텍스트마이닝과 의미연결망을 활용한 재난 유형별 이슈 분석

Analysis of Issues Based on Disaster Type Using Text Mining and a Semantic Network

김태진* · 음미령** · 박상현***

Kim, Tae Jin*, Eum, Mi Ryeong**, and Park, Sang Hyun***

Abstract

Recently, the government has been increasingly communicating with the public in response to their opinions on state administration and policy projects. To examine the practicality of the public's suggestions, this study investigated issues by disaster type, based on information from major media channels and comment data from the news. An analysis of the frequency of appearance, text mining (TF-IDF, LDA, and sentiment analysis), and the semantic network was performed by extracting the comment data of articles on the themes of "disaster" and "evacuation," published from January 2010 to May 2020. The analysis results showed that news articles centered on these themes increased rapidly from 2017. The main disasters in Korea were those of "fire," "typhoon," "forest fire," "radioactivity," and "earthquake," in order of enormity. Of the total negative words pertaining to "radioactivity" disasters, 43% were negative-sentiment words, and the semantic network analysis revealed that the terms "typhoon," "forest fire," and "earthquake" were connected to "radioactivity" disasters. This study is meaningful as it identifies issues by type of disaster and factors of anxiety expressed by the public using news and comment data, without conducting surveys and interviews.

Key words : TF-IDF, LDA, Sentiment Analysis, Semantic Network

요 지

최근 정부는 국민과의 소통을 강조하면서 국정 운영 및 정책사업에 국민의 의견을 반영하고 있다. 실제 활용성이 더욱 커지면서 본 연구는 국내 주요 언론사의 뉴스와 언론사의 댓글 데이터를 바탕으로 재난 유형별 이슈를 살펴보았다. 2010년 1월부터 2020년 5월까지 보도된 뉴스 중 '재난'과 '대피'를 주제로 보도된 기사와 기사의 댓글 데이터를 추출하여 출현빈도, 텍스트마이닝(TF-IDF, LDA, 감성분석), 의미연결망 분석하였다. 분석의 결과 '재난'과 '대피'의 뉴스 기사가 2017년 이후 급격히 증가하고 있었으며, 우리나라의 주요 재난은 '화재', '태풍', '산불', '방사능', '지진' 순으로 나타났다. 재난 유형별 부정적 감성어는 '방사능' 재난이 전체의 43%에 해당하는 부정어가 출현하는 것을 확인하였으며, 의미연결망을 분석한 결과 '태풍', '산불', '지진' 단어에서 방사능 재난과 관련된 의미가 연결되는 부분을 확인하였다. 본 연구는 기존 설문조사와 인터뷰를 하지 않고 뉴스와 댓글 데이터를 활용하여 재난 유형별 이슈를 파악하고 국민들의 불안 요소를 확인할 수 있다는 점에서 의의가 있다.

핵심용어 : 역문서빈도, 잠재 디리클레 할당, 감성분석, 의미연결망

1. 서 론

매년 우리나라는 재난대응력과 회복력을 강화하는 사업

과 정책을 펼치고 있지만, 여전히 관리체계에 의구심이 있다는 국민적 평가와 언론사의 부정적인 발표가 잇따르고 있다. 이에 정부는 국민과의 소통을 강조하면서 국정 운영 및

*교신저자, 정회원, 국립재난안전연구원 안전연구실 연구원(Tel: +82-52-928-8127, Fax: +82-52-928-8149, E-mail: kimtj88@korea.kr)

Corresponding Author, Member, Researcher, Safety Research Division, National Disaster Management Research Institute

**국립재난안전연구원 안전연구실 연구원

Researcher, Safety Research Division, National Disaster Management Research Institute

***정회원, 국립재난안전연구원 안전연구실 공업연구관

Member, Senior Research Officer, Safety Research Division, National Disaster Management Research Institute

정책사업에 국민의 의견을 반영하고자 다양한 시도들이 시작되고 있다. 행정안전부는 2019년부터 국민수요 맞춤형 생활안전 연구개발사업을 매년 추진하는 계획을 발표하고 1차 공모를 통해 국민의 의견을 반영한 연구사업이 추진되고 있으며, 국민권익위는 1년 2개월(‘18.1.1.~‘19.2.28.)간 민원 정보분석시스템에 수집된 민원을 분석한 결과를 토대로 미세먼지 제도개선 방안을 마련하였다. 이처럼 국가의 정책을 수립하는 과정에서 국민의 의견을 반영하는 것은 국민적 신뢰를 고양할 수 있는 수단과 더불어 효과적인 정책을 시행할 수 있음을 확인할 수 있다. 이러한 장점으로 Lee (2015)는 국민행복의 구현이 핵심 국정과제 중 하나로 강조됨에 따라 국민이 무엇을 원하는지 등의 여론을 시의성 있고 정확하게 파악하기 위한 용도로 빅데이터를 활용해야 한다고 제시하였으며, Min and Shim (2020)은 국민 의견을 수집한 빅데이터를 바탕으로 우리나라 교육 분야별 문제점을 지적하고 시사점을 도출한 사례가 있다.

위에서 언급한 바와 같이 국민 의견의 중요성과 활용성이 커지면서 다양한 의견을 확인할 수 있는 수단으로 뉴스 기사와 뉴스 기사의 댓글 데이터를 주목하였다. 뉴스 댓글은 2002년부터 주요 언론사에서 제공된 서비스로 기사에서 전달하는 내용에 자신의 의견을 개진하며 의사소통이 가능한 공간으로 이용자의 적극적인 의사 표현과 상호작용 가능한 새로운 플랫폼으로 자리 잡았으며(Marlow, 2004), 댓글을 통해 사회적 이슈를 파악하고 기사만으로 알 수 없었던 정보를 제공하는 등 긍정적인 효과가 있음을 밝혔다(Lee and Lee, 2009). 본 연구는 이러한 효과를 바탕으로 뉴스와 댓글 데이터를 분석하여 재난 유형별 이슈를 파악하고자 하였다. 반면, 뉴스 댓글의 경우 누구나 쉽게 작성할 수 있는 자유로움으로 인해 기사와 무관한 악성 댓글, 광고성 스팸, 허위 사실 등이 반복적으로 나타나면서 부정적인 견해가 있음을 확인하였다(Lee and Kim, 2009).

그럼에도 댓글을 통해 의견을 표명하고 거리낌 없는 표현을 통해 중요한 속성을 파악하는 순기능을 확인하였다(Choi et al., 2008; Kim and Kim, 2005). 그간 국내에서 재난 관련 텍스트 분석의 학술발표 및 연구는 Table 1과 같이 6건 정도 있다. 6건 중 5건은 ‘재난대응 용어분석’, ‘재난상황메시지 분석’, ‘포항지진’, ‘미래 재난’, ‘재난 관련 감성도’와 같이 특정한 주제를 지정한 연구를 수행하였으며, 나머지 1건은 텍스트 기반 ‘재난재해 이벤트 발생’ 가능성을 예측하는 연구로 확인되었다. 본 연구와 유사한 주제로 연구를 수행한 Seo and Yoo (2017)의 경우 특정한 사건인 ‘옥시’, ‘문지마범죄’, ‘경주지진’을 중심으로 충격도를 분석하기 위한 목적으로 연구의 대상과 분석 기법에서 본 연구와는 다소 차이가 있음을 확인하였다.

2. 연구방법

2.1 자료수집 방법 및 수집결과

본 연구의 자료수집 대상은 Table 2와 같이 뉴스 기사와 댓글 데이터이다. 시간적 범위는 대형 언론사에서 댓글 기능을 제공하기 시작한 2010년 01월 01일부터 2020년 05월 05일까지 설정하였으며, 자료수집 방법으로는 빅카인즈(BIGKINDS)와 웹 크롤링(Web Crawling) 기술을 활용하였다. 빅카인즈는 한국언론진흥재단(Korea Press Foundation)에서 운영하는 시스템으로 54개의 주요 언론사의 뉴스 메타데이터를 활용하여 기사를 수집할 수 있으며, 웹크롤링은 웹에서 원하는 정보를 수집하는 기술로 댓글 데이터의 HTML 구조를 파악하고 프로그래밍 언어를 통해 자동으로 수집하였다.

2.2 분석 방법 및 절차

본 연구에서는 재난 유형별 이슈를 파악하기 위하여 텍스트마이닝 기법과 의미연결망 분석을 수행하였다. 텍스트마

Table 1. Literature Review

Thesis Title (Published Date)	Contributor
Analyzing Disaster Response Terminologies by Text Mining and Social Network Analysis (2016.03)	Kang et al. (2016)
Design and Implementation of Disaster Message Analysis Model through Text Analysis (2017.08)	Yoo et al. (2017)
Text based Event Detection Method using Word Embedding Model: For disaster events (2019.05)	Min et al. (2019)
A Review on Disaster Response through Critical Discourse Analysis of Newspaper Articles - Focused on the November 2017 Pohang Earthquake (2019.06)	Lee et al. (2019)
An Exploratory Approach on the Futures Disaster Keyword (2018.11)	Cho (2018)
Opinion Mining Analysis Using Disaster SNS Data (2017.05)	Seo and Yoo (2017)

Table 2. Data Collection Method

Collection Object	Temporal Scope	Collection Method
News Article	2010.01.01.~2020.05.05.	BIGKINDS
News Comment Data		Web Crawling

이닝은 텍스트를 분석하는데 널리 사용하는 방법으로 인간이 일상적으로 사용하는 자연어로 구성된 뉴스 기사, 책, SNS, 보고서 등 다양한 형태의 비정형 데이터에서 자연어 처리 기술(Natural Language Processing, NLP)을 활용하여 정보를 추출 및 가공한다. 텍스트마이닝 기법은 크게 단어, 토픽, 감성분석으로 나뉘는데 재난 유형별로 핵심어와 이슈가 어떻게 나타나는지를 확인하고자 단어와 토픽, 감성분석을 수행하였다. 단어 분석은 정보검색에 널리 사용되는 알고리즘인 TF-IDF 분석을 수행하였으며, 토픽 분석은 문서에서 내재된 토픽을 찾아내는 LDA 분석을 수행하였다.

TF-IDF 분석은 단어 빈도(Term Frequency, TF)와 역문서 빈도(Inverse Document Frequency, IDF)를 바탕으로 중요한 단어를 추출할 수 있으며, Eq. (1)과 같이 간단한 수식을 통하여 단어의 중요도에 따른 단어 분석을 수행할 수 있다. 여기서 t 는 특정 단어, f 는 단어의 출현 횟수, N 는 전체 문서의 수, n 는 특정단어(t)가 출현한 문서 수를 의미한다. TF-IDF는 N 개의 문서에서 특정 단어 t 가 f 번 출현할 경우 n 개의 출현한 문서를 계산하는 수식으로 $\log(\frac{N}{n})$ 는 t 가 출현한 문서가 많을수록 흔히 등장하는 용어라고 판단하여 중요도가 낮아진다.

$$TF-IDF = \frac{f}{t} \log\left(\frac{N}{n}\right) \quad (1)$$

Latent Dirichlet Allocation (LDA) 분석은 문서와 단어를 행렬로 구성하여 잠재(Latent)적인 주제(Topic)를 찾아내어 문서가 지닌 특징을 발견하기 위한 개념으로 Fig. 1과 같이 문서들이 다양한 토픽(Multiple Topics)으로 이루어져 있다고 가정하고 단어에서 드러나지 않는 내재적 토픽을 매개변수로 하여 특정 문서의 단어 빈도와 분포 등 단어의 생성과정을 확률로 추론하는 확률모델(Statistical Model)이다(Blei, 2012).

LDA 분석은 Eq. (2)의 수식으로 나타낼 수 있다. 여기서 D 는 말뭉치(Corpus) 공간에서 단어들의 집합을 의미하며, 말뭉치는 비정형 데이터를 컴퓨터가 인식할 수 있도록 구성

Fig. 1. The Intuitions Behind LDA

된 형태이다. K 는 하이퍼파라미터(Hyperparameter)에 해당하며, 문서 전체의 주제 수를 설정, N 은 문서의 공간이다. 하이퍼파라미터는 모델링 과정에서 사용자가 직접 설정하는 값을 뜻한다. $p(Z_{d,n} | \theta_d)$ 는 문서에서 나타내는 주제를 확률로 계산할 수 있으며, $p(W_{d,n} | \beta_{1:k}, Z_{d,n})$ 는 단어와 주제 사이의 출현 확률로 계산한다. 계산된 단어의 출현확률값은 말뭉치로부터 얻은 주제의 분포를 값으로 확인하여 중요한 주제를 추출하고, 추출된 주제에 대한 단어를 뽑는 과정이다.

$$p(\beta_K, \theta_D, Z_D, W_D) = \quad (2)$$

$$\prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \left\{ \prod_{n=1}^N p(Z_{d,n} | \theta_d) p(W_{d,n} | \beta_{1:k}, Z_{d,n}) \right\}$$

감성분석(Sentiment Analysis)은 텍스트에 등장하는 단어를 중심으로 감성이나 논조(Tone)를 분석하기 위해 감성사전을 활용한 방법으로 문서의 극성(Polarity)을 파악하는 것이 그 목적이다. 감성사전은 해당 단어가 보편적으로 어떤 감성을 표현하는지 미리 정해놓은 사전을 기반으로 다양한 방법론을 활용하여 분석한다. 다만, 감성에 관한 연구는 합의된 형태가 존재하지 않으며, 감성 정보를 추출하는 방식과 사용하는 감성사전에 따라 극성을 분류하는 것은 연구자의 관점에 따라 매우 다른 결과가 나타날 수 있다.

본 연구는 긍정과 부정 클래스에 속하는 단어를 추출하여 감성의 극성을 파악하려는 것이 아니라 재난 유형별 이슈를 파악하기 위한 목적으로 심리학에서 사용되는 부정적 감성 단어를 기반으로 출현빈도와 그 강세를 확인하고자 하였다. 이에, 심리학에서 주로 사용하는 Plutchik (1980)의 감성바퀴를 참고하였다. 플러치크는 Fig. 2와 같이 수레바퀴에 유사한 감성의 개념과 감성의 강도를 색상의 관계로 표현한

Fig. 2. Plutchik's Wheel of Emotions

고 있는데, 기본적인 감성 8가지를 배치하고 중심에서 멀어질수록 약한 감성이 배치되고 마주보는 감성은 반대로 대립하는 감성들이 배치된다. 플러칙은 총 15개의 부정적인 단어를 제시하고 있으나, 다소 함축된 단어로 구성되어 있어 댓글 데이터를 분석하기에는 다소 부족하다. 예를 들어 플러칙에서 제시하는 부정적 감성단어 “분노”, “격노”, “혐오”와 같은 명사만을 출현빈도로 분석하기보다는 “열받는다”, “어이없는”, “짜증난다”와 같은 구어체적인 단어가 추가되어 분석하는 것이 더 효과적인 방법이다. 이는 뉴스 기사와 달리 댓글 데이터는 대체로 짧은 문장으로 이루어져 있는 특성과 편의를 위해 명사보다는 목적격 조사를 생략하거나 어휘를 축약하는 경우가 많기 때문이다.

플러칙에서 제시하는 감성은 강도를 기반으로 공학적 처리가 쉬운 장점으로 감성분석에 활용되고 있으며(Yang et al., 2014; Fang et al., 2018), 감성을 일부 수정하는 연구(Yoon and Kwon, 2010)와 감성어에 대한 극성을 분류하는 연구(Chang, 2012) 등 활용도를 높이기 위한 목적으로 많은 응용 연구가 존재하지만, 단어의 극성을 분류하는 것은 연구자의 관점에 따라 달라지기 때문에 기존 문헌을 인용하여 분석하기에는 다소 어려움이 있다. 본 연구에서는 Table 3과 같이 플러칙의 부정적 감성 단어를 기반으로 유사한 명사, 형용사, 동사를 구성하여 탐색적 이슈를 분석하고자 하며, Table 3은 플러칙의 부정적 감성 단어들 사이에서 출현빈도가 10,000개 이하로 나타나는 단어는 탈락시켰다. 실제 “격노”와 같은 단어는 분노와 화남을 뜻하지만, 온라인 상에서는 주로 사용되지 않아 분석하기에는 다소 부족하기 때문이다.

의미연결망은 단어와 단어 사이의 상관관계를 분석하여 어떻게 연결되어 있는지를 시각적으로 확인할 수 있는 분석 기법으로 직관적인 데이터 분석을 위하여 수행하였다. 본 연구에서는 피어슨 상관계수를 활용하여 단어와의 관계와 거리를 선형관계로 나타내었다. 의미연결망의 경우 텍스트 분석 기법으로 널리 알려진 방법론으로 이론의 경우 기존 문헌으로 대체하고자 분석과정 위주로 설명하였다.

본 연구의 분석 절차로는 Fig. 3과 같다. 먼저 빅카인즈를 활용하여 수집된 뉴스 기사를 전처리와 분석 단계를 수행하였다. 이 과정에서 주요 재난 유형을 선정하고 선정된 재난 유형을 기준으로 댓글 데이터를 최종적으로 수집하였다.

수집된 댓글 데이터는 기초적인 텍스트마이닝 분석, LDA, 감성분석을 수행하여 재난 유형별 출현빈도와 강세를 확인하였다. 분석을 수행하는 도구로는 프로그래밍 언어인 파이썬(Python)을 기반으로 자연어 분석 라이브러리 Konlpy, Scikit-learn, Gensim을 활용하였다. 여기서 Konlpy는 문장을 형태소(명사, 조사, 동사, 부사)로 분리하여 기초분석을 수행할 수 있으며, Scikit-learn은 TF-IDF 분석을 위한 머신러닝 라이브러리, Gensim은 LDA 분석을 수행할 수 있는 라이브러리이다. 추가로 LDA 분석결과를 시각적으로 확인하기 위한 라이브러리 pyLDAvis를 사용하였다.

3. 데이터 분석

3.1 뉴스 데이터를 활용한 재난 유형 선정

분석에 앞서 재난 유형이 다양함에 따라 모든 유형의 재난을 고려하기에는 현실적으로 어려운 한계가 있으며,

Fig. 3. Analysis Process

Table 3. Negative Sentimental Words

Categorize	Word
Fearful & Angry	Fear, Fearful, Danger, Dangerous, Annoy, Annoyance, Angry, Anger, Aghast, Incompetent, Incompetence
Anxiety & Complaint	Concern, Concerned, Worry, Anxiety, Anxious, Nerve, Nervous, Discontent, Dissatisfy, Discontent, Problem, Problematic, Degenerate, Degenerative
Negative	Illegal, Illegally, Illegality, Deterioration, Deteriorate, Deteriorative, Lack, Lacking, Ridicule, Ridiculous, Fault, Faulty, Lackadaisical

어떠한 재난 유형을 분석 대상으로 선택하는지는 연구자마다 중요도가 다르다. 이러한 선택적 분석에 있어서 우리나라의 주요 재난은 뉴스 기사에 자주 나타날 것이라는 전제로 뉴스 기사에서 분석의 대상을 선정하고자 한다.

재난 유형별로 발생한 이슈와 상황, 문제점을 파악하고자 ‘재난’, ‘대피’를 검색어로 정하였으며, 빅카인즈에서 제공하는 뉴스 메타데이터에서 ‘재난’, ‘대피’ 검색어 기반의 뉴스 기사를 1차적으로 추출한 결과 총 491,311건(‘재난’ 349,635건, ‘대피’ 141,676건)의 뉴스 데이터를 확인하였다. 수집 결과 ‘재난’, ‘대피’ 뉴스 기사의 건수 추이를 살펴보면, Fig. 4와 같이 매년 소폭 증가하는 추세로 ‘재난’ 관련 뉴스의 경우 2020년이 급격히 증가한 것으로 확인되었으며, ‘대피’ 관련 뉴스의 경우에는 2019년이 가장 높게 나타났다.

앞서 수집한 뉴스 기사는 재난 유형을 선정하기 위하여 전처리의 과정이 필요하다. 전처리 과정에서는 뉴스의 기사를 형태소로 분리하여 대량의 텍스트로 변환하고 분석에 필요하지 않은 불용어를 삭제하는 과정이다. 이 과정에서 491,311건의 뉴스 기사를 대상으로 총 76,988,093개의 사용 단어(중복 포함)를 수집하였다. 수집된 단어를 바탕으로 재난 유형을 선정하기 위하여 TF-IDF 분석을 하였다. TF-IDF는 문서에서 특정 단어가 비정상적으로 많이 출현하는 경우 문서의 가중치를 역으로 낮추면서 단순 고빈도로는 단어의 중요도를 비교하기 어려운 문제에서 해당 단어가 얼마만큼 중요한지를 수치로 비교할 수 있는 장점이 있다. 한 예로 모든 문서에 ‘오늘’이라는 단어가 자주 나타나게 된다면, 빈도 기준으로는 중요한 단어로 해석할 수 있지만, TF-IDF의 경우 전체 문서의 비중을 계산하여 흔히 나타나게 되는 단어일 경우 가중치를 낮추어 탈락한다. 반대로 ‘누출’, ‘불씨’와 같이 재난 유형에 따라 특정하게 나타나는 단어의 경우 단어의 빈도는 전체적으로 낮지만, 재난 유형에 따라 중요하다고 판단하게 된다. Table 4는 분석의 결과로 TF-IDF 값의 경우 ‘화재’, ‘태풍’, ‘산불’, ‘방사능’, ‘지진’ 등의 순으로 나타났으며, 빈도의 경우 ‘화재’, ‘지진’, ‘태풍’, ‘산불’, ‘방사능’ 등의 순으로 TF-IDF와 빈도의 순위가 다소 다른 결과를 보여주었다.

Fig. 4. Disaster and Evacuation News Frequency

3.2 재난 유형별 탐색적 단어 분석

본 연구에서는 뉴스 기사의 TF-IDF 값을 기준으로 5개의 재난 유형을 선정(‘화재’, ‘태풍’, ‘산불’, ‘방사능’, ‘지진’)하였으며, 댓글 데이터를 기준으로 LDA 분석과 감성어 출현빈도를 확인하였다. 재난 유형별 댓글 데이터는 총 11,445,524건으로 Fig. 5를 살펴보면, 2017년 이후로 급격히 증가하는 추세를 확인할 수 있다.

3.2.1 LDA를 활용한 토픽모델링 분석

LDA 분석은 Fig. 6과 같이 2가지의 결과를 확인할 수

Table 4. Word Frequency and TF-IDF Value

Rank	Disaster	TF-IDF Value	Count
1	Fire	0.0624	71,466
2	Typhoon	0.0218	23,739
3	Forest Fire	0.0184	13,860
4	Radioactivity	0.0179	13,480
5	Earthquake	0.0151	44,567
6	Intense Heat	0.0122	7,101
7	Landslide	0.0112	5,505
8	Explosion	0.0109	11,492
9	Flood	0.0109	7,863
10	Heavy Rain	0.0104	12,565

Fig. 5. Frequency of News Comment by Disaster Type

Fig. 6. LDA Analysis Visualization

있는데 왼쪽의 경우 Intertopic Distance Map (IDM)으로 단어를 밀집 벡터(Vector)의 형태로 표현하는 워드 임베딩(Word Embedding) 방식으로 토픽의 군집 형성을 시각적으로 확인할 수 있으며, 오른쪽의 경우 토픽별로 단어가 형성되

는 확률을 계산하여 그 값과 토픽을 형성하는 단어를 확인할 수 있다.

재난 유형별 대표적인 3개의 토픽을 살펴보면 Table 5와 같다. 화재의 Topic 1은 ‘안전’, ‘훈련’, ‘대응’ 등 화재와

Table 5. LDA Analysis for Disaster (Ranked 3rd by Probability)

Disaster	Topic	Key Word
Fire	Topic 1 (14.8%)	Safety, Support, Disaster, Training, Business, Plan, Operation, Response, Fire, Situation, Promotion, Area, House, Seoul City, Facility, Household, Living, Provision, Manamenet, Preparation, Institution, Strengthening, Install, Restore, Participation, Experience, Progress, Prevent
	Topic 2 (13.4%)	Fire, Occurrence, Accident, Cause, Damage, Extinguishing, Building, Investigation, Fire Department, Explosion, Evacuation, Smoke, Vehicle, Factory, Site, Work, Flames, Police, Authorities, Personnel Damage, Estimation, Fire truck, Seoul City, Fire station, Extinguishing
	Topic 3 (9.5%)	Fire, Safety, Installation, Accident, Inspection, Occurrence, Facility, Management, Building, Firefighting, Construction, Using, On-site, Market, Result, Gas, Risk, Check, Sprinkler, Electricity, Indicate, Official, Building, Jecheon City, Operation, Sejong hospital, Measures, Architecture, Investigation
Typhoon	Topic 1 (15.3%)	Damage, Area, Business, Disaster, Typhoon, Outbreak, Kompas, Safety, Preparation, Construction, Plan, Sejong City, Countermeasure, Site, Management, Install, Facility, Promotion, Prevention, Response, Residents, Situation, Disaster, Inspection, Restoration, Flooding, Earthquake, Operation, River
	Topic 2 (10.6%)	Typhoon, Impact, Meteorological Administration, Forecast, Seoul City, Nationwide, Dianmu, Jeju island, Northward, Prospect, Province, Korean peninsula, Region, Temperature, Weather, Forecast, Strong wind, Average year, Precipitation, Highest, Korea, Occurrence, Possibility, Gwangju City, Deagu City, Wind, South
	Topic 3 (8.2%)	Sea, Marine, Village, Wind, Busan City, Tree, Jeju City, Strong wind, Nakdong River, Velocity per second, Typhoon, Wave, Ship, Sinking, Coast, South, Fishing boat, Street tree, Breakwater, Seogwipo, West coast, Location, Sea area, Tsunami
Forest Fire	Topic 1 (25.5%)	Forest fire, Outbreak, Prevention, Forest, Prevention, Extinguishing, Spring season, Forest Service, Period, Activity, Area, Operation, Countermeasure, Monitoring, Fire, Incineration, Forest fire prevention, Install, Reinforce, Plan, Authorities, Risk, Public officials, Deployment, Campaign, Villager, Field, Village, Concentration, Watchman
	Topic 2 (22.1%)	Project, Region, Support, Promotion, Job, Disaster, Plan, Management, Field, Safety, Response, Economy, Creation, Budget, Market, Expansion, Provision, Facility, Government, Education, Growth, Creation, Work, Operation, Institution, Development, Nationwide, Society
	Topic 3 (14.9%)	Weather, Wildfire, Temperature, Seoul City, Nationwide, Area, Wind, Fire, Region, Outbreak, Dry, Predicted, Subzero, Impact, Evolution, Neighborhood, Caution, Daegu City, Weekend, Meteorological Administration, East coast, Hillside, Cloud, View, Dry war ning, Cold, Ulsan City, Estimate
Radio activity	Topic 1 (13.8%)	Government, Conference, President, People, Leader, National, Legislator, Policy, Assertion, Security, Indicate, Discussion, Strengthening, Emphasis, Safety, Demand, Korea International, Committee, Candidate, Representative, Information, Plan, Opinion, Minister, Policy
	Topic 2 (10.9%)	Nuclear power plant, Accident, Fukushima City, Radioactivity, Japan, Outbreak, Breakdown, Nuclear power, Government, Spill, Leak, Stop, Chernobyl City, Power plant, Groundwater, Pollution, Situation, Residents, Safety, Tsunami, Danger, Damage, Nuclear accident, Pollution water, Tokyo electric power, Earthquake, Sea, Buildings
	Topic 3 (7.6%)	Radioactivity, Substances, Radiation, Investigation, Result, Detection, Uranium, Measurement, Cesium, Environment, Check, Pollution, Standards, Inspection, Radioactive, Impact, Level, Underground, Safety, Region, Install, Detention, Exposure, Authorities, Korea, Concentration
Earthquake	Topic 1 (13.8%)	Disaster, Safety, Earthquake, Response, Training, Facility, Management, Plan, Preparedness, Occurrence, Situation, Countermeasure, Education, Promotion, Strengthening, Institution, Arrangement, Fire, Install, Target, Operation, Region, School, Government, Equipment, National, Field, Information, Fire department
	Topic 2 (10.9%)	Earthquake, Occurrence, Scale, Strong earthquake, Damage, Chile, Area, Building, Aftershock, Meteorological Administration, Rescue, Debris, Warning, Neighborhood, Confirmation, Situation, Observation, Estimation, United States, Contact, Point, Geology, Local time, Vibration, Overseas korean, Life, Detection
	Topic 3 (7.6%)	Haiti, Support, Relief, Earthquake, Delivery, Activity, Donation, Damage, Emergency, Recovery, Fundraising, Dollar, Local, Reconstruction, Catastrophe, help, Participation, Social, Aid kit, International, Visit, Scheduled, Pain, Hope, Child, Plan, Medical

관련된 대비와 대응 방안과 관련된 토픽이 형성되는 것으로 해석할 수 있으며, Topic 2는 ‘건물’, ‘대피’, ‘서울’ 등 도심지역에서 발생하는 화재와 관련된 토픽이 형성되는 것으로 해석된다. Topic 3은 ‘시설’, ‘건물’, ‘스프링쿨러’ 등 화재의 원인과 관련된 토픽이 형성되는 것으로 해석할 수 있다. 태풍의 Topic 1은 ‘피해’, ‘대비’, ‘복구’ 등 태풍과 관련된 대비와 예방, 피해와 관련된 토픽이 형성되는 것으로 해석할 수 있으며, Topic 2는 ‘기상청’, ‘예보’, ‘강수량’ 등 기상과 관련된 토픽이 형성되는 것으로 해석된다. Topic 3은 ‘바다’, ‘강풍’, ‘해일’ 등 해안을 중심으로 태풍의 피해와 관련된 토픽이 형성되는 것으로 해석된다. 산불의 Topic 1은 ‘진화’, ‘산림청’, ‘대책’ 등 산불 발생, 진화, 예방과 관련된 토픽이 형성되는 것으로 해석할 수 있으며, Topic 2는 ‘사업’, ‘재난’, ‘경제’ 등 산불과 관련된 사업과 경제적 토픽이 형성되는 것으로 해석된다. Topic 3은 ‘날씨’, ‘바람’, ‘건조’ 등 날씨와 산불과의 관련된 연관 토픽이 형성되는 것으로 해석할 수 있다. 방사능의 Topic 1은 ‘국가’, ‘안보’, ‘정책’ 등 핵 안보와 관련된 토픽이 형성되는 것으로 해석할 수 있으며, Topic 2는 ‘원전’, ‘고장’, ‘쓰나미’ 등 원전 사고와 관련된 토픽이 형성되는 것으로 해석된다. Topic 3은 ‘우라늄’, ‘세슘’, ‘오염’ 등 원전사고 발생시 위험요소와 관련된 토픽이 형성되는 것으로 해석할 수 있다. 지진의 Topic 1은 ‘안전’, ‘훈련’, ‘대비’ 등 지진과 관련된 재난대응 훈련 토픽이 형성되는 것으로 해석할 수 있으며, Topic 2는 ‘강진’, ‘칠레’, ‘교민’ 등 해외에서 발생한 지진과 관련된 토픽이 형성되는 것으로 해석된다. Topic 3은 ‘구호’, ‘성금’, ‘복구’ 등 구호 활동과 관련된 토픽이 형성되는 것으로 해석할 수 있다.

3.2.2 재난 유형별 재난대피 이슈 분석

앞서 탐색적 단어 분석과 토픽모델링 분석을 기초로 재난 유형별 재난대피 이슈를 파악하고자 감성분석과 의미연결망 분석을 수행하였다. 재난 유형별 부정적 감성어 출현빈도를 살펴보면 11,445,524건의 댓글 데이터에서 부정적 감성어는 Table 6과 같이 총 904,254개(7.9%)의 부정적 감성어를 확인하였다. 부정적 감성어는 불안과 불만을 뜻하는 감성어가 394,465개(43.6%)로 다른 부정적 감성어와 비교하여 높게 출현하는 특징을 확인할 수 있었다. 재난 유형별 부정적 감성어 빈도를 살펴보면 방사능이 386,980개(42.8%)로 가장 많이 나타났으며, 2순위는 지진 223,699개(24.7%), 3순위는 화재 157,061개(17.4%), 4순위는 태풍 95,101개(10.5%), 5순위는 산불 41,413개(4.6%)로 확인되었다.

재난 유형별 부정적 감성어가 가장 많이 출현한 구간의 이슈를 파악하기 위하여 부정적 감성어 출현빈도를 월별로 구성하고 가장 높은 구간 3개의 Point를 선정하여 뉴스를 분석하였다. ‘화재’ 관련 부정적 감성어 출현빈도는 Fig. 7와 같이 특정 사건에 의해 빈도가 높은 추세를 확인할 수 있다. 가장 높은 Point 1은 2017년 12월로 총 23,094개의

부정적 감성어를 확인하였으며, Point 2는 총 12,740개(2018년 1월), Point 3은 총 7,329개(2016년 8월)를 확인하였다. Point 1은 충북 제천 스포츠센터 참사와 관련된 내용, Point 2는 밀양 세종병원 화재와 관련된 내용, Point 3은 특정 차량의 화재와 관련된 내용을 파악하였다.

부정적 감성어가 가장 높은 Point 기간에 작성된 댓글 데이터를 추출하여 대중들의 이슈를 확인하고자 의미연결망 분석하였다. 의미연결망을 살펴보면 Fig. 8과 같이 ‘소방’ 단어가 가장 높은 연결성으로 중심적인 단어임을 확인할 수 있으며, ‘화재’, ‘건물’ 단어와 연결되어 의미연결망이 형성되었다.

Table 6. Negative Sentiment Analysis Result

Disaster	Categorize	Rank	Count	Ratio
Fire	Fearful & Angry	8	46,415	5.1%
	Anxiety & Complaint	6	59,254	6.6%
	Negative	7	51,392	5.7%
	Sum		157,061	17.4%
Typhoon	Fearful & Angry	11	31,411	3.5%
	Anxiety & Complaint	10	42,421	4.7%
	Negative	12	21,269	2.4%
	Sum		95,101	10.5%
Forest Fire	Fearful & Angry	14	14,811	1.6%
	Anxiety & Complaint	13	17,757	2.0%
	Negative	15	8,845	1.0%
	Sum		41,413	4.6%
Radio activity	Fearful & Angry	2	143,328	15.9%
	Anxiety & Complaint	1	172,912	19.1%
	Negative	5	70,740	7.8%
	Sum		386,980	42.8%
Earthquake	Fearful & Angry	4	77,500	8.6%
	Anxiety & Complaint	3	102,121	11.3%
	Negative	9	44,078	4.9%
	Sum		223,699	24.7%
Sum			904,254	100%

Fig. 7. Negative Sentimental Word (Fire)

연결망이 가장 높은 ‘소방’ 단어를 중심으로 의미를 살펴 보면 Fig. 9와 같이 ‘장비-인력-지방-부족’, ‘예산-증원’과 ‘제천’과 ‘밀양’ 화재사고와 관련되어 ‘초기-대응’, ‘책임자-대처-필요-처벌’ 단어가 연결되어 소방 활동의 문제와 책임자의 대처 능력에 대한 이슈를 확인하였으며, 이는 소방력 강화와 시설관리 담당자의 화재 관련 대처방안 및 교육의 필요성을 대중들이 요구하고 있음을 유추할 수 있다.

‘태풍’ 관련 부정적 감성어 출현빈도는 Fig. 10과 같이 매년 소폭 상승하는 추세로 장마와 태풍철에 급증되는 것으로 확인되었다. Point 1은 2019년 10월로 총 3,227개의 부정적 감성어를 확인하였으며, Point 2는 총 2,870개(2018년 8월), Point 3은 총 2,716개(2017년 5월)를 확인하였다. Point 1은 태풍 미탁과 관련하여 제주, 전남의 피해, Point 2는

Fig. 8. Semantic Network (Fire)

Fig. 9. 'Fire Fighting' Word-Centered Network

태풍 솔릭과 제주, 서해, 목포, 광주의 피해와 관련된 내용을 파악하였으며, Point 3은 태풍으로 인해 울산지역의 집중호우가 발생한 내용을 확인하였다.

‘태풍’의 의미연결망은 Fig. 11과 같이 ‘태풍’ 단어가 가장 높은 연결성으로 중심적인 단어임을 확인할 수 있으며, ‘일본’ 단어와 연결되어 파생적인 의미가 연결되는 의미연결망이 형성되었다.

연결망이 가장 높은 ‘태풍’ 단어를 중심으로 의미를 살펴 보면 Fig. 12와 같이 ‘일본’의 경우 ‘싫다-지진-불매’ 단어와 같이 재난과 연관된 내용보다는 정치적 발언에 대한 이슈를 확인하였으며, ‘방사능-오염-방류’ 단어가 연결되어 방사능 오염에 대한 이슈를 확인하였다.

‘산불’ 관련 부정적 감성어 출현빈도는 Fig. 13과 같이 17년 이후로 그 빈도와 강세가 급격히 증가하는 것으로 나타났다. Point 1은 2019년 4월로 총 5,568개의 부정적 감성어를 확인하였으며, Point 2는 총 1,979개(2017년 5월),

Fig. 10. Negative Sentimental Word (Typhoon)

Fig. 11. Semantic Network (Typhoon)

‘방사능’ 관련 부정적 감성어 출현빈도는 Fig. 16과 같이 빈도와 강도가 급격히 증가하는 것으로 확인되었다. Point

Fig. 15. ‘People’, ‘Forest Fire’ Word-Centered Network

관한 내용, Point 2는 한빛원전 원자로 정지 및 방사성 물질 누출 사고, 방사능 방재훈련 관련 서면 처리 이슈, Point 3은 일본산 수입 식품에서 방사성 물질이 검출된 내용을 확인하였다.

‘방사능’의 의미연결망은 Fig. 17과 같이 ‘원전’ 단어가 가장 높은 연결성으로 중심적인 단어임을 확인할 수 있으며, ‘일본’, ‘국민’ 단어와 연결되어 2개의 군집된 의미연결망이 형성된 것을 확인할 수 있다. 연결망의 특징으로는 3가지 토픽이 시각적으로 나뉘는 것을 확인할 수 있으며, 실제 ‘원전’, ‘일본’, ‘국민’ 단어 사이의 거리가 있어 내용적인 측면에서 차이가 있음을 유추할 수 있다.

연결망이 가장 높은 ‘원전’ 단어를 중심으로 의미를 살펴 보면 Fig. 18과 같이 ‘탈원전’, ‘전기’ 단어가 연결되어 파생적인 연결망 구성을 확인하였다. ‘탈원전’의 경우 정책에 대한 찬반 의사와 관련된 내용 일부와 ‘사고-위험-지진’ 단어가 연결되어 방사능 재난에 대한 이슈를 확인하였으며, ‘전기’의 경우 ‘요금-에너지-신재생’ 단어가 연결되어 에너지 정책

Fig. 17. Semantic Network (Radioactivity)

Fig. 18. 'Nuclear Facility' Word-Centered Network

과 관련된 내용을 파악하였다. ‘걱정-무섭다-죽다’ 단어와 같이 일부 부정적인 의미는 ‘중국-후쿠시마-위험-사고-신고리’ 단어가 연결되어 국제적인 원전 이슈와 연결되는 것을 확인하였다.

‘지진’ 관련 부정적 감성어 출현빈도는 Fig. 19와 같이 빈도와 강도는 실제 지진이 발생한 시점으로 급격히 증가하였다. Point 1은 2016년 9월로 총 100,733개(2016년 9월), Point 2는 총 97,215개(2017년 11월), Point 3은 총 76,810개(2018년 6월)를 확인하였다. Point 1은 경주 지진으로 인하여 불안감과 피해와 관련된 내용, Point 2는 포항 지진과 여진에 관한 내용, Point 3은 경북 의성, 포항 및 동해안 일대 동시 지진 발생으로 인한 피해 내용을 확인하였다.

‘지진’의 의미연결망은 Fig. 20과 같이 ‘지진’ 단어가 높은 연결성으로 중심적인 단어임을 확인할 수 있으며, ‘국민’, ‘원전’, ‘일본’ 단어와 연결되어 의미연결망이 형성되었다.

연결망이 가장 높은 ‘지진’ 단어를 중심으로 의미를 살펴 보면 Fig. 21과 같이 ‘일본’의 경우 ‘한반도-지진-대비-대피’ 단어가 연결되어 일본 지진과 관련하여 한반도에서 발생하는 지진을 대비해야 하는 이슈를 확인하였으며, ‘국민’의 경우 ‘우리나라-재난-대비-불안’ 단어가 연결되어 지진 재난

Fig. 19. Graph of Negative Sentimental Word (Earthquake)

Fig. 20. Semantic Network (Earthquake)

Fig. 21. ‘Earthquake’ Word-Centered Network

과 관련하여 대비와 대응력에 대한 걱정하는 이슈, '나라'의 경우 '내진-부설-사고-원자력-원전-중단' 단어가 연결되어 지진에 대한 내진과 원전 중단으로 인한 사고에 관한 내용을 확인하였다.

4. 결론

일반적으로 특정 주제를 대상으로 국민 인식 수준을 파악하기 위해서는 그간 설문지나 인터뷰를 통하여 분석하였으나, 다양한 집단의 생각과 이슈를 파악하기가 쉽지 않았다. 이에 본 연구에서는 국민이 생각하는 재난 유형별 이슈를 살펴보고자 뉴스와 뉴스 댓글 데이터를 분석하여 텍스트마이닝과 의미연결망 분석을 수행하였으며, 다음과 같은 결과를 확인하였다.

첫째, 2017년 이후 ‘재난’과 ‘대피’ 관련 뉴스 건수가 급격히 증가하는 것을 확인할 수 있었으며, 이는 각종 재난과 사고의 이슈가 빈번해지고 있음을 의미한다.

둘째, 텍스트마이닝 분석을 수행한 결과 우리나라에서 발생하는 주요 재난의 경우 ‘화재’, ‘태풍’, ‘산불’, ‘방사능’, ‘지진’의 순으로 나타났으며, 재난 유형별 토픽의 경우 ‘화재’와 ‘지진’은 재난훈련과 안전에 대한 토픽이 형성되었으며, ‘태풍’과 ‘산불’은 대비와 예방에 대한 토픽이 형성되었다. 반면, ‘방사능’의 경우 원자력 정책과 후쿠시마 사고와 관련된 토픽이 형성되어 타 재난과 비교하여 다소 다른 토픽이 형성됨을 알 수 있었다.

셋째, 최근 3년 동안 모든 재난 유형에서 부정적으로 인식되는 감성이 출현빈도가 큰 폭으로 증가하는 추세를 확인하였으며, 이 중에서도 방사능 재난이 전체의 약 43%의 해당하여 가장 부정적인 재난 유형으로 확인되었다.

넷째, 재난 유형별 의미연결망 분석 결과 ‘태풍’, ‘산불’, ‘지진’ 재난에서 방사능 재난과 관련된 부정적 의미연결망을 확인하였으며, 이는 방사능과 관련된 국민의 관심이 증가하고 있음을 시사한다.

결론적으로 최근 발생하는 재난과 관련된 뉴스 건수가 증가하고 있으며, 이 중에서도 ‘방사능’ 재난 관련 부정적 감성어와 출현빈도가 가장 큰 것으로 확인되었다. 또한, 원자력 정책과 사고에 대한 뉴스 토픽이 형성되었음에도 실제 댓글 데이터에서는 복합재난의 가능성과 큰 피해에 대한 ‘걱정’, ‘무섭다’, ‘죽다’와 같이 우려하는 의미의 연결망이 형성되었다. 즉, 재난과 관련 없는 원전 정책 뉴스 기사에서도 국민은 재난과 사고를 걱정하고 불안한 감정의 댓글을 작성하고 있었으며, 다른 재난 유형에서도 이러한 특징이 나타났다. 이러한 분석을 바탕으로 재난 유형별 국민이 염려하는 이슈를 찾아내어 해결하는 문제해결형 재난안전 사업을 시행하거나 정책 의사결정 과정에서 시사점을 제공할 수 있다고 판단된다. 또한, 사회 전반에 확인되지 않은 허위정보 및 가짜뉴스 확산 문제를 해결하는 방안으로 키워드별 이슈를 분석하여 허위정보 유무를 확인하고 선제적인 대응을 통하여 허위정보 확산 방지 및 잠재적 피해를 예방하는 방법으로도 활용할 수 있다고 판단된다.

다만 본 연구는 구조화되지 않은 다양한 의견들 사이에서 유용한 정보를 추출할 수 있는 새로운 방법론을 제시하고 있으나, 연구자의 디지털 문해력의 차이에 따라 일부 사람들의 의견이 편향되게 반영될 수 있다는 점과 언론사의 성향이나 성향에 따른 댓글 데이터가 분석 결과에 영향을 미친다는 점을 인지하였다. 이러한 단점을 보완할 수 있는 분석 기법이나 연구 방법론이 추가로 진행될 필요가 있다.

References

- Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, Vol. 55, No. 4) pp. 77-84.
- Chang, M.S. (2012). Empirical Sentiment Classification Using Psychological Emotions and Social Web Data. *Journal of Korean Institute of Intelligent Systems*, Vol. 22, No. 5, pp. 563-569.
- Cho, W.Y. (2018). An Exploratory Approach on the Futures Disaster Keyword. *Journal of the Korea Association for Futures Studies*, No. 2, pp. 169-174.
- Choi, D.S., Choi, S.E., and Choi, Y.J. (2008). How do replies affect forming public opinion in Internet portal news. *Journal of Political Communication*, Vol. 8, pp. 309-356.
- Fang, X., Wang, G., and Liu, Q. (2018). Text sentiment analysis based on multi-granularity joint solution. *Proceedings of the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis*, pp. 315-321.
- Kang, S.K., Yu, H., and Lee, Y.J. (2016). Analyzing Disaster Response Terminologies by Text Mining and Social Network Analysis. *Information Systems Review*, Vol.

- 18, No. 1, pp. 141-155.
- Kim, K.N., and Kim, J.Y. (2005). Analyzing Readers' Comments in Internet Newspaper Ohmynews. *Korean Journal of Broadcasting and Telecommunication Studies*, Vol. 19, No. 3, pp. 7-41.
- Korea Press Foundation. 'Disaster', 'Evacuation' keywords search in 2010~2020. www.bigkinds.or.kr
- Lee, C.H., and Lee, H.Y. (2009). The Pattern of Portal News Use among Portal Users and Their Recognition of Portal as a Press. *Korean Association For Communication & Information Studies*, pp. 177-211.
- Lee, K.J. (2015). Presenting the starting point for using big data to establish regional policies tailored to national happiness. *National territory*, pp. 150-151.
- Lee, K.S., and Kim, B.S. (2009). A study of a type of communication and speech acts to be involved to postings in The Internet Forum and Discussion. *Discourse and Policy in Social Science*, pp. 25-50.
- Lee, Y.S., Jeon, H.S., Lee, K.M., Min, B.H., and Choi, Y.S. (2019). A Review on Disaster Response through Critical Discourse Analysis of Newspaper Articles - Focused on the November 2017 Pohang Earthquake. *Journal of the Korea Society of Disaster Information*, Vol. 15, No. 2, pp. 223-238.
- Marlow, C. (2004). Audience, structure and authority in the weblog community. *In International communication association conference*, Vol. 27, pp. 1-9.
- Min, J.W., and Shim, J.K. (2020). A Study on Analysis of National Petition Data for Deriving Current Issues in Education. *Journal of Creative Information Culture (JCIC)*, Vol. 6, No. 2, pp. 57-64.
- Min, K.H., Lee, J.S., Yu, K.Y., and Kim, J.Y. (2019). Text based Event Detection Method using Word Embedding Model: For disaster events. *Proceedings of Korean Society for Geospatial Information Science*, pp. 109-110.
- Plutchik, R. (1980). *Emotion: Theory, research, and experience, Theories of emotion, vol. 1*. New York: Academic.
- Seo, M.S., and Yoo, H.H. (2017). Opinion Mining Analysis Using Disaster SNS Data. *Proceedings of Korean Society for Geospatial Information Science*, Vol. 2017, No. 5, pp. 3-6.
- Yang, M., Peng, B., Chen, Z., Zhu, D., and Chow, K.P. (2014). A topic model for building fine-grained domain-specific emotion lexicon. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 421-426.
- Yoon, A.S., and Kwon, H.C. (2010). Component Analysis for Constructing an Emotion Ontology. *Korean Journal of Cognitive Science*, Vol. 21, No. 1, pp. 157-175.
- Yoo, E.J., Yang, M.S., and Shim, H.S. (2017). Design and Implementation of Disaster Message Analysis Model through Text Analysis. *Korea Intelligent Information Systems Society*, pp. 62-63.

Received	May 25, 2021
Revised	May 26, 2021
Accepted	June 3, 2021