언급된 공문 총 391개를 대상으로 민원 유형의 중분류 비중을 분석한 결과(그림 6), '교통차량통행 불편'(91건, 23.3%), '기타'(76건, 19.4%), '소음'(43건, 11%), '토지/용지'(27건, 6.9%), '침수/토사유출'(5.4%) 등 순으로 나타났다. 전체 공문 대상으로 분석한 결과와 비교하면 '소음' 유형의 비중이 커진 것을 확인할 수 있었다. 이는 소음 문제가 공사 중지에 영향을 주는 주요한 리스크 요인이 됨을 시사한다. 앞서 도출된 상위 4가지 민원 유형('기타'유형 제외)을 바탕으로 지방청별 현황을 살펴보면(그림 7), 서울청과 부산청에서 상위 4가지 민원 유형은 전체의 55% 이상을 차지한 반면, 대전청에서는 30% 미만으로 가장 낮은 비중을 차지하였다.

'제3자피해' 민원은 민원 유형 분류 대분류가 '인적·물적 피해', '환경피해', '안전피해' 유형인 민원으로 정의하였으며, 해당되는 공문은 총 14,892개이다.

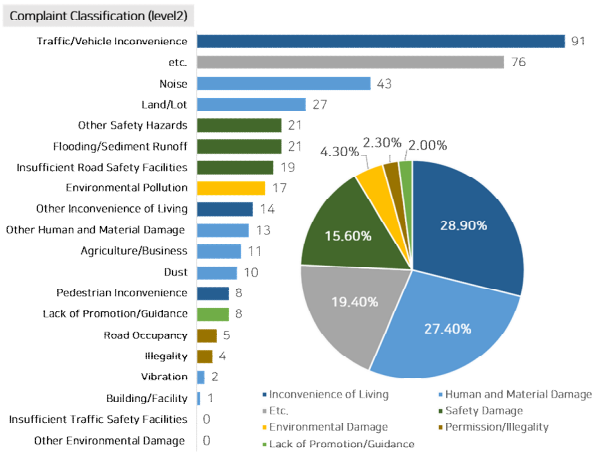


그림 6. '공사 중지' 관련 공문의 민원 유형 분류 및 비율
Fig. 6. Classification and portion of civil complaint types in documents related to 'construction suspension'

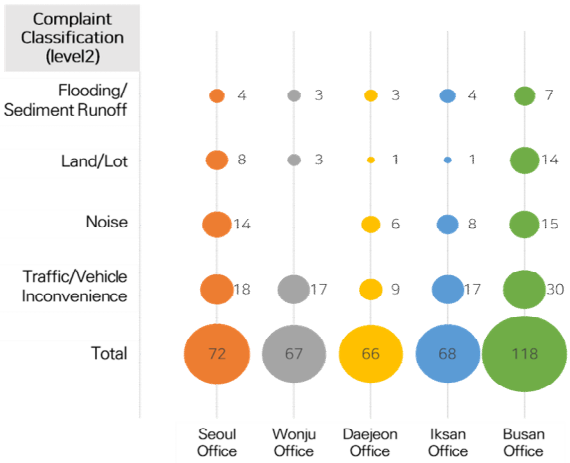


그림 7. '공사 중지' 관련 공문의 지방청별 주요 민원 유형 빈도
Fig. 7. Frequency of civil complaint types by local office in documents related to 'construction suspension'

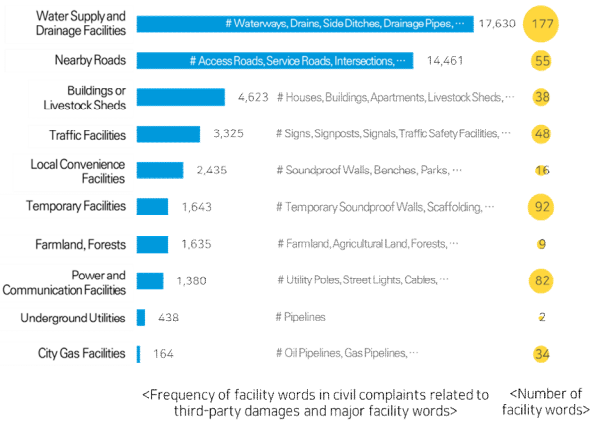


그림 8. '제3자피해' 관련 민원 공문의 중점관리시설 빈도
Fig. 8. Frequency of priority-managed facilities mentioned in documents related to 'third-party damage'

민원 공문에 표현된 시설물은 출현 빈도가 높을수록 해당 공문의 민원 이슈와 연관성이 높다고 해석할 수 있다. ‘제3자피해’ 민원과 연관성이 높은 시설물 현황을 분석하고자 본 연구에서는 건설공사 사업관리방식 검토기준 및 업무수행지침(국토부 고시 제2023-153호)’에 명시된 중점관리시설 10가지 유형에 해당하는 시설물 553개를 목록화하였다. ‘제3자피해’ 관련 민원 공문을 대상으로 중점관리시설에 해당하는 시설물의 출현 빈도를 분석한 결과(그림 8), ‘제3자피해’ 민원 이슈와 연관성이 높은 중점관리시설 순위는 ‘급·배수시설’(17,630건), ‘인근의 도로’(14,461건), ‘건조물 또는 축사’(4,623건), ‘교통시설물’(3,325건), ‘지역편의시설’(2,435건) 등 순으로 나타났다.

Ⅴ. 결 론

본 연구에서는 국도 건설 현장의 실데이터를 분석하여 도로 건설공사의 민원 유형에 대한 분류체계를 정의하고, 분류체계에 따라 민원 문서를 자동으로 분류하는 모델을 개발하였다. 이를 위해 국도 건설사업에서 축적된 공문을 대상으로 민원 유형별 수동 레이블링 및 데이터 증강을 통해 30,439개의 학습 데이터셋을 구축하고 KoELECTRA 모델을 전이학습하였다. 또한, 학습된 모델을 이용하여 2015년부터 2021년까지 5개 지방청의 공문 37,926개를 민원 유형에 따라 분류하고, 분류된 데이터를 바탕으로 중점 민원 이슈에 대해 시범으로 분석하였다.

본 연구는 대량의 민원 문서를 자동으로 분류하고 데이터 기반 분석 접근 방법을 제시함으로써 민원 정보를 체계적으로 관리하고 활용성을 제고한 점에서 의의가 있다. 본 논문에서 제시한 민원 공문 분석의 관점은 통계적 수치 분석보다 제안한 모델을 통해 유의미한 정보를 생산할 수 있는가의 측면에서 유용성을 평가하는 데 초점을 두었다.

향후에는 분석 정보의 신뢰성을 높이기 위해서 다양한 실험을 통한 모델의 성능향상과 종합적인 분석 체계를 구축할 필요가 있다. 나아가 민원 관리 실무에서의 활용도를 높이기 위해서 문서의 민원 유형을 비롯하여 문서의 내용 측면에서 핵심 민원 정보를 추출 및 연계 분석할 수 있도록 민원 분석 서비스를 개발할 예정이다.

References

[1] J. H. Lee, J. B. Mun, C. J. Lee, and S. M. Yun, "A Study of Complaint and Conflict Factors in Construction Projects", Proceedings of the Korean Institute of Building Construction Conference, Vol. 21, No. 1, pp. 279-280, May 2021.

[2] S. J. Im, "Role Analysis of Construction Client in Industrial Accident Prevention using Text Mining", Doctoral dissertation, Chungbuk National University, Cheongju, Korea, 2020.

[3] H. Y. Kim, Y. E. Jang, H. B. Kang, J. W. Son, and J. S. Yi, "A Suggestion of the Direction of Construction Disaster Document Management through Text Data Classification Model based on Deep Learning", Korean Journal of Construction Engineering and Management, Vol. 22, No. 5, pp. 73-85, Sep. 2021. https://doi.org/10.6106/KJCEM.2021.22.5.073.

[4] Y. Kim, S. Bang, J. Sohn, and H. Kim, "Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers", Automation in Construction, Vol. 134, pp. 104061, Feb. 2022. https://doi.org/10.1016/j.autcon.2021.104061.

[5] J. Liu, H. Luo, W. Fang, and P. E. Love, "A contrastive learning framework for safety information extraction in construction", Advanced Engineering Informatics, Vol. 58, pp. 102194, Oct. 2023. https://doi.org/10.1016/j.aei.2023.102194.

[6] P. Jafari, M. A. Hattab, E. Mohamed, and S. AbouRizk, "Automated extraction and time-cost prediction of contractual reporting requirements in construction using natural language processing and simulation", Applied Sciences, Vol. 11, No. 13, pp. 6188, Jul. 2021. https://doi.org/10.3390/app11136188.

[7] T. Ko, H. D. Jeong, and G. Lee, "Natural language processing-driven model to extract contract change reasons and altered work items for advanced retrieval of change orders", Journal of Construction Engineering and Management, Vol. 147, No. 11, Nov. 2021. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002172.

[8] H. Lee, W. Lee, B. Jo, H. Lee, S. Oh, S. You, M. Nam, and H. Lee, "Research on ITB Contract Terms Classification Model for Risk Management in EPC Projects: Deep Learning-Based PLM Ensemble Techniques", KIPS Transactions on Software and Data Engineering, Vol. 12, No. 11, pp. 471-480, Nov. 2023. https://doi.org/10.3745/KTSDE.2023.12.11.471.

[9] D. Zhang, M. Li, D. Tian, L. Song, and Y. Shen, "Intelligent text recognition based on multi-feature channels network for construction quality control", Advanced Engineering Informatics, Vol. 53, pp. 101669, Aug. 2022. https://doi.org/10.1016/j.aei.2022.101669.

[10] K. Jeon, G. Lee, S. Yang, Y. Kim, and S. Suh, "Dynamic building defect categorization through enhanced unsupervised text classification with domain-specific corpus embedding methods", Automation in Construction, Vol. 157, pp. 105182, Jan. 2024. https://doi.org/10.1016/j.autcon.2023.105182.

[11] S. Baek, S. H. Han, S. Yun, J. Lim, and J. Nam, "A Study for Conflict in Public Construction Projects Based on Online News", Proceedings of the Korean Institute of Building Construction Conference, Vol. 21, No. 1, pp. 277-278, May 2021.

[12] J. Baik, S. Chung, and S. Chi, "Issue Identification of Overseas Construction Markets from News Articles Based on BERTopic", Journal of Construction Automation and Robotics, Vol. 2, No. 2, pp. 21-26, Jun. 2023. https://doi.org/10.55785/JCAR.2.2.21.

[13] J. Park, M. Cho, S. H. Eom, and S. K. Park, "Quantification of Schedule Delay Risk of Rain via Text Mining of a Construction Log", Journal of Civil and Environmental Engineering Research, Vol. 43, No. 1, pp. 109-117, Feb. 2023. https://doi.org/10.12652/Ksce.2023.43.1.0109.

[14] S. H. Eom, G. Cha, S. K. Park, S. Park, and J. Park, "Analysis of Potential Construction Risk Types in Formal Documents Using Text Mining", Journal of Civil and Environmental Engineering Research, Vol. 43, No. 1, pp. 91-98, Feb. 2023. https://doi.org/10.12652/Ksce.2023.43.1.0091.

[15] T. Chang, J. Kim, S. Chi, S. Im, and J. Seo, "Textual Analysis of Civil Complaint Data for Understanding User Experience and Satisfaction on Public Facilities", KSCE 2020 Convention, pp. 518-519, Oct. 2020.

[16] Ministry of the Interior and Safety, "Civil Petitions Treatment Act (Act No. 18748)", 2022.

[17] J. C. Nam, "The Analysis on Occurrence Factors and Proposal for a Classification System of Construction Civil Appeal", Master's thesis, The Graduate School of Chung-Ang University, Seoul, Korea, 2002.

[18] J. Lee, "A Study on the Risk Assessment of Civil Appeals in Construction Project", Master's thesis, The Graduate School of Semyung University, Jecheon, Korea, 2005.

[19] Ministry Of Construction & Transportation, "Manual for Handling Civil Complaints When Promoting Public Construction Projects", Apr. 2007.

[20] H. Jo, "Improvement of the Effectiveness in Public Grievances Management through Applying Construction Design Value Engineering", Master's thesis, The Graduate School Pusan National University, Busan, Korea, 2018.

[21] J. S. Won and J. Y. Shin, "AI Learning Data Construction Method for Automated Classification of Complaint Types at Road Construction Sites",

Proceedings of 2023 Summer Conference of Society for Computational Design and Engineering, Jeju, Korea, pp.232-234, Aug. 2023.

[22] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks", arXiv preprint arXiv:1901.11196., Jan. 2019. https://doi.org/10.48550/arXiv.1901.11196.

[23] Hugging face, koelectra-base-v3-discriminator, https://huggingface.co/monologg/koelectra-base-v3-discriminator [accessed: Apr. 25, 2024]

[24] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators", arXiv preprint arXiv:2003.10555, Mar. 2020. https://doi.org/10.48550/arXiv.2003.10555.

[25] National Institute of Korean Language, https://kli.korean.go.kr/corpus/main/requestMain.do?lang=ko [accessed: Apr. 25, 2024]

[27] J. Y. Shin and J. S. Won, "A Study on the Method of Deriving Complaint Request Information for Developing Information Service of Civil Complaint for Road Construction Sites", Proceedings of KIIT Conference, Jeju, Korea., pp. 213-216, Jun. 2023.

저자소개

신 재 영 (Jaeyoung Shin)

2015년 2월 : 한양대학교 실내건축디자인학과(이학사)
2017년 2월 : 한양대학교 실내건축디자인학과(이학석사)
2020년 9월 ~ 현재 : 연세대학교 실내건축학과 박사과정
2017년 3월 ~ 현재 : 한국건설기술연구원 미래스마트건설연구본부 전임연구원
관심분야 : 실내건축(설계, 실내동선), 인공지능, BIM(Building Information Modeling)

원 지 선 (Jisun Won)

2003년 2월 : 경희대학교 토목건축공학부 (공학사)
2005년 2월 : 경희대학교 건축공학과 (공학석사)
2024년 2월 : 경희대학교 건축학과 (박사수료)
2005년 12월 ~ 현재 : 한국건설기술연구원 미래스마트건설연구본부 수석연구원
관심분야 : 건설 데이터 표준, 인공지능, 자연어처리, BIM(Building Information Modeling)