

Original Article

안전신문고를 이용한 재난 예측 방법론 제안

이준¹ · 신진동² · 조상명³ · 이상화^{4*}

¹한국교통연구원 안전방재연구센터 부연구위원, ²국립재난안전연구원 안전연구실 시설연구관, ³행정안전부 안전정책실 생활안전정책관, ⁴목원대학교 산학협력단 책임연구원

Research Suggestion for Disaster Prediction using Safety Report of Korea Government

Jun Lee¹, Jindong Shin², Sangmyeong Cho³, and Sanghwa Lee^{4*}

¹Associate Research Fellow, Transport Safety & Disaster Prevention Research Center, Korea Transport Institute (KOTI)

²Senior Research Officer, Safety Research Division, National Disaster Management Research Institute (NDMI)

³Director General, Life Safety Policy Bureau, Ministry of the Interior and Safety (MOIS)

⁴Senior Researcher, Industrial-Academic Cooperation Foundation, Mokwon University

요약

안전신문고는 2014년부터 운영되고 있으며, 2019년 7월까지 약 1백만 건의 누적신고건수가 존재한다. 본 연구에서는 정보화시대가 되고 있는 현 시점에서 약 116만 건이 넘는 안전신문고의 신고내용을 분석하여 국민의 소리와 관심이 과연 얼마나 힘이 있고 의미가 있는지 확인하고자 한다. 특히, 예측능력에 관심을 두고 있는데, 과연 안전신문고의 신고내용이 향후 일어날 수 있는 재난과 연관성이 있는지 확인하고자 하였다. 이를 위해 연구진은 안전신문고에 신고된 자료를 텍스트로 받아 자연어 분석 방법(Natural Language Processing)론에 의해 분석하였다. 이를 토대로 안전신문고 분석 기간 동안의 신문기사를 분석하여 안전신문고와 신문 기사 내용 간의 상관관계를 분석하였다. 그 결과 응답 및 확인 관련 보고서의 수가 증가함에 따라 몇 달 내 사고가 발생하였으며, 사회의 불안에 대해 사전에 보고된 안전신문고의 내용을 분석하면 미래 재난 예측에 활용될 수 있을 것이라 판단된다.

핵심용어: 안전신문고, 텍스트 마이닝, 예측, 자연어 분석(NLP), 단어구름

ABSTRACT

Anjunshinmungo (The safety e-report) has been in operation since 2014, and there are about 1 million cumulative reports by June 2019. This study analyzes the contents of more than 1 million safety newspapers reported at the present time of information age to determine how powerful and meaningful the people's voice and interest are. In particular, we are interested in forecasting ability. We wanted to check whether the report of the safety newspaper was related to possible disasters. To this end, the researchers received data reported in the safety newspaper as text and analyzed it by natural language analysis methodology. Based on this, the newspaper articles during the analysis of the safety newspaper were analyzed, and the correlation between the contents of the newspaper and the newspaper was analyzed. As a result, accidents occurred within a few months as the number of reports related to response and confirmation increased, and analyzing the contents of safety reports previously reported on social instability can be used to predict future disasters.

Keywords: Anjunshinmungo (safety e-report), Text mining, Prediction, NLP, Wordcloud

*Corresponding author: Sanghwa Lee, soboru2@mokwon.ac.kr

Received: 14 November 2019, Revised: 12 December 2019, Accepted: 16 December 2019



1. 서론

2014년 9월 30일 안전신문고는 안전한 생활환경 조성과 국민의 목소리를 정부가 직접 듣는다는 취지 안전행정부(현 행정안전부)에서 출범하였다. 안전신문고를 통해 2019년 7월까지 약 116만 건의 신고가 접수되었다. 초기에는 인터넷을 이용한 신고가 대부분을 차지하였지만, 모바일 앱과 사진 업로드 기능이 가능해지면서 하루 5,000여 건의 신고가 접수되기도 하였다. 모든 신고는 접수, 처리, 결과에 이르는 모든 과정을 등록된 휴대폰으로 전달하고 있다. 2019년 4월부터 행정안전부는 안전신문고를 통해 4대 불법주·정차 신고제를 시행하고 있으며, 실제로 과태료를 부과하면서 안전신문고의 역할은 더욱 커지고 있다.

안전신문고의 역사는 조선시대에도 있었는데, “신문고”라는 단어 자체가 백성이 국가에 직접 고하는 형식으로 전 세계적으로도 유례가 없는 국민과 국가의 직접소통방식의 하나이다. 안전신문고는 안전과 관련한 다양한 국민의 의견을 국가가 직접 청취하고 처리하는 기능을 계승한 것이라 할 수 있다.

본 연구에서는 정보화시대로 넘어가고 있는 현 시점에서 약 116만건이 넘는 안전신문고의 신고내용을 분석하여 국민의 소리와 관심이 과연 얼마나 힘이 있고 의미가 있는지 확인하고자 한다. 즉, 안전신문고 내용을 분석하여 특이점과 시사점을 찾아보고자 한다. 특히, “사고 예측능력이 가능한가”라는 부분에 연구진과 행정안전부의 관심이 집중되는 만큼 안전신문고의 신고내용이 향후 일어날 수 있는 재난과 연관성이 있는지 확인하고자 하였다.

안전과 관련한 민원신고 성격을 가진 안전신문고는 신고자가 텍스트 형식으로 민원을 입력하고, 사진을 업로드하여 현재 또는 미래의 위험요소에 대하여 신고하는 형식이다. 이러한 방법은 국민들이 쓰는 보편적인 언어와 기술내용에 의지하기 때문에, 때로는 오차가 있고, 같은 표현도 달리 쓰거나 오류가 있는 자연어(Natural Language) 형태이다.

본 연구에서는 이러한 서술형 텍스트를 분석하기 위하여 텍스트 마이닝의 형태소 분석 방법론(Cho et al., 2018)을 이용하여 안전신문고의 내용을 분석하였으며, 동일한 기간의 신문기사를 검색(Bae et al., 1998)하여 안전신문고 신고내용과 상관성을 분석하였다. 분석 범위는 안전신문고가 운영된 2014년 9월~2019년 7월까지 접수된 자료를 대상으로 하였으며, 해당 자료는 행정안전부에서 제공하였다. 또한 본 분석은 텍스트 중심으로 연구를 진행하였기 때문에 업로드된 이미지들은 분석에서 제외하였다.

2. 본론

2.1 안전신문고 분석 방법

2.1.1 형태소 분석 방법론

본 연구에서는 인공지능에 기반한 자연어 분석 방법론(Natural Language Processing ; NLP)을 이용하여 Window 환경에서 Python 3.0을 사용하여 분석하였다. Python은 문법이 간결하고 표현구조가 인간의 언어와 비슷하여 초보자도 쉽게 배울 수 있으며, 풍부한 라이브러리가 있어 웹 개발, 데이터 분석, 머신러닝, 그래픽, 학술 연구 등 여러 분야에서 활용되고 있는 프로그래밍 언어이다(Kang et al., 2017).

자연어처리는 그 속성상 언어학적 측면에서의 연구와 컴퓨터 공학적 측면에서의 연구가 동시에 이루어질 때 접근 가능하다(Hawng., 1998). 자연어처리란 컴퓨터를 이용하여 사람 언어의 이해, 생성 및 분석을 다루는 인공지능 기술로 일상 생활 언어를 형태 분석, 의미 분석, 대화 분석 등을 통하여 컴퓨터가 처리할 수 있도록 변환, 이 처리한 결과물을 사람의 편의성에 입각하여 텍스트, 음성, 그래픽 등으로 시각화하는 작업을 뜻한다(Kang et al., 2017).

자연어처리는 언어학을 근간으로 하며, 언어학은 말소리를 연구하는 음운론(Phonology), 단어와 형태소를 연구하는 형태론(Morphology), 문법과 맥락/담화를 각각 논의하는 통사론(syntax), 의미론(Semantics) 등 세부 분야가 존재한다(Fig. 1 참조).¹⁾

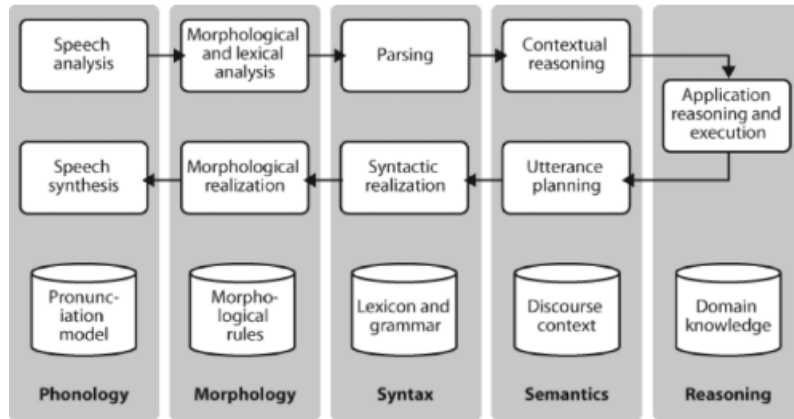


Fig. 1. Overall procedure of natural language processing

자연어처리 절차 또한 위와 같은 단계와 유사하게 4단계(형태소 분석, 구분 분석, 의미 분석, 담화 분석)로 진행된다. 형태소란 의미가 있는 최소의 단위로 문법적, 관계적인 뜻을 나타내는 단어 또는 단어의 부분을 의미하며, 형태소 분석은 단어(또는 어절)를 구성하는 각 형태소를 분리하고, 분리된 형태소의 기본형 및 품사 정보를 추출하는 것을 의미한다. 텍스트를 형태소 단위로 분리하는 방법은 크게 2가지로 다음과 같다.²⁾

- 방법 1 : 어떤 단어는 어떤 품사라 정의하고, 이를 이용하여 해당 품사로 분리하는 방법(사전을 활용하는 방법)
- 방법 2 : 모델을 이용해서 학습시키는 방법(문장에서 단어 구별)

구문 분석은 문장 즉, 일련의 문자열을 자연어의 경우 형태소로 분리하고, 해당 구문론적 관계를 도식화하여 명확히 하는 분석절차로 문장의 구조와 형태를 파악하는 단계이다.

의미분석은 문장의 의미를 이해하는 단계이며, 담화 분석은 문장의 숨은 의미나 대화의 내용을 이해하는 단계이다.³⁾

형태소 분석은 언어 특성에 따라 적절한 분석방법을 사용해야 한다. 중국어, 일본어의 경우 띄어쓰기를 하지 않기 때문에 단어 분리 문제가 중요하며, 핀란드어의 경우 굴절이 심하기 때문에 형태소의 원형 복원을 강조하여야 한다. 한국어는 영어와 달리 여러 형태소가 합쳐져 하나의 어절(단어)를 이루기 때문에 형태소 분리 문제가 매우 중요하다. 좀 더 설명을 하면 한국어를 띄어쓰기로 문장을 토큰화⁴⁾시켜도 매번 다른 조사가 붙어서 처리가 힘들고 번거로워진다. 또한 한국어는 영어와 비교하여 띄어쓰기가 어려운 편이며, 그러한 이유로 실제 사람들이 이걸 철저히 지키지 않는다. 따라서 한국어 텍스트를 형태소로 단위로 나누어서 분석을 수행하여야 한다(Yoo, 2018).

Python에서 형태소를 분리하기 위해 가장 많이 사용하는 패키지는 코엔엘파이(KoNLPY)로 앞서 설명한 텍스트를 형태소 단위로 분리하는 방법 중 첫 번째 방법을 이용하고 있다. 현재 대부분의 자연어처리 기술은 사용자가 많은 영어와 중국어를 기반으로 발전했기 때문에 한국어 자연어처리 기술은 이에 비해 뒤떨어진다. 코엔엘파이 패키지는 한국어 정보처리를 위

1) <https://ratsgo.github.io/natural%20language%20processing/2017/03/22/lexicon/>
 2) <https://tariat.tistory.com/559>
 3) <https://nlp4us.tistory.com/entry/%EB%91%90%EB%B2%88%EC%A7%B8-%EC%9E%90%EC%97%B0%EC%96%B4%EC%B2%98%EB%A6%AC>
 4) 토큰화(Tokenization)란 주어진 코퍼스(corpus)에서 토큰(token)이라 불리는 단위로 나누는 작업을 의미하며, 토큰의 단위가 상황에 따라 다르지만, 보통 의미있는 단위로 토큰을 정의한다(Yoo, 2019).

한 것으로 복잡 미묘한 한국어 텍스트에서 유용한 특성을 추출하기 위해 개발되었던 한국어 형태소 분석기를 활용하여 한국어 자연어 처리에 필요한 각종 사전, 말뭉치, 도구 등을 제공하고 있다(Kang et al., 2017).

코엔엘파이 패키지는 다양한 형태소 분석 방법을 제공하고 있으며, 형태소 분석에서 형태소를 비롯한 어근, 접두사/접미사, 품사 등 다양한 언어적 속성의 구조를 파악해야 한다. 코엔엘파이 패키지에서 제공하는 형태소 분석기는 총 5개 (jhannanum(한나눔), kkma(꼬꼬마), komoran(코모란), twitter(트위터), mecab(메깅))로 각 클래스간 성능은 다음의 Table 1과 같다. 성능으로 판단하자면 Mecab이 우수하고 그 다음은 Twitter가 괜찮은 것으로 보인다.

시간적인 성능 이외에 “아버지가방에들어가신다”라는 문장을 각 클래스들이 어떻게 태그하는지 살펴보면 다음의 Table 2와 같다.

형태소 분석기 중에서 제일 효율이 괜찮은 것은 Mecab으로 판단되나 Mecab 같은 경우 윈도우 7을 지원하지 않을 뿐만 아니라 윈도우 10에서는 빌드 등의 과정을 거쳐야만 사용할 수 있다.⁵⁾

데이터 가공 과정에서 필요한 코엔엘파이 함수 목록 및 기능에 대한 간단한 설명은 다음의 Table 3와 같다(Kang et al., 2017).

Table 1. KoNLpy package Performance between classes

Division	Preload time (secs)	Class pos method execution time (secs)
Kkma	5.6988	35.7163
Komoran	5.4866	25.6008
Hannanum	0.6591	8.8251
Twitter	1.4870	2.4714
Mecab	0.0007	0.2838

Source: <http://konlpy.org/ko/v0.4.3/morph/>

Table 2. “아버지가방에들어가신다” KoNLpy Pie Analysis (Example)

Kkma	Komoran	Hannanum	Twitter	Mecab
아버지가방에들어가/N	아버지/NNG	아버지가방에들어가신다/NNP	아버지/NNG	아버지/NOUN
이/J	가방/NNG		가/JKS	가방/NOUN
시니다/E	에/JKM		방/NNG	에/Josa
	들어가/VV		에/JKB	들어가신/Verb
	시/EPH		들어가/VV	다/Eomi
	니다/EFN		신다/EP+EC	

Table 3. KoNLpy package function

Function	Explanation
Twitter/Kkma/Komoran/Hannanum/Mecab	Analyze sentence morphology
Counter	Count the frequency of the words entered
nouns	Extract nouns
morphs	Parse the syntax into morphological units
pos	Parses a phrase into words
pprint	Print Unicode more easily
concordacne	Find the part that matches the entered phrase

5) <https://hyrama.com/?p=463>

코엔엘파이 모듈을 이용하여 구문을 분석하면, 우리나라의 5언 9품사보다 많은 36개 품사로 구분된다(Yoon et al., 2019). 안전신문고는 신고자가 문장으로 내용을 작성하기 때문에 명사와 동사의 활용이 많아 어근 정보를 파악하기 위해서는 구문 분석을 반드시 해주어야 한다.

예를 들어 “깨뜨리셨더군요”라는 문장을 코엔엘파이로 분석해보면 어근과 접속사, 어미를 구분하여야 “깨졌다”라는 어근 정보를 비로소 얻을 수 있다. 이에 대한 자세한 설명은 다음의 Fig. 2와 같다(Kim, 2003).

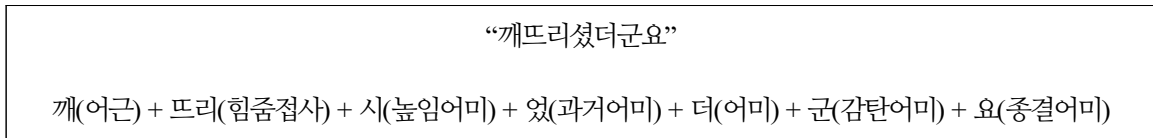


Fig. 2. “깨뜨리셨더군요” KoNLPY Pie Analysis Process (Example)

이와 같이 문장을 형태소로 구분을 하고 명사 또는 동사의 원형을 찾아서 어근을 찾는 방식으로 문장 의도를 대표적인 명사로 파악하는 방식을 사용하였다(Jalaj, 2019).

2.1.2 안전신문고 분석 결과

분석 대상은 2014년부터 2019년 7월까지 안전신문고에 접수된 약 116만건을 대상으로 하였으며, 연도별 신고건수, 신고 내용에 있는 문단 수와 및 단어 수는 Table 4와 같다. 2014년 초기에는 신고건수가 1,488건에 그쳤지만, 2019년 4대 불법주·정차 신고제가 시행되면서 폭발적으로 그 수가 증가하였다. 총 분석한 신고건수는 약 116만건, 문단 수 약 139만개, 약 530억 개의 단어를 활용하였다.

Table 4. Data size

Year	The number of reports	Number of Paragraph	Sum of word Frequency
2014	1,488	4,713	24,171,211
2015	74,123	276,265	1,084,011,768
2016	152,768	320,229	3,002,669,540
2017	226,919	375,717	21,806,487,560
2018	236,002	416,568	27,118,094,224
2019	473,139	533,907	97,602,276,212
Total	1,164,439	1,393,492	53,035,434,303

Fig. 3은 연도별 안전신문고 신고건수를 워드클라우드(wordcloud)로 분석한 결과이다. 2014년에는 “차량(도로교통 등)”과 “시설(도로시설 등)”에 대한 신고내용이, 2015년에는 “차량” 이외에 “아이(어린이 등)”, “안전” 관련 신고 내용이 가장 많았다. 2016년에는 “안전”, “위험”, “차량” 등이 신고가 많았으며, “주차” 관련 신고내용이 등장하기 시작하였다. 2017년도부터는 앱의 활용으로 “갤러리”와 “촬영”, “휴대폰”과 같은 단어의 증가가 두드러진다.

특히, 4대 불법주정차 신고제를 시행한 2019년에는 “불법”, “정차”, “신고”라는 단어가 많아졌으며 이중 “불법”이란 단어의 신고 내용 등장은 4대불법주·정차 신고들이 늘어나면서 크게 증가한 것이다(전체 신고의 60% 이상 차지). 또한 안전신문고 신고 초기에 불만사항 중심의 민원에서 불법주·정차 신고 유형이 생기면서 신고자가 합법과 불법을 구분하여 신고하는 성향의 변화를 관찰할 수 있게 되었다(예 불법~, 과 불법은 아니지만~이라는 구분이 추가 발생).

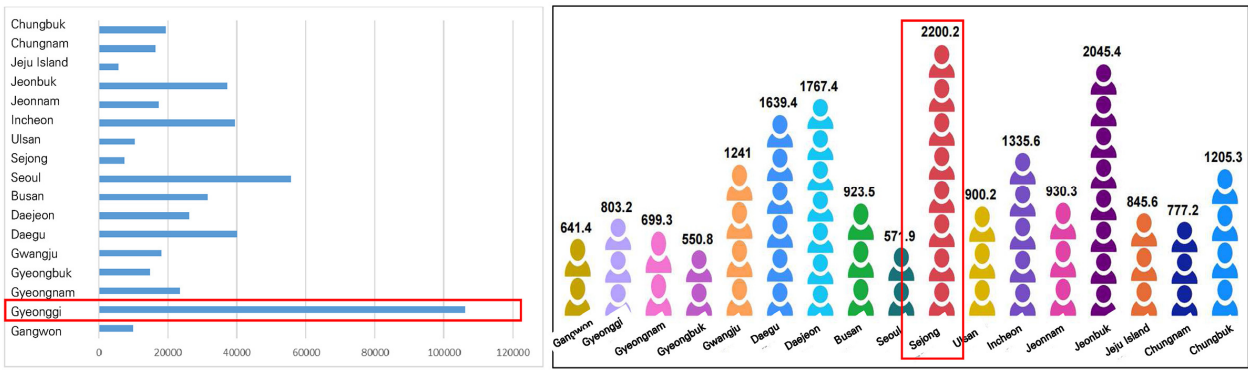


Fig. 5. (Left) Number of reports by region (2019), (Right) Number of reports by region per 100,000 population (2019)

2.1.3 뉴스기사 스크랩 분석

안전신문고의 안전 이슈가 사회적으로 어떤 위험요소와 관계가 있는지 분석하기 위해서는 가정이 필요하다. 안전신문고는 신고자가 위험을 감지한 경우 위험에 대하여 사전에 신고하는 특성이 있으며, 뉴스기사와 같은 언론보도는 재난안전의 사회적 이슈에 대하여 사건 발생이후에 급격히 후속보도가 증가하는 특성이 있다. 본 연구에서는 다음과 같이 가정하였다.

- 가정 1 : 안전신문고의 신고 이슈는 국민이 신고한 위험요소를 직접 나타내고 있으며, 신문기사는 사회의 위험 이슈를 반영한다. 단어의 노출빈도는 관심도와 같다.
- 가정 2 : 안전신문고는 국민이 자발적으로 위험요소에 대한 사전적인 알림 기능을 하고 있으며, 신문기사는 발생한 재난과 사고에 대하여 후속적으로 보도한다.

이러한 가정을 바탕으로 안전신문고의 신고내용과 신문기사의 보도내용을 비교하기 위해 뉴스 스크랩 및 분석을 진행하였다.

뉴스 스크랩은 안전신문고 운영 기간과 동일기간을 설정하여 네이버 뉴스 기사를 월별로 “재난”, “안전” 관련 기사 1,000 건씩 조회 수가 많은 기사를 대상으로 수집하였다. 노출빈도를 비교할 단어는 안전신문고로 신고된 상위 20개의 단어로 한정하여 비교하였다.

안전신문고의 월별 노출빈도와 뉴스기사의 월별 1,000건을 대상으로 노출빈도를 추출하여 2014년부터 2019년까지 5년 간 추출된 총 20개 단어의 노출빈도는 Table 5와 같다.

앞서 기술한 바와 같이 노출빈도 20개를 대상으로 5년의 분석기간 동안 1개월 별로 안전신문고의 신고내용과 신문기사의 발생빈도를 분석하였다. 월별 뉴스기사의 20개 단어에 대한 노출빈도는 Table 6과 같다.

2.2 안전신문고 분석 결과

동일한 기준의 단어간 노출빈도를 비교하기 위해서 2014년부터 2019년까지 월별로 안전신문고 신고내용에서 상위 노출된 20개 지정단어를 먼저 추출하였다(Table 5).

또한 동일단어에 대하여 뉴스에서 사용된 단어의 빈도수(Table 6)를 추출하였으며, 월별로 두 빈도수 간의 상관분석을 진행하였다.

상관분석은 연속 변수로 측정된 두 변수간의 선형관계를 분석하는 기법으로 한 변수가 증가하면 다른 한 변수도 선형적으로

Table 5. Frequent 20 words in AnjunSinmungo and Newspaper 2014~2019

Rank	AnjunSinmungo		Newspaper	
	Word	Frequency	Word	Frequency
1	Safety	12,354,875	Safety	3,686,578
2	Disaster	11,458,252	Disaster	2,545,605
3	Accident	9,536,584	Accident	2,034,715
4	Danger	8,885,214	Danger	868,857
5	Prevention	6,584,412	Prevention	558,143
6	Check	6,571,225	measure	382,382
7	Infrastructure	4,365,562	Drown	381,860
8	Measure	3,256,952	Infrastructure	332,265
9	Fire	3,200,521	Conflagration	323,687
10	Prepare	2,532,210	prepare	294,294
11	Weak	2,210,320	weak	227,318
12	Thawing	825,202	Fire	225,862
13	Conflagration	824,325	Response	223,901
14	Drown	405,088	Caution	197,743
15	Response	404,055	Thawing	185,259
16	Caution	398,524	Insurance	171,567
17	Training	388,215	Collapse	169,897
18	Collapse	388,003	Catastrophe	161,980
19	Insurance	101,025	Weak	149,695
20	Catastrophe	100,852	Training	141,960

Table 6. Word frequency in Newspaper 2019. Jan ~Jul

	Jan.	Feb.	Mar.	Apr.	May	Jun	Jul.
Safety	74,382	76,020	76,657	75,383	77,203	74,746	77,567
Disaster	52,417	53,782	54,055	53,418	54,601	52,872	55,147
Accident	38,755	39,392	39,756	39,119	39,938	38,937	40,029
Danger	17,987	18,442	18,624	18,351	18,715	18,351	18,897
Prevention	11,132	11,678	11,769	11,314	12,042	11,314	12,042
Measure	9,282	9,373	9,464	9,373	9,737	9,373	9,737
Infrastructure	7,753	8,117	8,117	8,026	8,117	8,026	8,117
Prevention	7,753	7,844	7,844	7,844	7,844	7,753	8,026
Conflagration	7,280	7,462	7,553	7,371	7,644	7,280	7,735
Prepare	6,461	6,825	6,825	6,825	7,007	6,825	7,189
Weak	5,824	6,006	6,097	5,915	6,188	5,915	6,188
Thawing	5,642	5,824	5,915	5,642	5,915	5,642	5,915
Fire	4,641	4,732	4,823	4,732	4,823	4,732	4,823
Response	4,467	4,467	4,467	4,467	4,467	4,467	4,467
Caution	4,004	4,004	4,186	4,004	4,186	4,004	4,186
Training	3,731	3,913	3,913	3,731	3,913	3,731	4,004
Collapse	3,640	3,640	3,640	3,640	3,640	3,640	3,640
Insurance	3,185	3,185	3,185	3,185	3,185	3,185	3,276
Catastrophe	3,094	3,367	3,367	3,367	3,367	3,276	3,549

로 증가 혹은 감소하는가를 나타내는 것이다. 상관분석은 앞서 설명한 바와 같이 두 변수간의 선형관계에 초점을 맞추기 때문에 선형관계를 갖는가, 선형관계를 갖는다면 어느 방향인가, 그 관계는 얼마나 큰가 등에 대해 분석한다. 상관분석에는 두 변수 사이의 선형적인 관계 정도를 나타내기 위해 상관계수(Correlation coefficient)를 사용한다. 상관관계를 나타내는 그림은 다음의 Fig. 6와 같으며, -1 ~ +1 사이의 값을 나타낸다. 일반적으로 1에 가까울수록 양의 선형관계, -1에 가까울수록 역의 선형관계가 있다고 판단하며, 0인 경우 상관관계가 없음을 의미한다.

해석의 차이가 있을 수 있으나, 0.6 이상의 분석 결과는 어느정도 양의 상관관계가 있음을 나타내며, 0.9 이상은 매우 강한 상관관계가 있음을 의미한다(Fig. 7 참조).

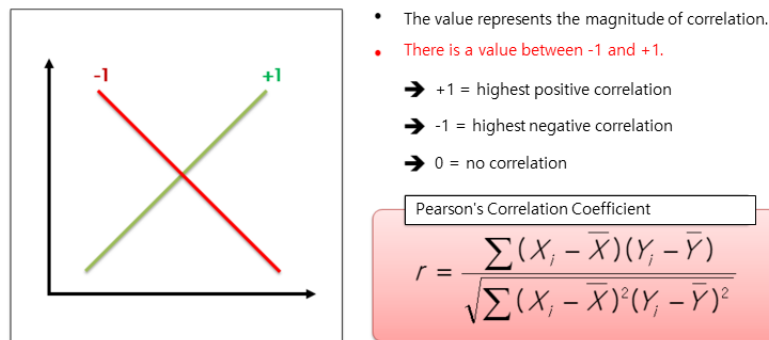


Fig. 6. Correlation

Perfect	+1	-1
Strong	+0.9	-0.9
	+0.8	-0.8
	+0.7	-0.7
Moderate	+0.6	-0.6
	+0.5	-0.5
	+0.4	-0.4
Weak	+0.3	-0.3
	+0.2	-0.2
	+0.1	-0.1
Zero	0	

Fig. 7. Correlation mean (Source : Christine Dancey and John Reidy (2011), Statistics without Maths for Psychology)

본 연구에서는 가정 2와 같이 신문기사의 단어노출은 사회의 이슈에 후속적으로 나타남을 가정하여, 1개월 차이를 두고 (즉 안전신문고 1월 자료와 뉴스기사 2월의 자료를 비교하여 1개월의 차이를 둠) 분석하였다. 여기서 의미가 있는 상관관계 분석결과만 발췌하면 다음의 Table 7과 같다.

본 상관관계는 최빈 20개 단어의 노출빈도가 증·감을 할 때 1개월 차이를 두고 안전신문고와 뉴스기사간의 노출빈도의 연관성을 판단하기 위한 과정이다.

안전신문고에 “조치(0.66)”, “확인(0.72)”이라는 단어의 빈도가 높아지면, 뉴스에서는 “안전”이라는 단어의 노출빈도가 높아지는 것으로 나타났다.

또한 안전신문고에서 “사고(0.68)”, “조치(0.65)”, “확인(0.58)” 관련 단어의 노출빈도가 높아지면 뉴스기사에서는 “위험” 관련 보도가 많아지는 것으로 나타났다. 안전신문고에 “확인(0.66)” 관련 신고가 많아지면, 신문기사에서는 “사고” 관련 내용이 증가하는 것으로 나타났다.

Table 7. Correlation result

	Safety_news	Danger-news	Accident-news
Danger shin-mum-go	0.56	0.21	0.43
Accidnet shin-mun-go	0.47	<u>0.68</u>	0.24
Response shin-mun-go	<u>0.66</u>	<u>0.65</u>	0.25
Check shin-mun-go	0.72	<u>0.68</u>	0.66

안전신문고 분석 결과 시사점을 도출하면 다음과 같이 3가지로 요약할 수 있다.

• 시사점 1

앞 절의 내용에서 살펴본 바와 같이 정확히 일치하는 뚜렷한 상관관계(0.9 이상의 상관관계)가 나타난 것은 없었지만, 이상치의 제거 또는 다양한 외부영향요인에 대한 처리 없이도 일부의 단어빈도가 0.6 이상의 상관관계가 나타나는 것을 확인할 수 있었다.

즉 안전신문고의 “조치”, “확인”, “사고” 등의 단어빈도가 높아지면, 신문기사에 등장할 수 있는 “안전” 관련 기사가 많아지거나, “위험”에 대한 기사가 많아져, 대형사고가 발생하는 것을 알 수 있었다. 이상치의 제거와 인과관계를 설명할 수 있는 사례를 찾는다면, 상관계수를 높일 수 있을 것이라 판단된다.

• 시사점 2

여기서 시간의 흐름을 살펴보면, 4절에서 가정한 것과 같이 안전신문고의 신고내용은 사고가 발생하기 전에 신고가 된 내용이고, 신문기사의 내용은 사고 발생이후 사회적 관심도가 높아져서 노출이 많아진 것이다. 실제 신고내용의 예를 들어 “가로등이 어두워 시야가 나쁘니 ‘조치’가 필요하다”는 ‘조치’ 신고가 안전신문고에 여러 번 발생하면, 이후에 큰 교통사고가 발생하고, 신문에서 “대형 교통‘사고’ 발생”이라는 ‘사고’ 및 ‘안전’ 관련 뉴스기사가 증가하게 되는 것이다.

다양한 사고와 그 내용특정은 어렵지만, 안전신문고에 출현하는 위험신호가 앞으로 발생한 재난을 예고할 수 있을 것이라는 기대가 생기는 대목이다. 본 연구에서는 분석의 한계로 더욱 높은 상관관계와 인과관계를 설명하기는 어려웠다.

• 시사점 3

안전신문고와 뉴스기사의 노출빈도를 표준화하여 시간대별로 표현하면 Fig. 8과 같이 나타난다.

상관관계가 다소 있음은 앞 절에서 이미 밝히고 있지만, 시간상에 노출빈도의 경향을 보면 또 다른 관계를 예측해 볼 수 있다. 안전신문고의 “안전” 이슈와 “화재”가 시간차를 두고 서로 급증하는 것을 볼 수 있다.

안전신문고의 ‘위험’ 관련 신고는 2015년 1월, 5월, 8월, 2016년 3월, 6월, 10월, 2017년 5월, 8월, 10월 등에서 신고빈도가 높아지는데, 뉴스에서 ‘화재’ 관련 기사는 1개월 또는 2개월 후에 노출 빈도가 높아지는 것을 알 수 있었다. 다만 안전신문고에 2019년 부터는 4대 불법주정차 신고를 받게되면서, 60% 이상이 불법주정차 신고가 접수되어 상대적으로 단어의 노출빈도가 편중되어 유의미한 패턴이 관측되지 않았다. 만약 우연한 관계가 아니라면, 안전신문고에 “위험” 신고가 많아지면, 수개월 내에 대형 화재가 발생한 것을 알 수 있다. 즉, 안전신문고가 대형 사고를 예측할 수 있는 예측력을 갖고 있는 가능성이 열린 것이다.

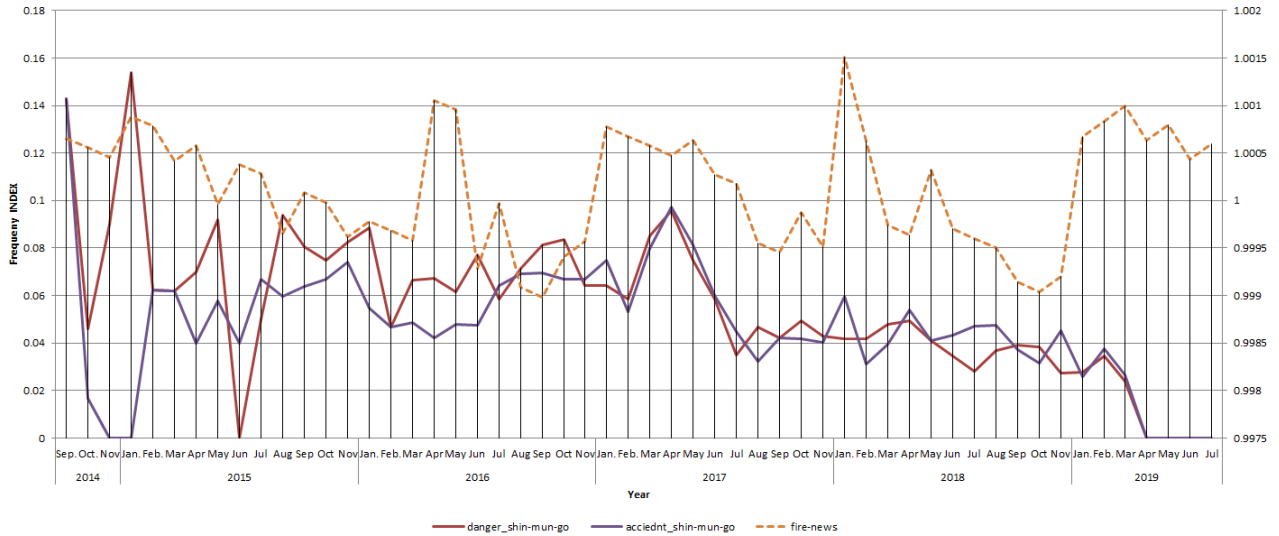


Fig. 8. Frequency index for Shin-mum-go and news contents

3. 결론

본 연구에서는 안전신문고의 신고 내용을 분석하고 해당 신고내용과 동일시점의 뉴스기사에서 재난안전 키워드의 상관 분석을 진행하였다. 안전신문고의 신고내용을 명사 단위로 분석을 진행 하였으며, 1개월 차이를 둔 시점의 신문기사의 노출 단어와 비교를 하였다. 특히, 최빈단어 20개를 비교한 결과 1개월 차이를 두고, 안전신문고에 “위험” 관련 신고가 증가하면 실제 사고가 발생하여 뉴스에서도 “안전” 관련 보도가 많이 발생할 수 있는 가능성을 찾을 수 있었다. 이러한 결과를 통해서 안전신문고가 재난과 재해발생 이전에 시그널을 주고 있을 가능성에 대하여 의미 있는 결과가 도출되었다.

안전신문고는 위험요소와 위해한 상황에 대해 국민이 국가에 신고하는 제도이며, 이는 미래의 재난발생 가능한 요소에 대하여 국가에 알리는 것이다. 물론, 한계점도 있다.

국민이 “지진”과 같은 자연재난을 예측하는 것은 어려울 것이다. 고도의 기술과 측정장비를 필요로 하기 때문이다. 하지만 국민들의 더 많은 눈이 재난을 예측하는 것이 효율적일 수도 있다. 보편적인 국민수준에서도 취약한(노후된) 구조물, 잘못된 표지판, 버스기사의 졸음운전 등 생활 주변에 사소한 위험요인들에 대한 인지를 할 수 있으며, 이는 공무원(정부)이 직접 수집하는 것보다도 훨씬 더 많은 감시자 활동으로 정보력이 월등히 앞서기 때문이다.

지금까지 재난에 대한 예측은 여러 분야에서 연구되어지고 노력하였지만, 재난을 예측한다는 것 자체가 어려웠다. 하지만 본 연구에서 분석한 안전신문고와 같이 사전에 위험 신호를 받을 수 있는 제도와 자료가 있다면, 충분히 미래의 재난예측에 활용이 가능할 것이라 판단된다.

본 논문은 분석의 한계로 더 높은 상관관계와 인과관계를 설명하기는 어려웠지만 향후 안전신문고를 활용한 재난/사고예측 인공지능 도입 연구를 수행한다면 이러한 한계를 극복할 수 있을 것이라 생각된다. 더불어 안전신문고에 대한 지속적인 모니터링, 평가 등을 통해 해당 자료를 다양한 측면에서 분석하여 정부 정책에 활용할 수 있도록 해야 할 것이다.

Acknowledgement

본 연구는 2019년 국립재난안전연구원에서 수행한 “안전신고의 성과 평가 모델개발 연구”의 지원으로 작성되었으며, 자료를 제공해 주시고 연구에 협조해 주신 행정안전부 안전정책실 이경자 행정사무관님께 감사드립니다.

References

- Bae, Y. M. (1998). A Study on the Preventive Education for Adolescents' Problems in Home Economics Education - Adolescents' Culture Analyzed with Newspaper Articles Concerning Adolescents. Korean Home Economics Education Association. 10(2): 131-144.
- Baek, Y. (2017). Text Mining Using R. Press Hanwool. pp. 171-178.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research. 3: 993-1022.
- Cho, J. Y., Kim, H. S., and Cho, M. J. (2018). A Study on the Change of Issues with Adolescent Problem by Using Text Mining - The Internet News Articles for the Years 2008 to 2018. Journal of Educational Innovation Research. 28(4): 461-487.
- Chung, B. M., Kim, T. W., Lee, J., and Chung, J. S. (2015). Twitter Topic Extraction and Topic Category Decision using LDA Model. Korea Institute of Information Sciencetists An Engineers, Proceeding. pp. 787-788
- Chung, D. Y., LEE, J. S., and Park, S. S. (2016). Technology Trend Analysis using Wordcloud. Confernce Proceeding of Korean Institute of Intellignet Systems. 26(1): 17-18.
- Dancey, C. and Reidy, J. (2011). Statistics without Maths for Psychology. Prentice Hall. 5th edition. p. 175.
- Hwang, H. S. (1998). Stemming Methodology for Natural Language Processing. Phrasebook. 37: 439-440.
- Kang, H. B., Kang, J. H., Rye, Y. N., Son, J. Y., Yoon, B. Y., and Lee, J. S. (2017). A Study on the Usability of Opinion Mining in Post-Occupancy Evaluation. The 17th Proceedings of the Architectural Institute of Korea Conference. 37(2): 1144-1146.
- Kim, Y. H. (2003). Social Netwok Analysis. Barkyoung Press.
- Lim, D. S. (2001). Error Types of Word-class Analysis in Korean-English Bilingual Children. Master Thesis. Yeonsei University.
- Park, S. J. (2015). A Topic Analysis of SW Education Textdata Using R. Korean Association of Information Education. 19(4): 517-524.
- Thanaki, J. (2019). Python Natural Language Processing (Advanced Machine Learning and Deep Learning Techniques for Natural Language Processing). Packt Press.
- Yoon, T. I. (2019). Text Analysis Using Python. Nulbom Press.

Korean References Translated from the English

- 강현빈, 강주희, 류예나, 손지연, 윤별이, 이준성 (2017). 거주 후 평가(POE) 방법으로써 오피니언 마이닝(Opinion Mining)의 활용가능성에 관한 연구. 제17회 대한건축학회 추계학술발표대회 논문집. 37(2): 1144-1146.
- 김용학 (2003). 사회 연결망 분석. 서울:박영사.
- 박선주 (2015). R을 활용한 SW교육 텍스트데이터 토픽분석. 정보교육학회논문지. 19(4): 517-524.
- 배영미 (1998). 가정과 교육에서의 청소년문제 예방교육을 위한 기초연구: 청소년관련 신문기사분석을 통해 본 청소년 문화. 한국가정과교육학회지. 10(2): 131-144.
- 백영민 (2017). R을 이용한 텍스트 마이닝. 한울. pp. 171-178.
- 윤태일, 이수안 (2019). 파이썬으로 텍스트 분석하기. 늘봄출판사.
- 임동선 (2001). 한국어-영어 이중언어사용아동의 품사별 오류유형 분석. 연세대학교 대학원 언어병리학 협동과정 석사학위 논문.
- 정덕영, 이준석, 박상성 (2016). 워드클라우드를 이용한 기술트렌드 분석. 한국지능시스템학회 학술발표 논문집. 26(1): 17-18.
- 정병문, 김태환, 이진, 김정선 (2015). LDA 모델을 이용한 토픽 추출 및 토픽 카테고리 판단. 한국정보과학회 학술발표논문집. pp. 787-788.
- 조주연, 김현숙, 조민 (2018). 텍스트마이닝을 활용한 청소년 문제의 이슈 변화 분석. 부산대학교 교육발전연구소, 교육혁신연구. 28(4): 461-487.
- 황화상 (1998). 자연언어 처리를 위한 형태소 분석 방법론. 어문논집. 37: 439-440.