Journal of the Korean Data Analysis Society (June 2020), 22(3), 999-1011.
https://doi.org/10.37727/jkdas.2020.22.3.999

텍스트 분석을 이용한 청원데이터의 주제 및 감성에 관한 연구

서혜선[1], 소진수[2]

요 약

본 연구는 국민들의 관심사와 참여가 많은 청와대 국민청원 게시글에 대한 텍스트 분석과 토픽 모델링을 실시하는데 있다. 17개로 분류되어져 있는 국민청원데이터는 개설된 이래 60여만 개의 게시글이 올라와 빅데이터 수준으로 축적되어 있는 상황이다. 이러한 청원 게시글에 대한 최초의 17개 분류가 적정한지를 텍스트 분석 기반으로 재분류해 보고자 한다. 이를 통해 국민청원 게시글에 대한 새로운 토픽(분류, 주제)들을 제안하고 단어들의 긍정어, 부정어 등을 고려한 감성 분석을 통해 각 토픽별 감성의 수준을 제시하고자 한다. 또한 선정된 토픽별로 청원글에 대한 국민참여수가 달라지는지, 그리고 각 청원글에 포함된 단어들의 감성 수준에 따라 국민참여수가 달라지는가를 연구하기 위해 일반화선형모형을 적용한 분석을 실시한다. 분석 결과 토픽과 감성에 따라 청원 게시글에 대한 국민동참의 참여도가 통계적 유의성하에서 달라짐을 확인하였다. 또한 청원 게시글에 참여를 올리는 데는 부정어의 사용보다 긍정어의 사용이 더 효과적이라는 사실을 확인할 수 있었다.

주요용어 : 토픽 모델링, 감성 분석, 국민청원, 빅데이터, 텍스트 분석.

1. 서론

청와대 국민청원 게시판은 2017년 8월 17일 현 문재인 정부 출범 100일을 맞이하여 신설되었으며, '국민이 물으면 정부가 답한다'라는 철학으로 정부와 국민들의 소통의 장의 역할을 해오고 있다. 30일 동안 20만 명 이상의 추천을 받은 청원에 대해서는 정부 및 청와대 책임자가 답변을 제공하는 것으로 되어있다. 이러한 청원데이터는 개설된 지 3년도 되지 않았는데 게시글이 60만 개에 달하고 있다. 청원데이터는 17개의 분류로 되어있으며 청원을 올리는 개개인이 가장 적합하다고 판단하는 분류영역으로 청원글을 올리고 있다. 개인의 주관적 판단으로 게시글의 분류를 선정하다보니 탐색적 분석을 해본 결과, 청원의 주요내용과 분류의 주제가 적절하게 매칭되지 않는 경우들이 많고 분류 간 구분이 모호하거나 청원내용에 적합한 분류의 범주가 빠져있는 경우들이 존재했다. 따라서 본 연구는 청원데이터의 텍스트 분석을 통해 최적의 토픽(분류, 주제)을 찾아내어 재분류를 실시하고 부정어와 긍정어 기반의 감성 분석을 함께 병행하고자 한다. 그리고 토픽과 감성에 따라서 청원글에 대한 국민참여수가 달라질 것이라는 가설하에 토픽과 감성이 청원글에 대한 국민참여수에 어떠한 영향을 주는지를 모델링하고자 한다. 이를 통해 국민청원 게시판의 청원 주제별 분류에 대한 새로운 기준을 제안하고자 하며 어떤 주제를 국민들이 개혁과 변화의 대상으로

[1]11159 경기도 포천시 호국로 1007, 대진대학교 컴퓨터공학과 조교수(교신저자). E-mail : jako403@daejin.ac.kr
[2]11159 경기도 포천시 호국로 1007, 대진대학교 컴퓨터공학과 석사과정. E-mail : AB219017@daejin.ac.kr
[접수 2020년 5월 20일; 수정 2020년 6월 12일; 게재확정 2020년 6월 13일]

생각하며 추천을 하는지에 대한 인사이트를 도출하고자 한다.

2. 이론적 배경 및 연구모형

2.1. 이론적 배경

국민다수의 관심사와 목소리가 반영된 청와대 국민청원 개설 이후, 게시된 청원데이터를 분석하는 많은 연구들이 진행되었다. Kim(2019)는 청와대 국민청원 페이지에 등록된 다양한 분야의 청원데이터를 분석하여 국민들의 현안을 파악하고자 하였다. 분류별 등록현황, 분류별 주요 이슈 등을 분석하여 국민들의 개혁 대상을 파악하고 그 의미를 도출하였다. 또한, Song, Park(2019)은 자연어처리를 통해 청원데이터에 나타난 의제와 감정을 분석하였으며, 그 결과 청원데이터에는 다양한 정책적 의제가 내포되어 있으며, 시민들의 슬픔과 분노와 같은 부정적 감정이 표출되는 것을 확인하였다. 그리고 이러한 의제 및 감정이 다른 국민들에게 얼마나 호응을 받는지를 분석하였다. 분석결과 슬픔이 분노 보다 더 일관된 효과를 보였으나 효과의 크기는 미미했다. 또한, 분노는 감정 자체의 효과 보다 다른 의제와 상호작용하는 경우 효과가 더 크게 나타난다고 하였다.

최근에 Hui, Kim(2020)은 청원의 내용과 키워드를 바탕으로 어떤 주제와 이슈들이 있었는지 계층적으로 분석하고 실제 카테고리와 비교하였다. 그 후, 딥러닝을 기반으로 청원동의 수가 20만을 넘어 답변을 받는 청원을 예측하는 모델을 개발하였다.

2.2. 토픽 모델링 및 감성 분석

토픽 모델링은 문서 집합으로부터 추상적인 토픽을 추출하는 텍스트 마이닝 기법 중 하나이다. 가장 대표적인 토픽 모델링 기법인 LDA(Latent Dirichlet Allocation)은 다수의 문서에서 잠재적으로 의미있는 토픽을 찾기 위한 확률론적 생성 모형이다(Blei et al., 2003; Blei, 2012; Kang et al., 2018). 토픽 모델링을 통해 SNS와 온라인 리뷰 등 다양한 유형의 비구조화 된 텍스트문서를 대상으로 한 연구가 다수 진행되었다. Michelson, Macskassy(2010)는 토픽 모델링을 통해 트위터 유저의 주요 관심 토픽을 분석하였으며, 이후 Hu et al.(2011)은 LDA 기법을 활용하여 시사 이슈에 관한 SNS 반응을 분석하였다. Jin et al.(2013)은 LDA 토픽 모델링을 적용하여 트위터 데이터에서 토픽의 변화 시점과 패턴을 파악하고자 하였다. 그 결과, 트위터가 언론 뉴스에 즉각적으로 반응하며 부정적 이슈를 빠르게 확산시키는 것을 확인하였다. Kang et al.(2013)은 대량의 신문 기사를 통해 주요 토픽을 추출한 후 매체별 토픽 구성의 차이를 분석하여 진보매체와 보수매체는 서로의 이데올로기에 따라 기사를 보도하는 경향성이 있음을 입증하였다. Park, Song(2013)은 43년간 발표된 국내 논문의 초록을 수집하여 문헌정보학 분야의 연구 동향 규명 및 연구 트렌드 파악을 위해 LDA 기반 토픽 모델링을 수행하였으며, Oh et al.(2016)은 구글 플레이 마켓에 업로드 되어있는 어플리케이션 설명글을 대상으로 LDA를 적용하여 50개 토픽으로 분류하고 토픽들을 3개 그룹으로 군집화하여 그 특징을 살펴보았다.

토픽 모델링에서 수행되는 절차는 3단계로 구성된다. 먼저, 토픽 모델링을 적용할 데이터를 수집한다. 다음 단계로 수집된 텍스트 데이터를 토픽 모델링에 적합한 형태로 변환하는 전처리가 수행되며, 마지막으로 비지도학습(unsupervised learning) 기반의 토픽 모델링을 통하여 토픽을 추출한다. 데이터 전처리는 (1) 토큰화(tokenization), (2) 불용어 삭제, (3) tagging, (4) lemmatization, (5)

vector- space representation으로 진행된다.

감성 분석은 소비자의 감성과 의견, 태도 등을 추출하는 텍스트 마이닝 기법 중 하나로써 오피니언 마이닝(opinion mining)이라고도 부른다. 최근 SNS와 전자 상거래, 온라인 커뮤니티 등 인터넷 매체의 발달로 인해 주관적인 요소가 포함된 텍스트 데이터가 범람하고 있으며(Han, Jin, 2014), 텍스트 데이터로부터 추출한 사용자의 감성이 기업의 마케팅에 적극 활용되면서 감성 분석의 중요성이 강조되었다(Kim, Jin, 2013; Jeong, Shon, 2018). 감성 분석은 연구자가 감성 어휘(sentiment terms)와 어휘의 긍정과 부정의 정도를 나타내는 극성(polarity)으로 구성된 감성사전을 구축하고, 이러한 감성사전을 이용하여 감성을 정량화한다. 정확한 감성 분석 결과를 도출하기 위해서는 감성사전과 같은 단어 집합이 매우 중요하다. Hu, Liu(2004)는 수년에 걸쳐 긍정 어휘 2,006개과 부정 어휘 4,783개를 추출하여 감성 분석을 수행하기 위한 영어로 구성된 Opinion Lexicon을 구축하였다. 또한 Wiebe et al.(2005)은 약 10,000개의 문장에서 나타나는 감성어휘의 사용 목적에 따라 감성과 감성 정도를 섬세하게 정의한 MPQA 감성사전을 만들었으며, Esuli, Sebastiani(2006)은 기존의 WordNet의 동의어 집합을 기반으로 준지도학습(semi-supervised)에 대한 분류 결과로 긍정, 중립, 부정 세 가지의 감성 정도를 구분하여 Senti WordNet 감성사전을 구축하였다. 한국어 감성사전에 관한 연구로는 An, Kim(2015)이 집단지성의 참여자가 단어의 감성 정도를 긍정, 부정, 중립 세 가지로 구분하여 투표하게 하고, 누적량에 따라 신뢰도가 높아지도록 하는 방법을 사용하여 오픈 한국어 감성사전을 구축하였다. 또한, Shin et al.(2016)은 주변 어휘를 계산하는 수식과 자동화된 기법으로 16,362개 어휘에 대하여 긍정, 부정, 중립 각각의 점수를 부여하여 구축한 KOSAC 감성사전이 있다. Park et al.(2018)은 국립국어원 표준국어대사전의 뜻풀이 분석을 통해 긍부정 어휘를 추출하여 총 14,843개의 1-gram, 2-gram, 관용구, 문형, 축약어, 이모티콘 등 어휘에 대한 긍정, 중립, 부정을 계산하여 KNU 감성사전을 구축하였다.

2.3. 연구모형

많은 기업들은 온라인 또는 소셜 미디어 참여를 통해 팬 페이지 및 공감 감정을 게시하여 소비자의 행동(좋아요, 댓글 및 공유)을 유인하고 있으며, 팬이 광고를 게시하는데 얼마나 효과적으로 다수 잠재 고객을 참여시키는가가 소셜 미디어 마케팅에서 승리하는 영향을 주는 가장 큰 요인의 하나가 되었다. 오늘날 소셜 미디어의 유용성은 소비자와 정보를 공유되는 방식에 달려 있다. 즉 게시물이 텍스트, 사진 및 비디오와 같이 동적 일수록 선호, 댓글, 답글 등의 공감이 많고 소셜 미디어 참여에 영향을 미친다고 하였으며(Luarn et al., 2015), 정보 및 social value와 같은 콘텐츠 유형도 소셜 미디어 참여에 영향을 준다고 하였다(Le, 2018; Luarn et al., 2015). 그러나 콘텐츠 유형은 게시물의 주제별로도 소셜 미디어 참여에 영향을 준다고 하였다(Di Gangi, Wasko, 2016). 이러한 맥락에서 본 연구에서는 국민청원 게시글의 토픽(주제, 분류)과 감성(긍정어 vs 부정어)과 국민참여수의 관계를 분석하고자 한다. 즉, 국민청원 게시물을 대상으로 GLM(Generalized Linear Model) 모델을 적용하여 토픽 모델링 LDA 기법으로 추출한 토픽과 감성 분석을 통해 계산된 감성점수가 국민청원 참여에 어떠한 영향을 미치는지를 다음과 같이 검증하고자 한다.

Hypothesis 1 : 국민청원 게시물의 토픽은 청원에 대한 국민참여수에 영향을 미친다.
Hypothesis 2 : 국민청원 게시물의 감성은 청원에 대한 국민참여수에 영향을 미친다.

3. 실증 분석

3.1. 토픽 모델링

국민청원 데이터를 수집하기 위해 Python 기반의 Beautifulsoup과 Selenium 패키지를 사용하여 국민청원 게시판 전용 웹 크롤러(web crawler)를 구축하였다. 2017년 8월 19일부터 2019년 6월 11일까지 약 22개월간 청와대 국민청원 게시판에 등록된 580,782개 청원데이터 중 삭제된 145,227개 청원을 제외한 435,555개의 청원데이터를 수집하였다. 또한, 국민청원 게시판에서 분류하고 있는 17개 분야 중 '기타' 분야를 제외한 16개 분야만 분석에 사용하였으며, 중복된 청원은 분석 대상에서 제외하였다. 최종적으로 분석에 사용한 청원데이터는 365,346개이며, 청원데이터는 청원 제목과 청원 내용, 청원 분야, 참여인원, 청원 등록일로 구성되어 있다.

최종 분석에 사용된 분야별 청원데이터는 Table 1과 같다. '기타' 분야를 제외한 16개 분야 중 정치개혁의 문서 수가 65,341개(17.88%)로 가장 많으며, 그 다음 인권/성평등 37,037개(10.14%), 안전/환경 33,830개(9.26%), 교통/건축/국토 29,571개(8.09%) 순으로 높았다. 반면, 농·산·어촌 분야의 문서 수는 2,073개(0.57%)로 매우 적었으며, 또한 저출산/고령화대책 3.771개(1.03%)와 반려동물 4,365개(1.19%)도 전체 데이터의 약 1%를 차지하는 수준으로 문서 수가 적었다.

Table 1. Number of documents in 16 categories

Variables		n	%
category	Economic democratization	16,968	4.64
	Transportation / Construction / Territory	29,571	8.09
	Rural area	2,073	0.57
	Culture / Art / Athletic / Press	19,334	5.29
	Future	19,646	5.38
	Pet	4,365	1.19
	Health / Welfare	26,152	7.16
	Growth engine	7,410	2.03
	Safety / Environment	33,830	9.26
	Diplomacy / Reunification / National defense	28,098	7.69
	Child / Education	26,485	7.25
	Human rights / Gender equality	37,037	10.14
	Job	24,077	6.59
	Aging society / Population	3,771	1.03
	Political reform	65,341	17.88
	Administration	21,188	5.80
total		365,346	100.00

토픽 모델링을 수행하기 위해서는 먼저 텍스트를 토픽 모델링에 적합한 형태로 전처리하는 단계가 선행되어야 한다. 그 첫 번째 단계는 토큰화 단계이다. Python의 Konlpy 라이브러리의 경우 5가지의 한국어 형태소 분석기(Kkma, Hannanum, Mecab, Komoran, OKT)를 제공하고 있으며, 그 중 본 논문에서는 Mecab 형태소 분석기를 사용하여 토큰화를 수행하였다. 두 번째 단계는 품사의 추출 단계이다. 일반적으로 핵심 키워드를 추출해내는 연구에서는 명사만을 추출하여 사용하며, 본 논문에서도 토픽 모델링을 적용하기 위해 청원 내용으로부터 명사만을 추출하여 사용하였다. 마지

막으로 한자, 숫자, 기호 등 의미가 없는 단어를 불필요한 단어로 간주하여 불용어 사전에 정의하고 삭제하였다.

Figure 1. Wordcloud of national petition page postings

Figure 1은 전체 청원 내용에서 추출한 명사의 출현 빈도를 기반으로 한 Wordcloud를 보여준다. 국민, 사람, 생각, 나라, 우리, 정부와 같은 단어가 많이 출현하였으며, 그 외에도 교육, 경제, 범죄, 연금, 일자리, 인권 등 다양한 주제에 대한 단어들이 혼재되어 나타나는 것을 확인할 수 있다. 이렇게 정제된 청원데이터로부터 다양한 토픽(주제)을 도출하기 위해 토픽 모델링을 수행하고자 한다.

Maier et al.(2018)는 토픽 모델링의 결과는 토픽 수에 의존하며 전체 문서 집합에서 적절한 토픽 수를 결정해야 한다고 하였다. Westerlund et al.(2018)에 따르면 최적의 토픽 수는 시행착오를 기반으로 결정하며, 주제 간의 중복을 피하고 주제의 해석성을 보장해야 한다. 또한, Calheiros et al.(2017)에 따르면. 토픽의 수가 적을수록 주제 간 중복을 피하고 구조적 유효성을 보장하는데 도움이 된다. 본 연구에서는 최적의 토픽 수를 선정하기 위해 단어별로 문서 간 거리를 k-means clustering과 Elbow 기법을 사용하여 측정하였다. Figure 2를 통해 최적의 토픽 수를 10개로 판단하였고, Python의 gensim 패키지에서 제공하는 LDA를 사용하여 분석을 진행하였다.

문서는 비구조화 형태인 텍스트 자료이기 때문에 분석을 수행하기 적합하도록 구조화된 자료로 변환되어야 한다(Oh, Jin, 2012). 문서를 구조화된 자료로 표현하는 방법에는 Harris(1954)가 제안한 BOW(Bag of Word)가 있다. BOW는 문서에 출현한 단어의 빈도수를 기반으로 측정되는 방법이다. 이 방법은 문서의 길이가 단어 빈도에 영향을 미친다는 문제가 있다. 대체로 문서의 길이가 길수록 단어의 빈도는 증가하고 문서의 길이가 짧을수록 단어의 빈도는 줄어들 수 있다. 이러한 문제점을 해결하기 위한 대안으로 TF-IDF(Term Frequency-Inverse Document Frequency)가 제안되었다. TF-IDF는 핵심 키워드를 추출하기 위한 척도로 활용이 가능하다. TF(Term Frequency)는 한 문서 내에서 특정 단어가 출현한 빈도수를 나타내며, 주어진 단어의 문서 내 출현 빈도가 높을수록 상대적으로 더 중요하다는 의미이다. IDF(Inverse Document Frequency)는 문서 집합의 전체 문서 수를 특정 단어가 나타는 문서의 수로 나눈 값이다. 따라서 높은 IDF 값을 가지는 단어는 문서 내에서

중요한 의미를 가지는 단어이다. TF-IDF는 두 값을 곱해서 사용하므로 어떤 단어가 해당 문서에 자주 등장하지만, 다른 문서에는 출현빈도가 낮을수록 높은 값을 가지게 된다. Table 2는 각 토픽에서 TF-IDF 값이 큰 주요 단어 15개를 정리하였다. 토픽 모델링 결과, 토픽1은 국민, 공무원, 사건, 경찰, 수사, 조사, 행정, 민원 등의 단어가 상위 단어로 출현하여 주제를 법률/수사/민원으로 정의될 수 있다. 마찬가지로 토픽2는 병원, 보험, 환자, 치료, 장애, 의료 등의 단어를 통해 주제를 건강/의료로 정의하였다. 토픽3의 선수, 올림픽, 축구, 국가, 대표, 감독 등의 단어를 통해 주제를 스포츠로 정의하였으며 토픽4는 대통령, 북한, 정부, 대한민국, 국회의원, 한국 등의 단어를 통해 정치/대북으로 정의하였다. 토픽5는 도로/교통과 연관이 있으며 토픽6은 아동/교육과 관련되어 있고, 토픽7은 시간, 사람, 기업, 임금, 회사. 일자리, 근로자 등 근로/일자리와 관련된 단어들이 주로 나타났다. 토픽8은 주택, 부동산, 아파트, 대출, 세금, 화폐, 시장 등 부동산/경제과 관련되어 있고 토픽9는 미세먼지, 중국, 문제, 사용, 환경 등 환경과 관련된 단어들이 주로 등장한다. 토픽10은 여성, 처벌, 피해자, 사건, 범죄, 청소년, 남성, 인권 등 인권/범죄와 관련되어 있어 총 10가지 토픽에 대한 주제를 도출 하였다. 이러한 10개의 토픽 중 Figure 3은 대표적인 6가지에 대한 wordcloud를 TF-IDF 값을 기반으로 표현하였다. 출현빈도를 기반으로 한 결과는 TF-IDF를 적용한 결과와 동일하였으며, 따라서 본 논문에서는 TF-IDF를 적용한 결과만 제시한다.

Table 3은 각 토픽별 청원데이터의 빈도 및 비율을 나타낸다. 정치/대북에 관한 문서 수가 84,951개(23.25%)로 가장 많으며, 인권/범죄가 55,691개(15.24%), 부동산/경제가 39,086개(10.70%)등의 순으로 많게 나타났다. 반면, 스포츠 13,335개(3.65%)와 건강/의료 16,386개(4.49%)는 해당 토픽에 포함된 청원데이터 수가 상대적으로 적다는 것을 알 수 있다.

Table 2. Importance of words in 10 topics by TF-IDF

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
words	TF-IDF	words	TF-IDF	words	TF-IDF	words	TF-IDF	words	TF-IDF
nation	0.323	hospital	0.427	athlete	0.561	nation	0.533	person	0.278
public officer	0.230	insurance	0.242	idea	0.211	country	0.277	idea	0.226
incident	0.211	patient	0.232	person	0.194	president	0.268	vehicle	0.219
police	0.182	person	0.216	Olympic	0.192	our	0.254	accident	0.177
investigation	0.181	treatment	0.205	soccer	0.173	idea	0.200	nation	0.155
research	0.176	disorder	0.194	country	0.171	person	0.191	time	0.151
administration	0.174	medical care	0.182	nation	0.171	North Korea	0.189	problem	0.148
person	0.149	idea	0.181	our	0.170	government	0.186	road	0.148
government	0.127	doctor	0.180	representative	0.167	South Korea	0.159	resident	0.123
civil complaint	0.126	nation	0.165	country	0.157	country	0.149	area	0.123
content	0.125	health	0.140	Korean	0.117	problem	0.107	taxi	0.119
president	0.114	surgery	0.122	coach	0.117	congressman	0.104	construction	0.115
judge	0.114	country	0.112	South Korea	0.116	Korea	0.096	safety	0.114
idea	0.113	support	0.109	match	0.113	politics	0.095	because	0.111
petition	0.109	our	0.092	association	0.096	economy	0.088	case	0.106

Table 2. Importance of words in 10 topics by TF-IDF (Continued)

Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
words	TF-IDF	words	TF-IDF	words	TF-IDF	words	TF-IDF	words	TF-IDF
child	0.443	time	0.287	nation	0.359	nation	0.270	female	0.355
school	0.324	person	0.226	housing	0.250	fine dust	0.260	person	0.334
student	0.295	idea	0.212	government	0.230	our	0.246	idea	0.274
idea	0.241	enterprise	0.189	real estate	0.225	idea	0.220	punishment	0.218
education	0.241	nation	0.188	pension	0.217	country	0.213	nation	0.188
teacher	0.202	wage	0.184	person	0.198	government	0.210	victim	0.179
time	0.166	work	0.174	policy	0.168	China	0.199	incident	0.172
person	0.148	company	0.167	apartment	0.161	problem	0.191	country	0.160
parents	0.136	public officer	0.152	ordinary person	0.145	use	0.149	crime	0.151
our	0.126	job	0.148	idea	0.145	Korean	0.145	petition	0.146
problem	0.121	government	0.147	loan	0.137	person	0.142	youth	0.145
day-care center	0.106	worker	0.144	tax	0.134	country	0.131	male	0.143
teacher	0.103	minimum	0.134	money	0.131	enterprise	0.128	society	0.140
our	0.103	employee	0.122	market	0.126	policy	0.111	our	0.123
country	0.102	policy	0.114	country	0.112	environment	0.108	human rights	0.110

Table 3. Number of documents in 10 topics

			n	%
Topic	1	Law / Investigation / Civil complaint	33,523	9.18
	2	Health / Medical care	16,386	4.49
	3	Sports	13,335	3.65
	4	Politics / North Korea	84,951	23.25
	5	Road / Transportation	29,554	8.09
	6	Child / Education	32,145	8.80
	7	Work / Job	35,460	9.71
	8	Real-estate / Economy	39,086	10.70
	9	Environment	25,215	6.90
	10	Human rights / Crime	55,691	15.24
		Total	365,346	100.00

국민청원 게시판에서 자체적으로 분류하는 '기타'를 제외한 16개 카테고리와 새롭게 도출한 10개 토픽 간의 교차분석을 실시한 결과는 아래의 Table 4와 같다. 토픽1(법률/수사/민원)으로 군집된 카테고리 중 정치개혁(47.5%)이 가장 많이 비율이 높았다. 토픽2(건강/의료)는 보건복지(59.2%)의 비율이 가장 높았으며, 토픽3(스포츠)은 문화/예술/체육/언론(56.7%)의 비율이 가장 높았다. 나머지 토픽에 대해서도 가장 비율이 높은 카테고리를 살펴보면, 토픽4(정치/대북)은 정치개혁(37.9%), 토픽5(도로/교통)는 교통/건축/국토(37.1%), 토픽6(아동/교육)은 육아/교육(53.6%), 토픽7(근로/일자리)은 일자리(42.6%), 토픽8(부동산/경제)은 교통/건축/국토(31.5%), 토픽9(환경)는 안전/환경(34.9%), 토픽10(인권/범죄)은 인권/성평등(37.3%)이 가장 높게 나타났다. 본 논문에서 LDA를 적용하여 기존의 16개 카테고리보다 적은 수의 토픽 10개를 새롭게 도출하였으며, 교차분석 결과 제안 토픽은 적은 수로도 기존의 16개 카테고리와 유사하며, 청원글의 주제를 잘 반영하고 있다고 판단된다 .

Table 4. Cross-tabulation between 16 categories and 10 topics

		Topic										
		1	2	3	4	5	6	7	8	9	10	Total
Category	Economic democratization	1,222 (3.6)	189 (1.2)	66 (0.5)	2,503 (2.9)	275 (0.9)	236 (0.7)	2,527 (7.1)	8,335 (21.3)	997 (4.0)	618 (1.1)	16,968 (4.6)
	Transportation / Construction / Territory	781 (2.3)	280 (1.7)	66 (0.5)	1,798 (2.1)	10,961 (37.1)	406 (1.3)	1,087 (3.1)	12,318 (31.5)	1,157 (4.6)	717 (1.3)	29,571 (8.1)
	Rural area	122 (0.4)	36 (0.2)	54 (0.4)	148 (0.2)	481 (1.6)	60 (0.2)	256 (0.7)	182 (0.5)	672 (2.7)	62 (0.1)	2,073 (0.6)
	Culture / Art / Athletic / Press	1,204 (3.6)	108 (0.7)	7,555 (56.7)	3,409 (4.0)	555 (1.9)	472 (1.5)	431 (1.2)	266 (0.7)	907 (3.6)	4,427 (7.9)	19,334 (5.3)
	Future	1,271 (3.8)	295 (1.8)	511 (3.8)	6,071 (7.1)	523 (1.8)	1,251 (3.9)	1,279 (3.6)	3,248 (8.3)	1,905 (7.6)	3,292 (5.9)	19,646 (5.4)
	Pet	141 (0.4)	260 (1.6)	1,901 (14.3)	262 (0.3)	411 (1.4)	111 (0.3)	37 (0.1)	38 (0.1)	114 (0.5)	1,090 (2.0)	4,365 (1.2)
	Health / Welfare	661 (2.0)	9,707 (59.2)	139 (1.0)	1,917 (2.3)	1,739 (5.9)	2,195 (6.8)	3,113 (8.8)	3,651 (9.3)	1,903 (7.5)	1,127 (2.0)	26,152 (7.2)
	Growth engine	295 (0.9)	80 (0.5)	117 (0.9)	1,506 (1.8)	236 (0.8)	232 (0.7)	984 (2.8)	1,978 (5.1)	1,671 (6.6)	311 (0.6)	7,410 (2.0)
	Safety / Environment	1,106 (3.3)	897 (5.5)	310 (2.3)	3,332 (3.9)	7,896 (26.7)	1,570 (4.9)	881 (2.5)	361 (0.9)	8,797 (34.9)	8,680 (15.6)	33,830 (9.3)
	Diplomacy / Reunification / National defense	1,152 (3.4)	782 (4.8)	1,068 (8.0)	20,347 (24.0)	680 (2.3)	604 (1.9)	840 (2.4)	177 (0.5)	1,428 (5.7)	1,020 (1.8)	28,098 (7.7)
	Child / Education	778 (2.3)	488 (3.0)	98 (0.7)	1,173 (1.4)	627 (2.1)	17,220 (53.6)	930 (2.6)	239 (0.6)	1,482 (5.9)	3,450 (6.2)	26,485 (7.2)
	Human rights / Gender equality	3,808 (11.4)	1,171 (7.1)	453 (3.4)	5,001 (5.9)	866 (2.9)	2,345 (7.3)	1,791 (5.1)	463 (1.2)	356 (1.4)	20,783 (37.3)	37,037 (10.1)
	Job	776 (2.3)	485 (3.0)	129 (1.0)	2,313 (2.7)	821 (2.8)	1,854 (5.8)	15,115 (42.6)	739 (1.9)	1,235 (4.9)	610 (1.1)	24,077 (6.6)
	Aging society / Population	61 (0.2)	438 (2.7)	19 (0.1)	242 (0.3)	148 (0.5)	1,573 (4.9)	352 (1.0)	624 (1.6)	120 (0.5)	194 (0.3)	3,771 (1.0)
	Political reform	15,909 (47.5)	508 (3.1)	676 (5.1)	32,179 (37.9)	1,033 (3.5)	1,054 (3.3)	2,702 (7.6)	3,508 (9.0)	1,156 (4.6)	6,616 (11.9)	65,341 (17.9)
	Administration	4,236 (12.6)	662 (4.0)	173 (1.3)	2,750 (3.2)	2,302 (7.8)	962 (3.0)	3,135 (8.8)	2,959 (7.6)	1,315 (5.2)	2,694 (4.8)	21,188 (5.8)
	Total	33,523 (100.0)	16,386 (100.0)	13,335 (100.0)	84,951 (100.0)	29,554 (100.0)	32,145 (100.0)	35,460 (100.0)	39,086 (100.0)	25,215 (100.0)	55,691 (100.0)	365,346 (100.0)

3.2. 감성 분석

본 연구에서는 KNU 한국어 감성사전을 사용하여 감성 분석을 수행하였다. KNU 감성사전은 다양한 분야에서 사용될 수 있는 범용 감성사전으로 국립국어원 표준국어대사전의 뜻풀이 분석을 통한 긍부정 어휘를 추출하여 총 14,843개의 1-gram, 2-gram, 관용구, 문형, 축약어, 이모티콘 등 어휘에 대한 극성 값이 계산되어 있다. 극성은 '매우 부정(-2)', '부정(-1)', '중립(0)', '긍정(1)', '매우 긍정(2)' 5가지로 구분되어 있다. 본 연구에서는 감성 분석을 통해 청원글의 극성을 positive(긍정, 매우긍정), negative(부정, 매우부정), neutral(중립) 3개로 분류하였다.

Table 5는 감성 분석을 통해 추출된 토픽별 감성어휘의 출현 빈도 및 비율을 보여준다. 청원데

이터의 특성상 비판 및 불평의 내용이 많아 모든 토픽에서 부정 어휘의 출현 비율이 긍정 어휘보다 높게 나타났다. 특히 인권/범죄의 부정 어휘가 308,189개(69.2%) > 건강/의료 116,183개(66.8%) > 법률/민원 153,714개(64.4%) > 도로/교통 139,936개(62.4%)의 순으로 평균 60.2% 보다 부정 어휘 비율이 높게 나타났다. 반면 스포츠는 39.947개(52.5%), 환경은 105,003개(54.2%), 근로/일자리는 155,310개(56.1%), 아동/교육은 169,801(56.2%), 부동산/경제는 153,974개(56.7%)로 부정 어휘가 평균보다 적게 나타났으며, 긍정 어휘의 출현이 상대적으로 많았다. 긍·부정의 분포가 대비되는 대표적인 토픽 4개를 Figure 4에 제시하였다. 토픽2 건강/의료와 토픽10 인권/범죄는 부정 어휘의 분포가 많은 것을 확인할 수 있으며, 토픽3 스포츠와 토픽9 환경의 경우 긍정과 부정 어휘가 상대적으로 고르게 분포되어있음을 알 수 있다.

Table 5. Frequency of positive and negative documents in each topic

		Sentiment			
		Negative	Neutral	Positive	Total
Topic	1	153,714(64.4)	7,014(2.9)	78,123(32.7)	238,851(100.0)
	2	116,183(66.8)	3,557(2.0)	54,288(31.2)	174,028(100.0)
	3	39,947(52.5)	2,027(2.7)	34,181(44.9)	76,155(100.0)
	4	327,839(57.6)	15,275(2.7)	226,405(39.8)	569,519(100.0)
	5	139,936(62.4)	7,141(3.2)	77,270(34.4)	224,347(100.0)
	6	169,801(56.2)	7,807(2.6)	124,287(41.2)	301,895(100.0)
	7	155,310(56.1)	6,042(2.2)	115,487(41.7)	276,839(100.0)
	8	153,974(56.7)	5,670(2.1)	111,800(41.2)	271,444(100.0)
	9	105,003(54.2)	5,394(2.8)	83,333(43.0)	193,730(100.0)
	10	308,189(69.2)	18,968(4.3)	118,449(26.6)	445,606(100.0)
	Total	1,669,896(60.2)	78,895(2.8)	1,023,623(36.9)	2,772,414(100.0)

Chi-square = 42,439, p-value < 0.001

3.3. 통계 모형

통계 모형의 종속 변수 '국민청원 게시물 참여 수'는 평균이 188.16이며, 분산이 36627423.29인 양의 정수를 가지는 이산형 변수이다. 따라서 일반화선형모형(GLM) 중 하나인 negative binomial regression 모델을 통해 과대분산 문제를 해결하고자 한다. 모델에서 연결 함수는 $g(u)=\log(u)$, 여기서 $\mu=E(y)$를 사용하며, 모수는 최대우도추정법으로 추정하였다. '토픽'과 '감성'이 국민청원 게시물 청원 참여수에 영향을 미칠 것이라는 가설하에 최적의 모델을 선택하기 위하여, AIC와 BIC 지표를 기반으로 한 변수소거법(backward elimination)을 사용하였다.

Table 6. Goodness of fit and model evaluation

Statistics	DF	Value	Value/DF
Deviance	365,345	481397.1030	1.3177
Pearson Chi-Square	365,345	67424033.541	184.5540
Log Likelihood		647175198.74	
AIC		3224918.2827	
BIC		3225047.9859	

Table 6는 최종 선택된 모델 및 모델 평가 기준에 대한 적합도를 보여준다. value/df는 1.3에 가

깝고 AIC는 3224918.28로 고려한 여러 모형들 중 AIC 값이 가장 작아 최적모형으로 선택하였다.

Table 7은 최적의 일반화선형모형에서 선택된 국민청원 게시물의 국민참여수에 영향을 주는 요인변수들이다. topic과 sentiment 모두 국민참여수에 영향을 주는 것으로 나타났다. 따라서 연구가설 H1과 H2는 통계적으로 유의한 가설임을 확인 할 수 있다.

Table 7. Significant factors to impact on the engagement

LR Statistics For Type 3 Analysis			
Source	DF	χ^2	Pr>ChiSq
Topic	9	8500.65	<.0001
Sentiment	1	14962.0	<.0001

Table 8의 토픽의 경우 토픽 2(건강/의료)와 토픽 10(인권/범죄)에서 국민참여수에 가장 영향을 주는 것으로 나타났으며 토픽 8(동산/경제)과 토픽 7(근로/일자리)이 가장 효과가 낮은 것으로 나타났다. 따라서 토픽의 유형에 따라 청원참여에 차이가 발생함을 확인할 수 있다. 감성(긍정어수-부정어수)의 경우 sentiment가 1 증가할 때, 국민청원 게시물 참여 수에 exp(-0.0818)=0.921만큼 영향을 미친다는 것을 알 수 있다. 따라서 가설 H2 또한 통계적으로 유의한 가설이었음을 다시 한 번 알 수 있다.

Table 8. The parameter estimates of the negative binomial regression model

Parameter		DF	Estimate	S.E	Wald χ^2	Pr>ChiSq
Intercept		1	5.2788	0.0104	256318	<.0001
Topic	1	1	-0.2554	0.0159	258.83	<.0001
	2	1	0.0593	0.0202	8.60	0.0034
	3	1	-0.0781	0.0221	12.47	0.0004
	4	1	-0.4163	0.0126	1084.20	<.0001
	5	1	-0.6088	0.0165	1362.28	<.0001
	6	1	-0.2453	0.0163	225.93	<.0001
	7	1	-0.8148	0.0161	2568.86	<.0001
	8	1	-1.2656	0.0153	6830.41	<.0001
	9	1	-0.4222	0.0177	566.86	<.0001
	10	0	0.0000	0.0000	.	.
Sentiment		1	-0.0818	0.0007	12968.0	<.0001

4. 결론

본 연구는 청와대 청원데이터의 '기타' 분류를 제외한 16개의 분류를 k-means clustering을 통해 차원이 축약된 최적의 토픽 수를 구하고, 토픽과 감성이 국민청원 게시글의 국민참여수에 영향을 미치는가를 분석하는 것이었다. 결론적으로 토픽이 어떤 내용인가에 따라 국민참여수에 차이가 발생함을 확인할 수 있었다. 건강/의료와 인권/범죄 토픽 등에서 상대적으로 국민참여가 높다는 것을 알 수 있었으며 감성의 경우 게시글 문장 내에 긍정어수-부정어의 수가 1 증가할수록 청원글에 대한 국민참여수가 증가한다는 사실을 확인 할 수 있었다. H2 연구가설을 정의할 때는 어휘의 긍정어보다 부정어가 증가할수록 국민참여수에 영향을 줄 것으로 가정했었으나, 연구결과 국민청원에

서 게시글 내의 긍정어휘의 중요도를 확인하게 되었다는 점이다. 따라서 동일한 주제라 하더라도 부정적인 어휘의 사용보다는 긍정 어휘를 사용하는 것이 보다 국민들의 공감대를 이끌어 내는데 역할을 하고 있다고 합리적인 추론을 해본다.

본 논문은 청와대의 국민청원에 대한 17개 대분류 기준의 새로운 대안을 제시해 보고자 하는 것에 의미를 부여할 수 있다 하겠다. 이러한 대안적 대분류 기준에 따른 중분류, 소분류 등에 대한 연구를 계속 진행해 보고자 한다. 또한, 감성 분석에서 긍정어간의 연관관계, 어떤 어휘들의 결합이 국민참여도를 끌어올리는지에 대한 연구를 하지 못한 것은 아쉬운 점으로 남는다. 따라서 어휘들 간의 연관관계(긍정, 부정, 감정 긍정어, 감정 부정어, 자극어, 반감어 등)를 재정의하고, 세분화하여 추가로 연구해볼 필요가 있다하겠다.

Reference

An, J. K., Kim, H. W. (2015). Building a Korean sentiment lexicon using collective intelligence, *Journal of Intelligence and Information System*, 21(2), 49-67. (in Korean).

Blei, D. M. (2012), Probabilistic topic models, *Communications of the ACM*, 55(4), 77–84.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research*, 3(4-5), 993-1022.

Calheiros, A. C., Moro, S., Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling, *Journal of Hospitality Marketing & Management*, 26(13), 675-693.

Di Gangi, P. M., Wasko, M. M. (2016). Social media engagement theory: Exploring the influence of user engagement on social media usage, *Journal of Organizational and End User Computing*, 28(2), 53-73.

Esuli, A., Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, 417-422.

Han, G. H., Jin, S. H. (2014). Introduction to big data and the case study of its applications, *Journal of The Korean Data Analysis Society*, 16(3), 1337-1351. (in Korean).

Harris, Z. S. (1954). Distributional structure, *Word*, 10(2-3), 146-162.

Hu, M., Liu, B. (2004). Mining and summarizing customer reviews, *International Conference on Knowledge Discovery and Data Mining*, 168-177.

Hu, Y., John, A., Seligmann, D. D. (2011). Event analytics via social media, *Proceedings of the 2011 ACM Workshop on Social and Behavioural Networked Media Access*, 39-44.

Hui, W. Y., Kim, H. H. (2020). Topic analysis of the national petition site and prediction of answerable petitions based on deep learning, *KIPS Transactions on Software and Data Engineering*, 9(2), 45-52. (in Korean).

Jeong, M. S., Shon, B. Y. (2018). Design and analysis of sentiment classification model for Korean music reviews based on convolutional neural networks, *Journal of The Korean Data Analysis Society*, 20(4), 1863-1871. (in Korean).

Jin, S. A., Heo, G. E., Jeon, Y. K., Song, M. (2013). Topic-network based topic shift detection on Twitter, *Journal of the Korean Society for Information Management*, 30(1), 285-302. (in Korean).

Kang, B. I., Song, M., H., Jho, W. S. (2013). A Study on opinion mining of newspaper texts based on topic modeling, *Journal of the Korean Library and Information Science Society*, 47(4), 315-334. (in Korean).

Kang, C. W., Kim, K. K., Choi, S. B. (2018). A topic analysis of abstracts in journal of Korean data analysis society, *Journal of The Korean Data Analysis Society*, 20(6), 2907-2915. (in Korean).

Kim, C. W. (2019). Major reform issues through the national petition data, *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, 9(2), 823-832. (in Korean).

Kim, J. S., Jin, S. H. (2013). A study on the application of opinion mining based on big data, *Journal of The*

Korean Data Analysis Society, 15(1), 101-113. (in Korean).

Le, T. (2018). Influence of WOM and content type on online engagement in consumption communities, *Online Information Review*, 42(2), 161-175.

Luarn, P., Lin, Y., Chiu, Y. (2015). Influence of Facebook brand-page posts on online engagement, *Online Information Review*, 39(4), 505-519.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology, *Communication Methods and Measures*, 12(2-3), 93-118.

Michelson, M., Macskassy, S. A. (2010). Discovering users' topics of interest on Twitter: a first look, *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, 73-80.

Oh, M. S., Kim, S. K., Kang, C. W., Kim, K. K., Choi, S. B., Jeon. Y. J. (2016). Topics classification of applications using the latent dirichlet allocation model, *Journal of The Korean Data Analysis Society*, 18(4), 1895-1903. (in Korean).

Oh, S. W., Jin, S. H. (2012). A Study on analysis of internet shopping mall customers' reviews by text mining, *Journal of The Korean Data Analysis Society*, 14(1), 125-137. (in Korean).

Park, J. H., Song, M. (2013). A study on the research trends in library & lnformation science in Korea using topic modeling, *Journal of the Korean Society for Information Management*, 30(1), 7-32. (in Korean).

Park, S. M., Na, C. W., Choi, M. S., Lee, D. H., On, B. W. (2018). KNU Korean sentiment lexicon - Bi-LSTM-based method for building a Korean sentiment lexicon, *Journal of Intelligence and Information Systems*, 24(4), 219-240. (in Korean).

Shin, H. P., Kim, M. H., Park, S. Z. (2016). Modality-based sentiment analysis through the utilization of the Korean sentiment analysis corpus, *EONEOHAG : Journal of the Linguistic Society of Korea*, 0(74), 93-114. (in Korean).

Song, J. M., Park, Y. D. (2019). What happens in the blue house online petition? : An analysis of the blue house online petition based on natural, *Korean Political Science Review*, 53(5), 53-78. (in Korean).

Westerlund, M., Leminen, S., Rajahonka, M. (2018). A topic modelling analysis of living labs research, *Technology Innovation Management Review*, 8(7), 40-51.

Wiebe, J., Wilson, T., Cardie. C. (2005). Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation*, 39(2), 164-210.

A Study on the Topic and Sentiment of National Petition Data Using Text Analysis

Hyesun Suh[1], *Jinsoo So*[2]

Abstract

The purpose of this study is to conduct text analysis and topic modeling on the posts of the National Petition of the Blue House, where there are many interests and participation of the people. The national petition data, which has been classified into 17, has accumulated over 600,000 posts since it was opened and accumulated at the level of big data. We would like to reclassify whether the first 17 classifications of these petitions are appropriate based on text analysis. Through this, we propose new topics (classifications) for the postings of the National Petition, and suggest sentiment levels for each topic through sentiment analysis considering the positive and negative words. In addition, analysis using generalized linear model (GLM) is conducted to study whether there is a difference in the participation number of the national participation posts according to the topics and the sentiments of words. As a result of the analysis, it was confirmed that the participation rate of citizens in the petition posts was different under statistical significance according to the topics and sentiments. Also, it was confirmed that the use of positive words was more effective than the use of negative words to increase participation in the petition post.

Keywords : topic modeling, sentiment analysis, national petition, big data, text analysis.

[1](Corresponding Author)Department of Computer Science Engineering Professor, Daejin University, 1007 Hoguk-ro, Pocheon-si, Gyeonggi-do 11159, Republic of Korea. E-mail : jako403@daejin.ac.kr

[2]Master's Course in Computer Science Engineering, Daejin University, 1007 Hoguk-ro, Pocheon-si, Gyeonggi-do 11159, Republic of Korea. E-mail : AB219017@daejin.ac.kr

[Received 21 May 2020; Revised 10 June 2020; Accepted 12 June 2020]