건설 현장 보행자의 안전 향상을 위한 텍스트 마이닝 기반 사고 데이터 분석

Enhancing Pedestrian Safety at Construction Sites: A Text Mining Approach to Extracting Accident Data

Abstract : Despite the significant risks to pedestrian safety posed by construction sites, the issue remains poorly managed due to the lack of sufficient statistical data. In response, this study proposed a method to extract statistical items on pedestrian accidents from news articles using a text mining approach. The proposed method was applied to analyze a total of 33 collected news articles about pedestrian accidents. The results showed a high accuracy rate of 93.9% in extracting location data. In classifying the type of accidents, which included slips and trips, falling objects, caught under objects, and inconvenience, the category of falling objects demonstrated high accuracy across all three morphological analyzers (okt, kkma, and soynlp). This research demonstrates the feasibility of collecting statistical data on pedestrian accidents near construction sites, which has not been previously compiled, by using news articles. This approach is expected to serve as valuable foundational data for enhancing pedestrian safety and formulating accident prevention policies.

Keywords : Pedestrian Safety, Construction Sites, Text Mining, Accidents

1. 서론

1.1 연구의 배경 및 목적

1.1.1 연구의 배경

최근 건설현장 주변 안전시설 미흡으로 인한 보행자 사고가 빈번하게 발생하고 있다. 2023년 5월 용인에서는 도보에 방치된 철판에 걸려 넘어진 보행자의 발목이 골절되는 사고가 발생하였으며, 2024년 1월 세종에서는 콘크리트 양생을 위해 천막을 덮어둔 집수정으로 보행자가 실족하는 사고가 발생하였다. 공중이용시설의 관리상의 결함을 원인으로 볼 수 있는 건설현장 보행자 안전사고는 중대재해 처벌 등에 관한 법률에서 정의하는 중대 시민 재해 중 하나임에도 불구하고, 피해자가 건설 근로자일 때를 의미하는 중대 산업재해와 비교하여 상대적으로 소홀하게 관리되고 있다. 이러한 원인은 첫째, 건설 현장 근로자의 사상 건수가 건설 현장 인근을 지나는 보행자의 사상 건수보다 월등히 많기 때문이다. 둘째, 보행자 사고는 책임소재가 다소 명확하지 않아 공사 주체와 관할 지자체가 서로 안전관리 책임을 미루는 관계로 대응 방안 마련에 어려움이 있기 때문이다. 셋째, 현재 건설현장 주변의 보행자 안전사고는 관련 통계조차 없는 상황으로 그 중요성이 간과되고 있기 때문이다.

본 연구는 건설현장 주변의 보행자 안전관리가 소홀한 여러 원인 중 마지막 항목을 해결하는 방안으로, 텍스트 마이닝을 사용하여 보행자 사고 관련 뉴스를 통계 데이터로 변환하는 방법을 제안하는 것을 목적으로 한다.

1.2 연구의 범위 및 방법

본 연구의 단계별 연구내용은 다음과 같다. 우선, 현재 건설작업자를 대상으로 수집되고 있는 사고 사례 데이터 현황을 검토하여 보행자 사고 통계 작성을 위해 필요한 항목들을 도출한다. 그 후 텍스트 마이닝 관련 선행연구 고찰을 통하여 뉴스 텍스트를 통계 항목으로 추출하는 방법을 수립한다. 연구 방법 개발이 완료되면, 실제 적용을 위하여 건설현장 주변 보행자 사고 관련 뉴스 기사를 수집하고 텍스트 마이닝을 수행하여 통계 항목을 추출한다. 마지막으로, 텍스트 마이닝을 위해 사용한 분석기의 성능을 평가하고 향후 활용방안을 제안한다.

* Corresponding author: Lee, Seulbi, Division of Architecture & Urban Design, Incheon National University, South Korea
E-mail: sblee@inu.ac.kr
Received May 15, 2024: revised ,
accepted ,

2. 이론적 고찰

2.1 건설공사 안전관리 종합정보망

국토교통부의 건설공사 안전관리 종합정보망[1]은 건설현장에서 반복적으로 발생하는 사고에 대하여 근본적인 원인을 규명하고 재발 방지 대책을 수립하기 위하여 사고 사례를 기록해야 한다는 취지에서 시작되었다. 사고 접수가 편리하도록 개발된 온라인 시스템과 더불어 관련법을 통해 신고 누락 시 과태료를 부과하고 있으므로 비교적 잘 관리되고 있는 상황으로, 2023년을 기준으로 235명의 사망자와 7,215명의 부상자가 보고되었다.

<Table 1>은 현재 건설공사 안전관리 종합정보망에서 건설사고 사례 통계 작성을 위해 수집하고 있는 주요 항목들과 그 예시를 정리한 것이다. 주요 항목으로는 사고명, 발생일시, 발생지역, 사고 유형, 사고 원인, 사망자 수, 부상자 수, 피해금액, 사고 발생 후 조치사항, 재발 방지 대책 등이 있다. 이때 사고명은 사고 내용을 한눈에 파악할 수 있도록 현장명과 사고 유형의 조합(예: 구미 공장 리모델링 사업 철골공사 안전망 설치 중 추락사고)으로 작성할 것이 권장된다. 사고 유형은 객관식 항목으로 떨어짐, 넘어짐, 물체에 맞음, 깔림, 끼임, 베임, 찔림, 부딪힘 등 중 하나로 선택해야 한다. 사고 원인의 경우 시공계획 미준수, 기계 장비 관리 미흡, 안전 수칙 미준수, 작업자의 불안전한 행동 등 중에서 하나를 선택한 후 구체적인 원인은 주관식으로 작성해야 한다.

Table 1. Items for statistics on construction worker accidents

Item	Example
Accident name	Fall accident during steel frame safety net installation in Gumi-si factory remodeling
Date and time	2023-10-23 10:35 a.m.
Location	Gyeongsangbuk-do, Gumi-si
Accident type	Fall
Cause of the accident	Unsafe behavior of workers (Failure to wear safety harness)
Number of fatalities	1
Number of injuries	0
Damage cost	N/A
Follow-up actions	Reporting to fire departments and transferring to hospital
Preventive measures	Enhancing safety training and monitoring

2.2 텍스트 마이닝

텍스트 마이닝은 구조화되어있지 않은 대규모의 텍스트에서 새로운 지식을 추출하는 분석 기법을 의미한다(Hotho et al. 2005). 건설 분야에서 텍스트 마이닝은 입찰서류, 계약서, 공사일지 등 다양한 단계에서 생성되는 문서를 분석함으로써 잠재적인 위험 요인을 식별하거나(Khalef and El-adaway 2021), 소셜 미디어나 뉴스 기사를 분석함으로써 건설 정책에 대한 주민의 의견을 수렴하는 연구 (Kinawy et al. 2017) 등이 있다.

Goh and Ubeynarayana (2017)의 연구에서는, 미국의 산업안전보건청(OSHA)에서 건설 사고 사례 보고서 중 사고 상황을 설명하는 서술형 문항을 텍스트 마이닝으로 분석하여 사고 원인(예: 물체 사이에 끼임, 붕괴, 추락, 낙하물에 부딪힘 등)을 추출하는 모델을 개발하였다. 해당 연구는 비교적 우수한 분류 성능으로 텍스트 마이닝의 유용성을 입증하였다는 기여가 있으나, 이미 각 항목이 잘 정리되어있는 사고 사례 보고서를 입력값으로 사용하였다는 점에서 개선의 여지가 있다. 또한, 한국어와 영어의 구조적 차이를 고려할 때, 한국어에 적합한 텍스트 마이닝 모델이 필요하다.

3. 텍스트 마이닝 모델 개발

본 연구의 텍스트 마이닝 모델 개발 순서는 다음 <Fig. 1>과 같이 정리된다. 가장 먼저 텍스트 데이터의 수집 단계에서는 건설현장 주변에서 발생한 보행자 사고에 관한 뉴스 기사를 수집한다. 그 후 전처리 단계에서는 다양한 한국어 형태소 분석기를 사용하여 뉴스 기사에서 수집된 말뭉치들을 형태소 단위로 나눈다. 또한, 통계 정보 추출 전 유의미한 형태소만을 남겨두기 위하여, 각 형태소 분석기에 적합한 방법으로 불용어 처리를 수행한다. 전처리 단계가 완료된 이후에는 보행자 사고 통계 작성을 위한 항목들을 추출한다. 추출 대상 항목은 건설공사 안전관리 종합정보망에서 수집하고 있는 통계 항목 중 지역, 사고 유형, 사고 원인으로 선정한다. 객관식 항목으로 정리가 가능한 지역과 사고 유형은 사용자 사전을 기반으로 추출하는 방법을 시도한다. 주관식 항목으로 작성되어 작성자의 의견이 반영될 수 있는 사고 원인의 경우 사고 유형에서 정의된 객관식 항목 단어들과의 연속적으로 사용되는 단어를 찾기 위한 N-grams 방법을 사용한다.

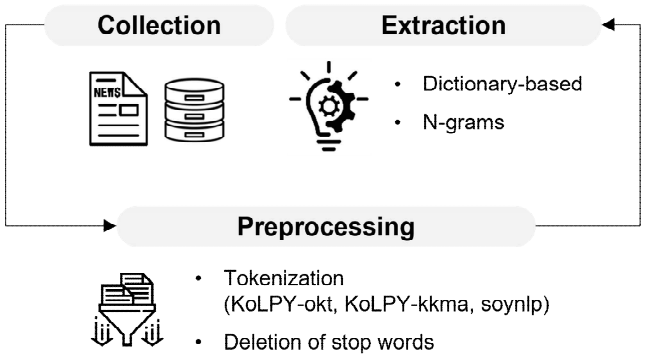

Fig. 1. Text mining process in this research

1) https://www.csi.go.kr

3.1 데이터 수집

본 연구에서는 건설현장 주변에서 발생한 보행자 사고에 관한 뉴스 기사를 수집하기 위하여, 구글과 같은 검색 엔진에 “건설현장”, “보행자”, “사고” 등을 키워드로 뉴스 기사를 수집하였다. 데이터의 품질을 높이기 위하여, 수집된 기사를 직접 읽어 정책이나 행정 조치에 관한 기사가 아닌 실제 사고 또는 통행 불편과 관련된 보도 기사만을 추출하였으며, 연구 결과의 편중을 방지하기 위하여 각 사고에 대하여 하나의 대표 기사만을 수집하였다. 최종적으로 2022년 1월 1일부터 2023년 12월 31일까지 발생한 총 33건의 보행자 사고에 관한 뉴스 기사가 수집되었다.

3.2 데이터 전처리

수집된 텍스트에서 유의미한 정보를 추출하기 위해서는 여러 문장으로 이루어진 말뭉치(Corpus)를 한국어의 가장 작은 단위인 형태소로 토큰화(Tokenization) 하는 과정이 선행되어야 한다. 본 연구에서는 가장 먼저 대표적인 한국어 처리를 위한 파이썬 패키지인 KoNLPY(박은정, 조성준 2014)를 사용하여 뉴스 기사를 형태소 단위로 분류하였다. 다양한 분석기 간의 성능을 비교하기 위해 KoNLPY의 분석기 중 빠른 처리 속도를 특징으로 하는 오픈소스 기반의 Okt(Open Korean Text) 분석기와, 높은 정확성을 특징으로 하는 KKMA 분석기를 각각 사용하였다.

KoNLPY는 한국어 형태소 분석을 위하여 다양한 분야에서 활발하게 적용되고 있으나, 일반적인 단어들로 학습된 모델이므로 새로 등장한 신조어나 전문용어의 경우 추출 성능이 떨어질 수 있다는 한계가 있다. 이를 보완하기 위하여 학습 데이터를 직접 지정할 수 있는 soynlp[2] 분석기를 추가로 사용하였다. 토큰화를 위한 학습 데이터로는 2023년 12월 1일부터 2024년 4월 30일까지 6개월간 건설공사 안전관리 종합정보망에 접수된 건설 작업자 사고 사례 총 998건에 대한 텍스트 자료를 사용하였다. 구체적으로 <Table 1>에서 설명하였던 통계 항목 중에서 사고명, 사고 유형, 사고 원인, 사고 발생 후 조치사항에 해당하는 5개의 텍스트를 학습 데이터로 사용하였으며, 학습 데이터에서 추출된 13,883개의 형태소 중 유효한 4,039개가 뉴스 기사의 토큰화에 사용되었다.

<Fig. 2>는 학습 데이터 중 사고 유형의 빈도를 막대그래프로 나타낸 것이다. 가장 빈번하게 발생한 사고 유형은 289건의 넘어짐이며, 떨어짐(작업자), 떨어지는 물체에 맞음, 물체와 부딪힘, 끼임, 깔림의 순서대로 사고가 자주 발생하였다. 188건으로 분류되는 기타 항목은 교통사고, 절단, 베임, 찔림, 화상, 질병 등을 포함한다.

2) https://github.com/lovit/soynlp

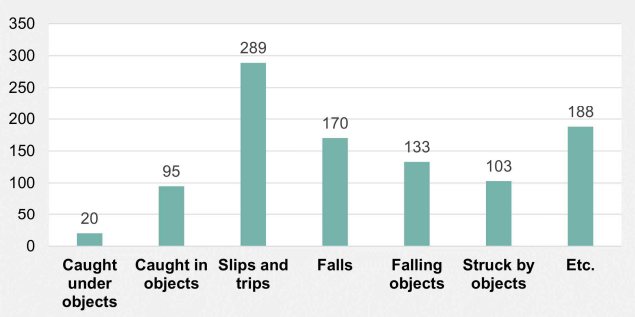

Fig. 2. Distribution of accident types in the training data

뉴스 기사의 토큰화 이후, 마지막 전처리 단계로써 특징 추출에 불필요한 조사, 접미사, 접속사 등을 제거하는 불용어 처리를 수행하였다. KoNLPY의 두 가지 분석기의 경우, 한국어 불용어 사전[3]을 참고하여 총 728개의 불용어를 제거하였다. 반면 soynlp의 경우, 학습 데이터에서 특정 형태소에 대하여 다음 형태소로 등장할 형태소의 빈도 확률을 나타내는 응집확률(Cohesion probability)이 0.1 이상이 되는 형태소를 제거함으로써 불용어를 처리하였다.

3.3 통계 항목 추출

1) 지역

본 연구에서 사용한 세 가지 종류의 분석기는 지역과 관련된 단어를 별도의 사전으로 포함하고 있지 않기 때문에, 건설현장 주변에서 발생한 보행자 사고의 지역을 특정하기 위하여 웹 크롤링을 위한 파이썬 패키지인 BeautifulSoup을 사용하여 지역 관련 단어 사전을 정의하였다. 크롤링 대상 웹 페이지로는 구글에서 “대한민국의 행정 구역” 검색 시 최상단에 등장하는 위키피디아[4]를 사용하였다. 위키피디아의 구조를 참고하여 생성할 데이터 프레임의 열은 지역, 구, 군, 시로 지정하였으며, 지역-구, 지역-군, 지역-시-구 세 가지의 조합 중 하나에 포함될 수 있도록 코드를 작성하였다. 최종적으로 토큰화된 형태소 중 지역, 구, 군, 시 중 어느 하나라도 식별되는 경우 현재 건설공사 안전관리 종합정보망에서 사용하고 있는 통계 분류 기준과 같이 지역 단위(1특별시, 6광역시, 6도, 3특별자치도)로 추출되도록 구성하였다.

2) 사고 유형

건설현장 주변 보행자 사고의 특성과, <Fig. 2>에서 살펴본 학습 데이터의 사고 유형 분포를 종합적으로 고려하여, 보행자 사고의 대표적 유형은 넘어짐, 떨어지는 물체에 맞음, 깔림의 세 가지를 선정하였다. 해당 사고 유형에 대한 사용자

3) https://gist.github.com/spikeekips/40eea22ef4a89f629abd87eed535ac6a
4) https://ko.wikipedia.org/wiki/대한민국의_행정_구역

사전은 국립국어원 표준국어대사전의 인접 어휘를 사용하여 <Table 2>와 같이 적용하였다. 구체적으로, 넘어짐과 관련된 단어 사전에는 '넘어'와 '걸려'가 포함된다. 떨어지는 물체에 맞음과 관련된 단어 사전에는 '떨어', '맞아', '맞았', '낙하', '쏟아'가 포함된다. 깔림과 관련된 단어 사전에는 '깔', '덮', '기울', '무너', '붕괴'가 포함된다. 그 외에도, 수집된 뉴스 기사 중에는 사고는 발생하지 않았으나, 건설현장으로 인해 통행 불편을 겪는 상황을 보도하는 사례가 포함되어있다. 이러한 뉴스 기사를 별도로 분류하기 위해 불편함을 사고 유형으로 추가하였다. 불편함을 추출하기 위해 사용된 어휘는 일반적으로 건설현장 보행자 사고에서 자주 등장하는 단어인 '불편', '불안', '점용'을 사용하였다.

Table 2. Korean Word Lexicon for Different Accident Types

Accident types	Keyword (Korean words)
Slips and trips	넘어, 걸려
Falling objects	떨어, 맞아, 맞았, 낙하, 쏟아
Caught under objects	깔, 덮, 기울, 무너, 붕괴
Inconvenience	불편, 불안, 점용

3) 사고 원인

구체적인 사고 원인의 경우 주관식으로 작성되는 항목이므로, 지역이나 사고 유형과 달리 사전을 기반으로 특정하기에는 어려움이 있다. 따라서 사고 원인을 추출하는 방법으로는 텍스트 마이닝 기법 중 단어의 순서를 기반으로 특정 단어의 집합을 추출하는 N-grams를 사용한다(Conway et al. 2009). N-grams에서 N은 숫자를 의미하는 것으로, 만약 3-gram이라면 말뭉치에서 토큰화된 형태소 중에서 연속하여 등장하는 3개의 단어를 집합으로 추출하는 방식으로 작동한다. 즉, 본 연구에서는 N-grams를 사용하여 <Table 2>에서 정의한 사고 유형별 관련 어휘들과 함께 등장하는 단어를 추출함으로써 사고의 원인 예측을 시도하였다.

4. 분석 결과

4.1 형태소 빈도 분석

<Table 3>은 세 가지 한국어 형태소 분석기를 사용하여 총 33건의 건설현장 주변 보행자 사고 뉴스 기사를 분석하였을 때 가장 많이 추출된 상위 20개의 형태소를 순서대로 정리한 결과이다. 표를 통해 KoNLPY의 두 가지 분석기의 경우 공통으로 '현장'이라는 단어를 가장 많이 추출하였음을 확인할 수 있다(okt: 151회, kkma: 158회). 반면, soynlp의 경우 118회로 기록된 '공사'라는 단어가 가장 많이 등장한 것으로 분석되었다. 분석기마다 단어의 빈도에는 다소 차이가 있지만, 세 가지 분석기 모두에서 빈번하게 출력된 상위 4개 단어는 '현장', '사고', '안전', '공사'로 일치하였다.

각 분석기의 특징을 살펴보면 okt 분석기의 경우 '사장'이라는 단어가 40회로 9위를 기록하였는데, 이는 '공사장'이 '공'과 '사장'으로 잘못 나누어진 결과로 추측할 수 있다. 또한 다른 분석기와 달리 okt 분석기에서는 떨어지는 물체에 맞음을 설명하기 위한 말뭉치를 kkma 분석기의 '떨어지'또는 soynlp의 '떨어'처럼 추출해내지 못하고 '떨어져', '떨어진', '떨어졌다' 등 비교적 일관성 없는 단어로 추출하였다.

KoNLPY의 또 다른 분석기인 kkma 분석기의 경우, okt 분석기와 비교하면 단어의 잘못된 분리나 의미 없는 단어의 추출과 같은 오류가 적은 편이었다. 그러나, 일반적으로 불용어 사전에 포함되지 않는 단어인 '한'이나 '면'과 같은 단어도 추출하는 부분은 보완이 필요하였다.

마지막으로 건설공사 안전관리 종합정보망에 접수된 건설작업자 사고 사례를 학습하여 단어 추출에 사용한 soynlp 분석기의 경우, KoNLPY의 학습 데이터와 비교하여 현저히 적은 단어를 학습했음에도 불구하고 유사한 성능을 보여주었다. 그뿐만 아니라 별도로 불용어 사전을 정의할 필요 없이, 응집확률이라는 지표를 사용하여 학습 데이터에서 출현 빈도가 낮은 단어들을 제외하는 방법을 사용하고 있어 okt 분석기에서 5위를 차지한 '했다'나 kkma 분석기에서 6위를 차지한 '한'같은 단어가 추출되지 않는다는 장점이 있다. 그러나 학습 데이터로써 건설 작업자 사고 사례를 사용하였기 때문에 다른 분석기와 달리 '피해'나 '시민'과 같은 단어는 추출되지 않았다. 또 다른 특징으로는 다른 두 분석기가 '공사'와 '현장'을 모두 분리한 것과는 달리 soynlp의 경우 일부 '공사현장'으로 추출되었다는 점이 있다.

Table 3. Top 20 most frequent Korean words

Rank	okt		kkma		soynlp	
	Word	Count	Word	Count	Word	Count
1	현장	151	현장	158	공사	118
2	사고	130	공사	124	사고	118
3	안전	115	사고	117	현장	109
4	공사	99	안전	103	안전	108
5	했다	92	건설	59	건설	55
6	건설	64	한	45	보행자	48
7	한	57	보행자	45	것으로	34
8	보행자	47	공사장	39	발생	32
9	사장	40	인도	36	인근	32
10	인도	37	도로	34	에서	29
11	발생	32	발생	32	떨어	28
12	인근	32	면	32	차량	27
13	하는	29	인근	32	으로	26
14	건물	29	건물	29	공사현장	26
15	해	29	떨어지	28	도로	25
16	차량	28	차량	28	작업	21
17	하는	28	지나	28	설치	20
18	피해	28	관계자	28	위험	18
19	관계자	28	피해	27	관리	18
20	시	27	시민	26	아파트	17

4.2 지역

<Table 4>는 세 가지 한국어 형태소 분석기를 사용하여 건설현장 주변 보행자 사고와 관련 뉴스 기사에서 언급하고 있는 지역을 추출한 결과와 정확도를 보여주고 있다. 세 가지 분석기 모두 33건 중 31건의 사례에 대하여 지역을 올바르게 추출함으로써 93.9%의 높은 정확도를 보여주었다.

세 가지 분석기에서 모두 공통으로 잘못 분류된 첫 번째 사례의 경우, 실제 광주 동구에서 발생한 사고를 경기도로 추출하였는데 이는 '광주' 라는 지명을 광주광역시뿐만 아니라 경기도 광주시에도 이중으로 기록하면서 오류가 발생하였다. 잘못 분류된 두 번째 사례의 경우, 전라남도 담양군에서 발생한 사고를 보도하는 뉴스였으나 뉴스 기사 자체에 '광주지법' 이라는 단어가 많이 등장함으로 인해 광주로 추출되었다.

Table 4. Comparative accuracy of extracted location

Location	Cases	okt		kkma		soynlp	
		Count	Accuracy	Count	Accuracy	Count	Accuracy
Seoul	5	5	100%	5	100%	5	100%
Busan	5	5	100%	5	100%	5	100%
Daegu	2	2	100%	2	100%	2	100%
Incheon	2	2	100%	2	100%	2	100%
Gwangju	2	1	50%	1	50%	1	50%
Daejeon	1	1	100%	1	100%	1	100%
Ulsan	1	1	100%	1	100%	1	100%
Gyeonggi	9	9	100%	9	100%	9	100%
Gangwon	1	1	100%	1	100%	1	100%
North Chungcheong	1	1	100%	1	100%	1	100%
North Jeolla	3	5	100%	5	100%	5	100%
South Jeolla	1	0	0%	0	0%	0	0%
Total	33	33	93.9%	33	93.9%	33	93.9%

4.3 사고 유형

<Table 5>는 보행자 사고 통계 항목 중 사고 유형을 추출한 결과와 정확도를 보여주고 있다. soynlp 분석기가 33건 중 22건의 사례에 대한 사고 유형을 추출함으로써 66.7%로 가장 높은 정확도를 기록하였으며, 21건의 사고 유형을 맞힌 okt 분석기가 정확도 63.6%로 그 뒤를 이었다. 큰 차이는 아니지만 kkma 분석기가 20건의 사고 유형을 맞혀 60.6%의 가장 낮은 정확도를 기록하였다.

사고 유형의 종류별로는 물체에 맞음이 soynlp와 okt 분석기 모두 100%, kkma 분석기는 88.9%로 가장 잘 추출되었다. 이는 물체에 맞는다는 표현에 사용되는 '떨어' 또는 '맞아' 등의 어휘가 다른 사고를 보도할 때는 잘 쓰이지 않기 때문으로 예측된다. 반면 보행자가 넘어지는 사고와 물체에 깔리는 사고는 추출 정확도가 비교적 낮았는데, 이는 건설현장의 물체가 넘어지면서 보행자가 깔리는 사고가 다수 포함된 관계로 실제 사고 유형은 깔림인데도 넘어짐으로 기록되는 경우가 발생했기 때문이다. 또 다른 오류 사례로는 넘어짐, 깔림과 관련된 단어가 같은 횟수로 출현하여 사고 유형이 하나로 대표되지 않은 경우가 있었는데, 이처럼 사고 유형이 중복으로 간주되는 경우를 구분하기 위하여 <Table 5>에 기타 항목을 추가하여 기록하였다. 기타 항목은 여러 사고 유형 관련 어휘를 포함하고 있음을 의미하므로, 기타 항목이 적을수록 형태소 분석기의 정확도가 높다고 할 수 있는데 전체 정확도와 마찬가지로 soynlp, okt, kkma 순으로 우수한 성능을 보였다. 주로 중복하여 등장하는 사고의 유형은 앞서 설명한 넘어짐과 깔림 외에도 넘어짐과 불편함, 깔림과 불편함 등이 있었다.

<Table 5>에서 분류되지 않은 항목에 해당하는 3건의 경우, 모두 건설현장 주변을 지나던 보행자와 덤프트럭이 충돌하여 발생한 교통사고였으며, 세 가지 분석기 모두 100%로 사전에 정의하지 않은 사고 유형에 해당하지 않는 경우 별도의 항목으로 잘 추출하는 것을 확인할 수 있었다.

Table 5. Comparative accuracy of extracted accident types

Location	Cases	okt		kkma		soynlp	
		Count	Accuracy	Count	Accuracy	Count	Accuracy
Slips and trips	2	1	50%	0	0%	1	50%
Falling objects	9	9	100%	8	88.9%	9	100%
Caught under objects	12	7	58.3%	7	58.3%	7	58.3%
Inconvenience	7	4	57.1%	5	71.4%	5	71.4%
Not categorized	3	3	100%	3	100%	3	100%
Etc.	-	9	-	10	-	8	-
Total	33	33	63.6%	33	60.6%	33	66.7%

4.4 사고 원인

사고 원인을 추출하기 위해서 사고 유형에서 정의한 키워드와 함께 등장하는 단어의 집합을 추출하는 방법인 N-grams를 사용하였다. N의 값은 3으로 지정하였으며, 세 가지 형태소 분석기 중 사고 유형 추출의 정확도가 가장 우수하였던 soynlp 분석기를 사용하였다. <Fig. 3>은 보행자가 떨어지는 물체에 맞는 사고와 물체에 깔리는 사고에 대한 N-grams 결과를 워드 클라우드로 표현한 것이다.

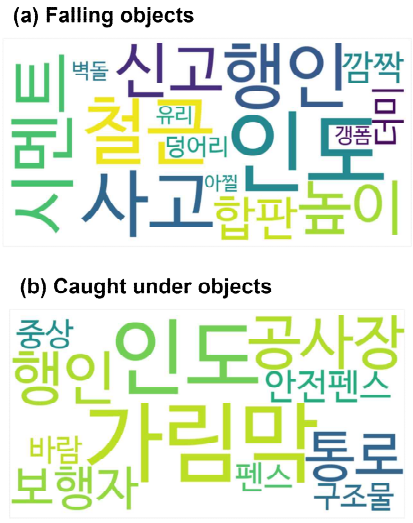

Fig. 3. Extracted cause of the accident

이를 통해 떨어지는 물체에 맞는 사고의 원인은 주로 건설현장의 자재임을 알 수 있었으며, 구체적으로는 '시멘트(13회)', '철근(12회)', '합판(10회)', '유리(6회)', '갱폼(6회)' 등이 사고를 유발했음을 확인할 수 있다. 반면 <Fig. 3>의 (b)에서 '가림막(22회)', '통로(16회)', '안전펜스(10회)', '펜스(8회)', '구조물(7회)' 이 출현한 것으로 미루어보았을 때 물체에 깔리는 사고의 원인의 경우 건설현장의 구조물들이 원인이 됨을 알 수 있다. 또한, 물체에 깔리는 사고의 경우 '바람' 이 7회 등장하였는데, 바람으로 인해 구조물이 넘어지면서 보행자가 깔리는 사고가 발생하였다는 것을 확인할 수 있다. 그뿐만 아니라 '중상(8회)' 이라는 단어를 통해 물체에 깔리는 사고가 발생하는 경우 보행자가 중상에 이르게 됨을 추측할 수 있다.

5. 결론

본 연구에서는 건설현장 주변에서 발생하는 보행자 사고에 관한 통계의 부재를 해결하기 위해 텍스트 마이닝을 적용하여 보행자 사고 관련 뉴스를 통계 데이터로 변환하는 방법을 제안하였다. 구체적으로 데이터 전처리 과정에서는 뉴스에서 수집된 말뭉치를 KoNLPY의 okt, kkma 분석기와 sonlpy 분석기를 사용하여 형태소로 토큰화하고 불용어를 제거하는 방법을 제안하였다. 건설현장 주변 보행자 사고의 통계 항목 추출 단계에서는 지역, 사고 유형, 사고 원인을 추출하는 방법으로 웹 크롤링 기반 사전 구축, 인접 어휘 기반 사전 구축, N-grams를 사용하는 단계별 방법을 기술하였다. 제안한 텍스트 마이닝 방법으로 총 33건의 보행자 뉴스 기사를 분석한 결과, 지역의 정확도 세 가지 분석기에서 모두 93.9%로 높게 나타났다. 사고 유형의 경우, 분석기를 기준으로는 sonlpy가 가장 좋은 성능을 보였으며 사고 유형을 기준으로는 떨어지는 물체에 맞는 사고가 가장 정확하게 추출되었다. 마지막으로 사고 원인의 추출을 위해 N-grams 분석을 수행한 결과 보행자가 떨어지는 물체에 맞는 사고의 주요 원인으로는 건축자재가, 보행자가 물체에 깔리는 사고의 주요 원인으로는 건축 구조물을 식별할 수 있었다.

본 연구는 현재까지 별도로 집계되지 않고 있는 건설현장 주변 보행자 사고의 통계를 과거 뉴스 기사를 사용하여 작성하는 방법을 제안함으로써, 보행자 사고의 원인을 규명하고 보행자 안전관리 강화를 위한 사고 예방 정책을 수립하는 데 기여한다. 향후 연구에서는 뉴스 기사 수집 기간의 범위를 확장하여 더 많은 텍스트 데이터를 기반으로 제안한 모델을 검증하고 보완할 것이다.

감사의 글

본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행한 성과입니다(No. NRF-2022R1F1A1072491).

References

박은정, 조성준. (2014). "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지" *제26회 한글 및 한국어 정보처리 학술대회 논문집.*

Conway, M., Doan, S., Kawazoe, A., and Collier, N. (2009). "Classifying disease outbreak reports using n-grams and semantic features." *International journal of medical informatics*, 78(12), pp. 47-58.

Goh, Y. M., and Ubeynarayana, C. U. (2017). "Construction accident narrative classification: An evaluation of text mining techniques." *Accident Analysis & Prevention*, 108, pp. 122-130.

Hotho, A., Nürnberger, A., and Paaß, G. (2005). "A brief survey of text mining." *Journal for Language Technology and Computational Linguistics*, 20(1), pp. 19-62.

Khalef, R., and El-adaway, I. H. (2021). "Automated identification of substantial changes in construction projects of airport improvement program: Machine learning and natural language processing comparative analysis." *Journal of management in engineering*, 37(6).

Kinawy, S. N., Bakht, M. N., and El-Diraby, T. E. (2017). "Mismatches in stakeholder communication: The case of the Leslie and Ferrand transit stations, Toronto, Canada." *Sustainable cities and society*, 34, pp. 239-249.

요약 : 본 연구는 건설현장 주변 보행자 안전사고가 중대 시민 재해에 해당함에도 불구하고, 관련 통계 데이터가 충분히 확보되지 않아 소홀히 관리되는 문제점을 해결하고자 텍스트 마이닝을 활용하여 뉴스 기사에서 보행자 사고 통계 데이터를 추출하는 방법을 제안하였다. 제안된 방법을 통해 수집된 총 33건의 보행자 사고 뉴스 기사를 분석한 결과, 지역 추출에 대한 정확도는 93.9%로 매우 높게 나타났다. 넘어짐, 떨어지는 물체에 맞음, 깔림, 불편함으로 분류되는 사고 유형 분류에서는 떨어지는 물체에 맞음이 세 가지 형태소 분석기(okt, kkma, soynlp)에서 모두 높은 정확도를 보여주었다. 본 연구는 그동안 집계되지 않았던 건설현장 주변 보행자 사고의 통계 정보를 과거 뉴스 기사를 통해 작성할 가능성을 보여주었으며, 향후 보행자 안전 강화 및 사고 예방 정책 수립에 중요한 기초자료로 활용될 수 있다.

키워드 : 보행자 안전, 건설현장, 텍스트 마이닝, 사고