**KSCE Journal of Civil and Environmental Engineering Research**
Vol. 43, No. 1: 91-98/ February, 2023
DOI: https://doi.org/10.12652/Ksce.2023.43.1.0091

ISSN 1015-6348 (Print)
ISSN 2799-9629 (Online)
www.kscejournal.or.kr

**Information Technology** **정보기술**

# 텍스트 마이닝을 통한 건설공사 공문 잠재적 리스크 유형 분석

엄세호* · 차기춘** · 박선규*** · 박승희**** · 박종호*****
Eom, Sae Ho*, Cha, Gichun**, Park, Sun Kyu***, Park, Seunghee****, Park, Jongho*****

# Analysis of Potential Construction Risk Types in Formal Documents Using Text Mining

**ABSTRACT**

Since risks occurring in construction projects can have a significant impact on schedules and costs, there have been many studies on this topic. However, risk analysis is often limited to only certain construction situations,and experience-dependent decision-making is therefore mainly performed. Data-based analyses have only been partially applied to safety and contract documents. Therefore, in this study, cluster analysis and a Word2Vec algorithm were applied to formal documents that contain important elements for contractors or clients. An initial classification of document content into six types was performed through cluster analysis, and 157 occurrence types were subdivided through application of the Word2Vec algorithm. The derived terms were re-classified into five categories and reviewed as to whether the terms could develop into potential construction risk factors. Identifying potential construction risk factors will be helpful as basic data for process management in the construction industry.

**Key words :** Text mining, Cluster analysis, Word2Vec, Potential risk, Formal document

**초 록**

건설프로젝트에서 발생되는 리스크는 공기지연 및 비용증가에 큰 영향을 끼치기 때문에 다양한 리스크를 파악하기 위한 노력이 이루어지고 있다. 그러나 시공단계의 리스크 분석은 공종 및 수행단계에 국한되거나, 경험 의존적 의사결정이 주로 수행되고 있다. 데이터 기반의 분석도 일부 사례에 적용되고 있을 뿐이다. 따라서 본 연구에서는 시공사 또는 발주처에 중요한 요인들이 포함되어 있을 것으로 판단되는 수발신공문을 대상으로 군집분석과 Word2Vec 알고리즘을 적용하였다. 군집분석을 통해 6개 유형으로 1차 분류를 수행하였으며, Word2Vec을 통해 157개의 공문 발생 유형을 도출하였다. 도출된 연관어의 속성별 분석을 위하여 새로운 5개의 범주를 적용하였으며, 이를 통해 공문 발생 유형이 잠재적인 건설 리스크 요인으로 발전 가능한지 검토하였다. 텍스트 마이닝을 통한 3단계의 공문 발생 유형 분석 결과는 건설현장의 공정관리를 위한 기초자료로써 도움 될 것으로 판단된다.

**검색어 :** 텍스트 마이닝, 군집분석, Word2Vec, 잠재 리스크, 공문

* 정회원 · 성균관대학교 글로벌스마트시티융합전공 석사과정 (Sungkyunkwan University · alalsp767@g.skku.edu)
** 정회원 · 성균관대학교 글로벌스마트시티융합전공 연구교수, 공학박사 (Sungkyunkwan University · ckckicun@skku.edu)
*** 종신회원 · 성균관대학교 건설환경공학부 교수, 공학박사 (Sungkyunkwan University · skpark@skku.edu)
**** 종신회원 · 성균관대학교 건설환경공학부 교수, 공학박사 (Sungkyunkwan University · shparkpc@skku.edu)
***** 종신회원 · 교신저자 · 성균관대학교 리질리언트에코스마트시티 연구교수
(**Corresponding Author** · Sungkyunkwan University · rhapsode@skku.edu)

**Received** October 19, 2022/ **revised** October 26, 2022/ **accepted** October 27, 2022

# 1. 서 론

## 1.1 연구의 배경 및 목적

건설현장에서는 다양하게 투입된 인력, 장비, 비용 등의 자원이 복잡하게 상호작용하여 원활한 공사 진행에 많은 영향을 끼치게 된다. 이러한 외부 환경 요인이 불확실성의 결과로서 건설프로젝트에 악영향을 주는 사건으로 발전하게 되면 이를 건설프로젝트 리스크로 정의할 수 있다(Al-Bahar, 1989). 건설프로젝트 리스크는 공기지연 및 비용증가에 큰 영향을 끼치기 때문에 현장에서의 리스크를 파악하기 위한 많은 노력이 이루어졌다.

Kang et al.(2002)은 건설공사 리스크 관리를 위한 분류체계를 도출하기 위하여 설문조사 및 분석을 통하여 정치·사회적, 기획, 입찰, 계약, 설계 및 시공단계 인자를 구분하였다. Kim et al.(2008)은 공종별 리스크 발생률이 전체 공종 중 26.1 %으로 높은 비율을 차지한 철근콘크리트 공사를 대상으로 기존 문헌 및 인터뷰 설문조사를 진행하고 AHP 기법을 이용하여 리스크 영향력을 분석하였다. Yoon et al.(2008)은 철근콘크리트 공종을 대상으로 AHP 기법 및 공정리스크 중요도 지수를 도입하여 공정리스크를 도출하고 관리 시스템을 제안하였다. Yang(2020) 및 Choi(2015)은 해외건설프로젝트 리스크 분류 체계를 실제 사례에 근거하여 도출하고 전문가 설문조사를 통하여 해외 공사에서의 건설 리스크를 도출하였다. Lee et al.(2018)에서는 공공 건설사업 내에서 공사비 증가, 공기 지연의 증가를 유발하는 갈등 요인에 대하여 갈등 유형을 분류하고 군집분석을 통해 유형을 분류하였다. 이처럼 분석 가능한 다양한 정보를 바탕으로 건설현장의 리스크를 파악하기 위한 노력이 있었지만, 일부 공종 및 수행단계에 국한되거나, 설문조사를 통한 실무자의 경험 의존적 의사결정을 기반으로 하고 있다.

최근 이러한 경험 의존적 의사결정에서 벗어나 데이터 기반의 객관적인 의사결정을 수행하고자 다양한 건설문서를 분석하는 텍스트 마이닝 기법이 주목받고 있다. 국내 사례에 대하여는 주로 안전사고와 관련된 텍스트 마이닝 연구가 수행되었다. Kang and Yi(2018)는 안전보건공단에서 제공하는 건설공사 재해사례 텍스트 데이터를 Word2Vec 알고리즘을 활용하여 시각화하였으며, Park and Kim(2021)은 한국산업안전보건공단의 텍스트 데이터를 웹크롤링하여 빈도분석과 중심성 분석을 통해 중요도를 산출하였다. Kim and Kim(2019)은 계절별로 건설현장에서 발생한 추락사고 인터넷 기사를 웹크롤링하여 주성분분석과 군집분석을 통하여 추락사고의 특징을 도출하였다. Yang and Lim(2021)은 국내 KCI 논문 등재지 및 학위논문을 바탕으로 건설 재해 연구 동향을 의미연결망을 통해 분석하였다. Shin and Chi(2014)은 미국 건설사고 리포트의 텍스트 마이닝을 통해 도출되는 단어 간의 연관성과 군집분석을 통해 텍스트에 포함된 경험지식의 정보화 가능성을 확인하였다. 안전사고 외에 Kim(2022a)은 해외건설프로젝트에 대하여 공기지연일과 공기지연요인의 분포를 4개로 구분하고 지연사유와 지연대책을 분석하여 사례를 나열하였다. Kim(2022b)는 아파트 건설산업 프로젝트 내에서의 정확한 공정관리를 위해 적극적인 작업분류체계 분류를 제안하였으며, 작업 상세내용 분석에 Word2Vec 등의 알고리즘을 활용하였다. 이외에 해외프로젝트의 계약문서를 텍스트 마이닝하여 입찰 리스크 및 사업 타당성을 분석할 수 있는 기반을 마련하였으며(Lee and Yi, 2017; Marzouk and Enaba, 2019; Son and Lee, 2019), 상하이에서 수행된 건설프로젝트의 감독일지를 분석하여 현재 리스크 관리에서 누락 가능성이 있는 고빈도-저심도(high-frequency-low-severity) 리스크를 정의하였다(Wang et al., 2020). 이처럼 텍스트 마이닝을 활용한 건설 리스크의 분석은 지금까지 소수의 전문가들이 범주화하거나, 비정형데이터로써 관리되던 건설 리스크 관련 데이터 산업을 발전시키고 있다(Kang and Yi, 2018). 다만, 비교적 대량의 데이터 수집이 용이한 건설현장 재해사고를 바탕으로 한 리스크 분석, 입찰단계에서 계약 문서 분석을 통한 사전 타당성 검토 등의 일부 사례에만 텍스트 마이닝 기술이 적용되고 있어 시공단계에서의 전반적인 리스크 도출에 한계가 있다.

건설현장에서 생성되는 건설프로젝트 문서 중 감독일지 및 공문 등은 시공단계에서 발생하는 다양한 현장 관리 요인이 포함되어 있으며, 이는 시공사 또는 발주처에 중요한 요인들이 포함되어 있을 확률이 상당히 높다. 그러나 민감한 공사정보로 인하여 정보 접근의 한계가 존재하며, 충분한 연구가 진행되지 못하였다. 따라서 본 연구에서는 국내 토목공사 현장의 공문을 수집하고 군집분석과 Word2Vec 알고리즘을 활용한 텍스트 마이닝을 수행하여 공문의 발생 유형을 살펴보고 잠재적인 건설 리스크 요인을 검토하고자 한다.

## 1.2 연구의 범위 및 방법

본 연구는 국내 토목공사 현장 10개 공구에서 시공사와 발주처 간의 수발신공문을 대상으로 텍스트 마이닝 기반 군집분석과 Word2Vec 알고리즘을 적용하였다. 수집된 총 파일은 12,227개이

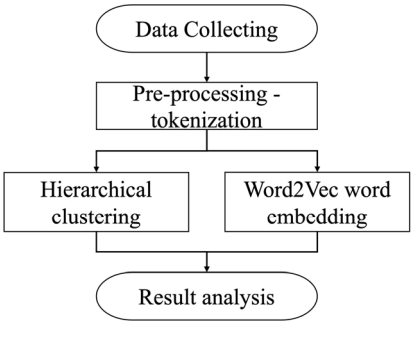



Fig. 1. Research Flow Chart

며, 분석 가능한 텍스트만을 추출하여 총 7,375개의 텍스트를 분석하였다. 공문 분석 체계를 Fig. 1에 나타내었다. 수집한 공문의 제목 및 내용에 대하여 형태소 분석 및 추출, 불용어 처리 등의 전처리를 수행한 후 군집분석을 수행하여 공문 발생 유형을 도출하였다. 이후 Word2Vec를 통해 공문의 제목을 학습시킨 후 도출된 군집을 기준으로 잠재 리스크 요인을 구성하였다.

## 2. 텍스트 마이닝

### 2.1 자료의 전처리

비정형 데이터인 텍스트는 문자, 단어, 구, 품사 등 다양한 요소로 구성되어 있기 때문에 문장으로부터 단어를 추출하고 문장부호 등을 분리할 수 있는 전처리 과정을 거쳐야 한다. 한글의 텍스트 마이닝은 형태소 분석을 통해 전처리를 수행한다. 다만 건설현장에서 발생된 건설문서의 경우 다양한 명사의 결합형태가 존재하며, 공문의 특성상 요약된 문장이 다수 존재하게 된다. 공문에 적합한 형태소 분석기를 채택하기 위하여 다양한 형태소 분석기의 비교가 필요하다(Kang and Yi, 2018). Fig. 2에 형태소 분석 절차를 나타냈다.

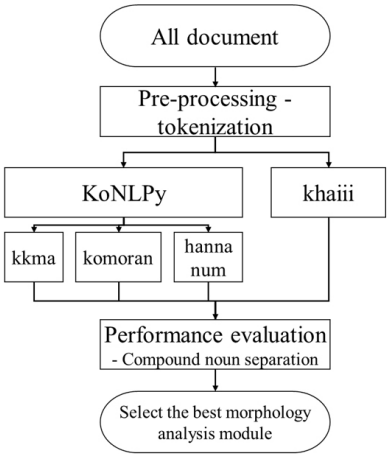



Fig. 2. Flow Chart of Selecting the Morphology Module

Table 1. Pre-processing Performance Comparison by Natural Language Processing Software

| KoNLPy | | | Kakao |
|---|---|---|---|
| kkma | komoran | hannanum | khaiii |
| a floating representation | kang (강) | a floating representation | above |
| land | ka (카) | is price fluctuations | last time |
| ready | geo | holder | significant |
| a related | lim (림) | and | work |
| holder | bu (부) | attaching | factory |
| and | a (아) | world | guidance |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 1955 pcs | 1057 pcs | 1580 pcs | 1036 pcs |

Table 1에 한글 데이터 분석에 사용되는 대표 라이브러리인 KoNLPy (Park and Cho, 2014)의 3가지 분석기와 카카오에서 공개한 khaiii (Kakao Hangul Analyzer Ⅲ) (kakao, 2020) 형태소 분석기를 비교하였다. 분석 대상 전체의 공문을 대상으로 형태소 분석 후 명사만 추출하였다. KoNLPy에서 제공하는 분석기들은 카카오의 khaiii 대비 도출시킨 명사의 개수가 많았으나, 공문에 요약된 문장형태를 잘 반영하지 못하여 불필요한 형태소로 분석하는 것으로 나타났다. 반면에 khaiii는 CNN (Convolution Neural Network) 알고리즘을 이용하여 형태소를 분석함으로써 비교적 공문의 명사를 정확하게 분석하고 불필요한 형태소는 잘 제외시킨 것으로 판단하여 본 연구에 적용하였다. 전처리 이후에는 복합명사로 인하여 명사가 분리된 건설 용어를 별도로 정의하였으며, 불용어를 제거하였다.

### 2.2 구조적 군집분석

군집분석은 유사한 데이터들을 서로 묶어주는 분석방법으로 구조적 군집분석은 텍스트 데이터를 트리 형태의 군집으로 나누어 데이터간의 거리가 가까운 대상들부터 결합하여 계층구조를 형성해 나갈 수 있다. 군집간의 거리를 측정하는 방법은 다양하며 본 연구에서는 군집내 증분과 군집 간 제곱합을 동시에 고려하는 와드법(Ward's method)을 사용하였다(Seo, 2019). 군집분석을 위하여 전처리가 완료된 단어를 대상으로 TF-IDF 빈도분석 방법을 우선 적용하였다. TF-IDF는 특정 문헌 내에서 단어 빈도가 높고, 전체 문헌에서 그 단어의 포함 문헌이 적을수록 값이 높아지는 특성을 가지고 있으며, 건설 리스크와 관계없이 반복적으로 발생하는 공문 내 단어를 효과적으로 제외시킬 수 있을 것으로 판단된다.

$$TF-IDF(t,d,D) = tf(t,d) \times idf(t,D) \tag{1}$$

$$idf(t,D) = \log\left(\frac{n}{1+df(t)}\right) \tag{2}$$

여기서 $tf(t,d)$는 특정 문서 $d$에서의 특정 단어 $t$의 등장 횟수이며, $D$는 전체 문서를 나타내며, $idf(t,D)$는 특정 단어 $t$가 등장한 문서의 수인 $df(t)$에 반비례하는 값이며, $n$은 총 문서의 개수이다.

### 2.3 Word2Vec

Word2Vec은 2013년 구글에서 발표된 연구로 단어의 의미는 그 단어 주변 분포로 이해될 수 있으며, 단어의 의미가 단어 벡터 안에 인코딩 될 수 있다는 2가지 가정을 통하여 중심단어로 주변 단어를 학습할 수 있는 워드 임베딩 방법으로, 최근 가장 많이

사용되는 모델이다(Mikolov et al., 2013). Word2Vec는 주변에 있는 단어들을 학습하여 중간에 있는 단어를 예측하는 CBOW (Continuous Bag of Words)방법과 중간 단어로 주변 단어를 예측하는 Skip-Gram 방식이 있으며, 본 연구에서는 군집분석에서 도출된 군집을 중심으로 공문에 분포된 단어를 예측하기 때문에 Skip-Gram 방식을 적용하였다. Word2Vec 적용 시에는 형태소 분석기를 적용하지 않은 상태로 학습을 진행하였다. 다만, 군집분석에서 정의하였던 복합명사를 통일하여 적용하고 공문에서 많이 사용되는 단어인 '무궁', '발전', '기원', '의거', '통보', '첨부' 등의 단어를 불용어 처리하는 과정을 적용하였다.

## 3. 텍스트 마이닝 기반 건설공사 공문 분석

### 3.1 구조적 군집분석 결과

Fig. 3에 한 개 공구에 대한 군집분석 결과를 나타내었다. 거리가 가까운 군집 간에 별도의 색상으로 구분되어 있으며, 건설현장의 구분이 가능한 일부 단어는 가렸다. 군집별 특징을 살펴보면 '자재(material), 신청서(application form), 지급(payment)', '처리(processing), 폐기물(waste)', '점검(inspection), 구조(structure), 검토(examination), 시행(implementation), 품질(quality), 환경(environment) 등', '현황(present status), 확인(confirmation), 관리(management), 안전(safe)', '계약(contract), 하도급(subcontracting)', '변경(change), 설계(designing), 승인(approval), 요청(request)'의 6가지 그룹으로 구분되었으며, 각 군집을 자재(construction material), 폐기물(waste), 점검/검토/조치(inspection/examination/act), 안전/관리(safe/management), 하도급(subcontracting), 설계변경(design change)으로 지정하였다.

Table 2. Clustering Results

| Cluster | Value |
|---|---|
| Design change | 10 |
| Inspection/Examination/Action | 10 |
| Construction material | 8 |
| Waste | 5 |
| Subcontracting | 4 |
| Safe/Management | 2 |

Table 2에 총 10개 공구에 대한 군집 및 도출된 횟수를 나타내었다. 각 공구에서 도출된 군집의 개수는 차이가 있었으나, A공구에서 나타난 6개 군집으로 구분 가능하였다. 설계변경과 점검/검토/조치와 관련한 사항은 모든 공구에서 도출되었으며, 자재, 폐기물, 하도급, 안전/관리의 순서로 군집이 분류되었다. 가장 적은 군집이 도출된 공구에서는 설계변경, 점검/검토/조치, 폐기물 등 3개 군집이 분석되었다.

### 3.2 Word2Vec 분석 결과

Word2Vec 알고리즘을 활용하여 각 공구별로 학습된 단어는 최소 443개에서 최대 1,997개, 평균 1495.7개의 단어가 학습되었다. 학습된 결과의 시각화를 위하여 구글 임베딩 프로젝터의 데이터

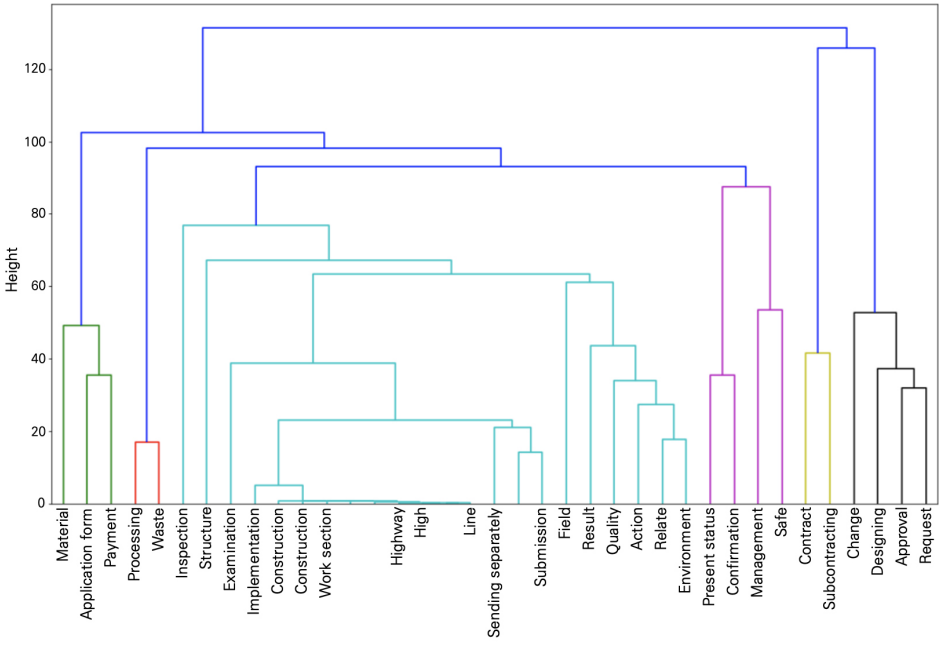
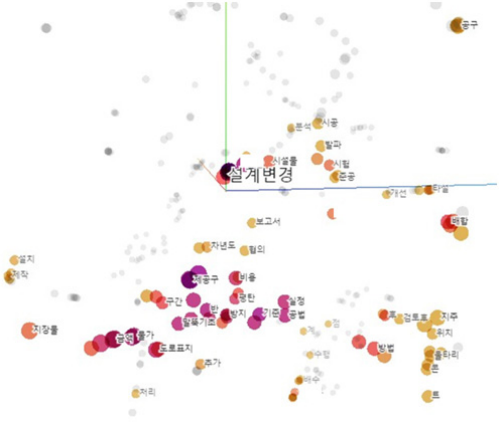
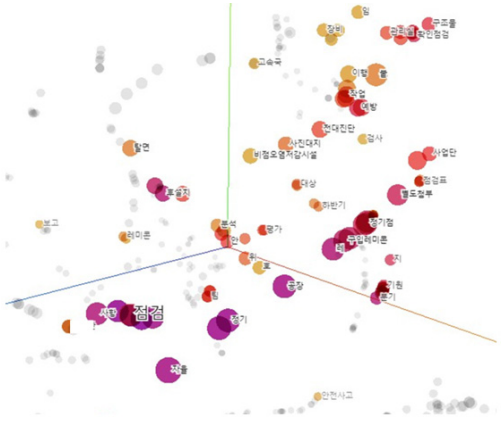



Fig. 3. Cluster Analysis of Formal Documents in Site A

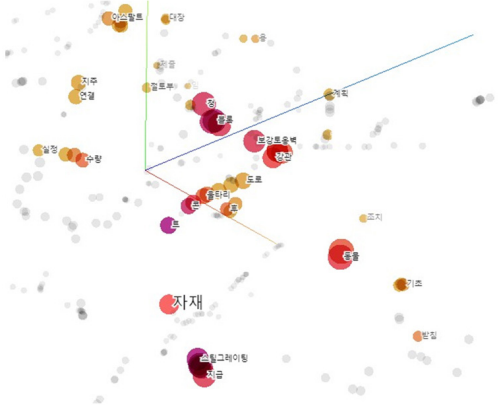
시각화 도구를 사용하였다(Smilkov et al., 2016). 시각화 방법으로는 t-SNE (Stochastic Neighbor Embedding)방법을 사용하였다. SNE 방법은 가우시안 분포를 확률분포로 전제하기 때문에 꼬리가 두텁지 않아 특정 개체에서 거리가 가까운 개체와 떨어져 있는 개체가 선택될 확률의 큰 차이가 발생하지 않는다. 따라서 가우시안 분포보다 꼬리가 두터운 t분포를 활용한 것이 t-SNE이며, Word2Vec 알고리즘에서 시각화하는데 많이 쓰인다(Kang and Yi, 2018). Fig. 4에 A공구의 설계변경, 점검/검토, 자재, 폐기물, 하도급, 안전과 연관어를 시각화하였으며, 군집분석과 동일하게 시공사를 유추할 수 있는 단어는 제외하여 나타내었다.



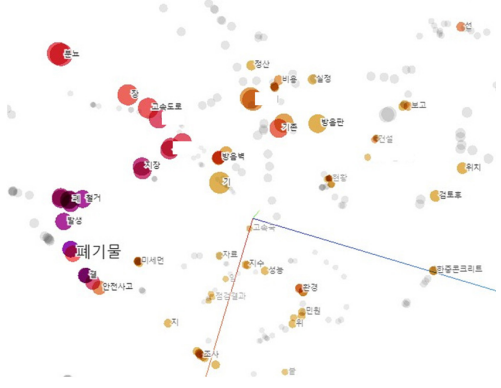

(a) Design change



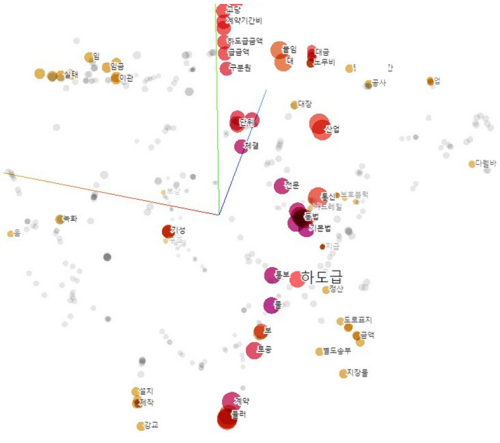

(b) Inspection



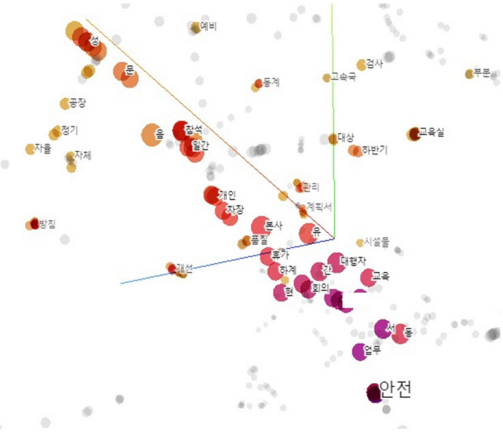

(c) Construction material




(d) Waste




(e) Subcontracting




(f) Safe

**Fig. 4.** Embedding Visualization of Word2Vec by t-SNE

설계변경에서 코사인 유사도를 기준으로 연관어를 분석한 결과 '반영(0.039)', '적용(0.055)', '승인(0.092) '등 설계변경이 수행된 경우, '기준(0.110)', '물가(0.139)' 등 외부 환경의 변화에 따라 설계변경이 필요한 경우에 관한 사례가 많은 것으로 유추된다. 점검과 관련한 단어는 '지적(0.056)', '결과(0.078)', '실시(0.142)' 등 점검 이후의 절차와 '자율품질(0.057)', '정기(0.098)', '분기(0.149)' 등 정기적이고 일반적인 절차 그리고 '공장(0.120)', '구입레미콘(0.189)' 등의 재료와 관련된 사례가 유추된다. 자재의 경우 특정 공종에 필요한 자재의 '물품명'과 '신청(0.146)' 등의 단어가 도출되었으며, 폐기물에서는 '처리(0.022)', '발생(0.094)', '의뢰(0.149)'등의 단어가 도출되었다. 하도급은 '기본법(0.138)', '의거(0.161)', '시행령(0.183)' 등 관련 기준 및 법령이 변경되었을 때에 대응 방안과 '체결(0.185)', '계약(0.207)', '대금(0.229)' 등 실제 계약을 위한 절차 사례가 많이 나타났다. 안전에서는 '관리자(0.055)', '안전관리자이탈계(0.133)', '대행자(0.211)' 등의 관리자의 업무 정보와 관련한 내용이 많이 도출되었다.

### 3.3 잠재적 건설 리스크 유형 분석

구조적 군집분석을 통해 6개의 군집으로 분류하였으며, 각 군집을 중심으로 Word2Vec를 활용한 워드 임베딩 학습 결과를 시각화로 표현하였다. 다만, 단순한 시각화의 경우 어떠한 양상을 보이는지 확인은 가능하지만 잠재적 건설 리스크 요인을 분석할 수 없는 한계점을 가진다. 또한 연관어는 단순히 중심단어 근처에 가까이 위치한 단어들로써 실제 리스크로써의 가치가 있는지 확인이 불가하다. 따라서 각 공구별 Word2Vec 학습결과에서 군집별로 상위 20개의 연관어를 별도로 추출한 후 일상적 현장관리 및 전처리 과정에서 거르지 못한 일부 단어들을 삭제하였다. 중복을 포함하여 총 279개의 연관어가 도출되었으며, 157개의 공문 발생 세부 유형으로 구분되었다. 이를 외부요인(external factor), 구조물/시설물(structure/facility), 재료(material), 민원(civil complaint), 장비(equipment) 등 5개의 범주로 구분하여 비교하였다. Fig. 5를 보면 구조물/시설물 관련은 34 %, 외부요인이 32 %, 재료가 31 %, 장비 및 민원 1 %를 차지하고 있으며, 재료의 경우 대부분 자재 관련 요인으로 구성되어 있는 것을 확인하였다. 외부요인에 따른 공문 유형은 안전 및 자재를 제외하고 고르게 분포되었으며, 구조물/시설물의 경우 설계변경이 전체의 63 %를 차지하는 것으로 나타났다.

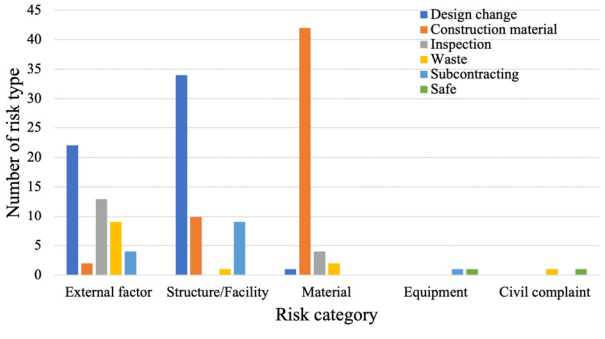



Fig. 5. Distribution Risk Types by Category

임의로 구분된 범주와 군집분석 유형에 따른 일부 연관어를 Tables 3~5까지 빈도 순으로 나타내었다. Word2Vec 분석결과를

Table 3. Potential Construction Risk of External Factors

| Category | First classification (clustering) | Second classification (Word2Vec) |
|---|---|---|
| External factor | Design change | Omission, Cost, Weak, Vegetation, Inhibition, Guide, Landscape, … |
| | Construction material | Production, Shock |
| | Inspection/ Examination/Action | Pointing out, Thawing season, Safe counterplan, Real condition, Factory, … |
| | Waste | Disturbance, Plants, Remove, Animal husbandry, … |
| | Subcontracting | Framework act, Laws and ordinances, Restoration, Enforcement ordinance |

Table 4. Potential Construction Risk of Structures

| Category | First classification (clustering) | Second classification (Word2Vec) |
|---|---|---|
| Structure/ Facility | Design change | Shoulder, Method of construction, Barrier, Cut slope, Format, … |
| | Construction material | Abutment, Road sign, Retaining wall, … |
| | Waste | Existing road |
| | Subcontracting | Girder, Structure construction, Soundproof wall, Earthwork, … |

Table 5. Potential Construction Risk of Materials, Equipment, and Complaints

| Category | First classification (clustering) | Second classification (Word2Vec) |
|---|---|---|
| Material | Design change | Interior material |
| | Construction material | Steel pipe, Reflector, Recycled aggregate, Guard fence, Lead fence, … |
| | Inspection/ Examination/Action | Hot weather concrete, Buy ready mixed concrete, Asphalt concrete, … |
| | Waste | Shotcrete, Waste concrete, … |
| Equipment | Safe/Management | Guarantee letter |
| | Subcontracting | Diamond grinding |
| Civil complaint | Safe/Management | Residual land |
| | Waste | Lot |

2차 분류로 구분하여 연관어를 분석하면 잠재적 건설 리스크 요인으로 발전 가능한 공문의 유형을 분석 가능할 것으로 판단된다. 외부요인에 따른 리스크 유형을 Table 3에 나타내었다. 설계변경의 경우 '누락(omission), 비용(cost), 취약(weak)' 등의 문제점 발견이나, '지침(guide)'등의 기준의 변경 또는 권고사항, '경관(landscape)' 등 구조물 또는 시설물의 형상 변경, '식생(vegetation), 억제(inhibition)' 등 공사 현장의 환경적 요인과 관련된 영향이 리스크로 발전 가능한 것으로 유추된다. 자재의 경우 '생산(production)'에 문제가 발생한 경우가 유추될 수 있으며, 점검/검토/조치의 경우 '지적(pointing out), 안전대책(safe counterplan), 실태(real condition)' 등 잘못된 현장관리에 따른 대처 마련, '공장(factory)' 등 현장 바깥의 원인 및 '해빙기(thawing season)' 등의 환경적 요인에 영향을 받는 것을 알 수 있다. 하도급의 경우 '기본법(framework act), 법령(laws and ordinances), 복원(restoration), 시행령(enforcement ordinance)' 등 건설 관련법에 많은 영향을 받았으며, 폐기물의 경우 연관어만을 가지고 뚜렷한 영향을 설명할 수 없었다.

Table 4에 구조물/시설물에 따른 리스크 유형, Table 5에는 재료, 장비, 민원에 따른 리스크 유형을 나타내었다. 구조물/시설물 범주의 경우 실제 구조물 및 공종 관련 연관어들이 도출되었다. 특히 폐기물의 경우 '기존도로(existing road)'라는 연관어가 도출되어 기존 구조물 또는 지장물에 대한 영향이 있음을 유추할 수 있다. 재료의 경우 실제 건설 현장에 쓰이는 재료 및 구조물의 부속품들에 관한 내용이 많이 도출 된 것을 확인할 수 있으며, 콘크리트 관련 연관어가 도출되어 현장관리에 있어 콘크리트의 품질관리 및 폐기물 처리 방안에 대한 영향을 많이 받는 것으로 나타났다. 장비 범주에서는 사용 장비의 안전을 확인하는 '보증서(guarantee letter)'확인과 같은 절차가 수행됨을 알 수 있었으며, 민원에서는 '잔여지(residual land)', '용지(lot)' 등 토지 사용과 관련된 민원이 제기되고 있음을 확인하였다.

## 4. 결 론

본 연구에서는 건설현장에서 생성되는 건설프로젝트 문서 중 시공사와 발주처간의 수발신공문을 대상으로 텍스트 마이닝 기반의 군집분석과 Word2Vec 알고리즘을 적용하여 공문 발생 유형을 도출하고 이를 잠재적인 건설 리스크 유형관점에서 분석하였다.

10개의 공구를 대상으로 군집분석을 수행한 결과 가장 적은 군집이 도출된 경우는 3개, 가장 많은 군집이 도출된 경우는 6개였다. 각 군집은 빈도 순으로 설계변경, 점검/검토/조치, 자재, 폐기물, 하도급, 안전/관리로 분류되었다. 군집분석만으로는 공문 발생 유형에 대한 설명이 불가능하여 Word2Vec 알고리즘을 활용하여 공문의 내용을 학습하였으며, 군집분석 1차 분류 유형을 기준으로 연관어를 도출하였다. 중복 연관어를 고려하여 157개의 잠재적 리스크 세부 유형으로 분류하였으며, 이를 외부요인, 구조물/시설물, 재료, 민원, 장비 등 5개의 리스크 범주로 구분하여 Tables 3~5까지 나타내었다.

단순한 텍스트 마이닝의 적용은 분석자의 주관적 판단이 많이 반영될 수 있기에 군집분석과 Word2Vec 알고리즘을 동시에 적용하여 공문 발생 유형을 분류하였다. 건설현장에서 발생 가능한 공문 유형을 도출하고 이를 주제어로 하여 연관어간의 유사점을 찾아냄으로써 건설현장에 발생하는 다양한 유형의 건설 관리 시나리오를 유추하였다. 유추된 시나리오는 건설프로젝트에 악영향을 끼칠 수 있는 잠재적인 리스크 발생 상황으로 간주될 수 있을 것으로 판단되며, 이를 미리 분석함으로써 공정관리를 위한 기초자료로써 도움 될 가능성이 높을 것으로 판단된다. 다만, 공문의 발생 조건이 무조건 건설 리스크로 발전하는 것이 아니기에 향후 잠재적 리스크 요인의 긍정/부정을 판단할 수 있는 추가 연구가 진행되어야 할 것으로 판단된다.

## 감사의 글

이 연구는 국토교통부/국토교통과학기술진흥원이 시행하고 한국도로공사가 총괄하는 "스마트건설기술개발 국가R&D사업(과제번호 22SMIP-A158708-03)"의 지원으로 수행하였습니다.

본 논문은 2022 CONVENTION 논문을 수정·보완하여 작성되었습니다.

## References

Al-Bahar, J. F. (1989). *Risk management in construction projects: A systemic analytical approach for contractors*, Ph.D. Dissertation, University of California Berkeley, Berkeley, California, USA.

Choi, J. W. (2015). *An analysis on regional differences of major delay factors in overseas architectural projects*, Master Dissertation, Hanyang University, Seoul, Korea (in Korean).

kakao (2020). *Kakao hangul analyzer Ⅲ*, Available at: https://github.com/kakao/khaiii (Accessed: October 18, 2022).

Kang, H. B. and Yi, J. S. (2018). "An analysis of public text data in construction disaster cases using Word2Vec-based data visualization." *Autumn Annual Conference of AIK, 2018, Architectural Institute of Korea*, Vol. 38, No. 2, pp. 567-570 (in Korean).

Kang, L. S., Kim, C. H. and Kwak, J. M. (2002). "Analysis for the importance of risk factors through the project life cycle." *Journal of the Architectural Institute of Korea Structure & Construction*, Vol. 17, No. 8, pp. 103-110 (in Korean).

Kim, J. S. (2022a). *Analysis of project delay using big data*, Master Dissertation, Hanyang University, Seoul, Korea (in Korean).

Kim, E. H. (2022b). *Automatic classification on the work breakdown structure of apartment construction projects: A machine learning*

*approach*, Master Dissertation, Sungkyunkwan University, Seoul, Korea (in Korean).

Kim, J. S. and Kim, B. S. (2019). "Characteristics analysis of seasonal construction site fall accident using text mining." *Korean Journal of Construction Engineering and Management*, Vol. 20, No. 3, pp. 113-121 (in Korean).

Kim, K. H., Kim, K. H., Lee, Y. S. and Kim, J. J. (2008). "A study about influence of risk factors in relation to construction cost increase and schedule delay on the reinforced concrete construction." *Journal of the Architectural Institute of Korea Structure & Construction*, Vol. 24, No. 5, pp. 165-172 (in Korean).

Lee, J. H. and Yi, J. S. (2017). "Predicting project's uncertainty risk in the bidding process by integrating unstructured text data and structured numerical data using text mining." *Applied Sciences*, Vol. 7, No. 11, pp. 1-15.

Lee, J. S., Kim, D. Y., Lee, C. J., Lee, J. H. and Han, S. H. (2018). "A research for clustering of conflict in public construction project." *Korean Journal of Construction Engineering and Management*, Vol. 19, No. 2, pp. 61-72 (in Korean).

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

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). *Efficient estimation of word representations in vector space*, Available at: https://arxiv.org/abs/1301.3781 (Accessed: October 18, 2022).

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

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

Seo, D. H. (2019). *Text mining with python*, bjpublic, Seoul, Korea (in Korean).

Shin, Y. J. and Chi, S. H. (2014). "Tacit knowledge informatization from text-based construction data." *Annual Conference of KICEM, 2014, Korean Journal of Construction Engineering and Management*, pp. 31-34 (in Korean).

Smilkov, D., Thorat, N., Nicholson, C., Rief, E., Viegas, F. and Wattenberg, M. (2016). *Embedding projector: Interactive visualization and interpretation of embeddings*, Available at: https://arxiv.org/abs/1611.05469 (Accessed: October 18, 2022).

Son, B. Y. and Lee, E. B. (2019). "Using text mining to estimate schedule delay risk of 13 oshore oil and gas EPC case studies during the bidding process." *Energies*, Vol. 12, No. 10, pp. 1-25.

Wang, G., Liu, M., Cao, D. and Tan D. (2020). "Identifying high-frequency-low-severity construction safety risks: An empirical study based on official supervision reports in Shanghai." *Engineering, Construction and Architectural Management*, Vol. 29, No. 2, pp. 940-960.

Yang, H. S. (2020). *Comparison of recognition on the risks affecting schedule delays and cost overruns in overseas civil construction projects*, Master Dissertation, Hanyang University, Seoul, Korea (in Korean).

Yang, S. W. and Lim, H. C. (2021). "Semantic network analysis on the research trends of construction accident." *Journal of the Architectural Institute of Korea*, Vol. 37, No. 6, pp. 231-236 (in Korean).

Yoon, Y. S., Suh, S. W., Park, M. S. and Jang, M. H. (2008). "Construction process based schedule risk management system." *Construction Engineering and Management*, Vol. 9, No. 4, pp. 101-110 (in Korean).