Journal of the Korea Convergence Society
Vol. 12. No. 4, pp. 31-42, 2021

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2021.12.4.031

텍스트 분석을 이용한 코로나19 관련 국내 논문의 주제 및 감성에 관한 융합 연구

허성민[1], 양지연[2*]
[1]금오공과대학교 응용수학과 학생, [2]금오공과대학교 응용수학과 부교수

A Convergence Study on the Topic and Sentiment of COVID19 Research in Korea Using Text Analysis

Seong-Min Heo[1], Ji-Yeon Yang[2*]
[1]Undergraduate Student, Dept. of Applied Mathematics, Kumoh National Institute of Technology
[2]Associate Professor, Dept. of Applied Mathematics, Kumoh National Institute of Technology

요 약 본 연구에서는 코로나19 관련 연구논문의 연구주제를 탐색하고 동향을 검토하고 있다. 또한 감성분석을 통해 부정적인 어조가 강한 경고가 되는 주제들을 알아본다. 잠재 디리슐레 할당(LDA)를 이용하여 총 8개의 토픽을 발견하였고, 이를 구조적 토픽 모델링(STM)과 비교하여 비교적 안정적인 결과임을 확인하였다. 또한 k-means 군집 알고리즘을 통해 각 토픽별로 세부 연구주제를 발견하였고 주성분 분석을 이용하여 이를 시각적으로 표현하였다. 감성분석을 통해 각 토픽별 긍정적, 부정적인 단어들을 살펴보고 감성점수를 계산하여 연구논문의 주된 어조를 파악하였는데, 특히 생물 의학 관련, 국제적 역학관계, 심리적 영향과 관련된 연구에서 부정적인 어조가 강한 것으로 나타나 해당 부문에 대해서 주의와 관심이 요구된다. 향후 연구자들이 연구의 방향성을 탐색하고 정책결정자들이 연구지원 사업을 결정하는데 기초자료로 활용될 수 있을 것이다.

주제어 : 코로나-19, 융합 연구, 텍스트 마이닝, 토픽 모델링, k-means 군집 알고리즘, 감성분석

Abstract The purpose of this study was to explore research topics and examine the trend in COVID19 related research papers. We identified eight topics using latent Dirichlet allocation and found acceptable validity in comparison with the structural topic model. The subtopics have been extracted using k-means clustering and plotted in PCA space. Additionally, we discovered the topics bearing negative tones and warning signs by sentiment analysis. The results flagged up the issues of the topics, Biomedical Related, International Dynamics and Psychological Impact. The findings could serve as a guideline for researchers who explore new research directions and policymakers who need to make decisions about which research projects to support.

Key Words : COVID-19, Convergence study, Text mining, Topic modeling, K-means clustering algorithm, Sentiment analysis

*This research was supported by Kumoh National Institute of Technology (202001950001).
*Corresponding Author : Ji-Yeon Yang(jyang@kumoh.ac.kr)
Received February 24, 2021 Revised March 22, 2021
Accepted April 20, 2021 Published April 28, 2021

1. 서론

2019년 12월 중국에서 발생한 코로나 바이러스는 현재까지 걷잡을 수 없이 퍼지고 있으며 전 세계는 유래없는 팬데믹 위기를 겪고 있다. 국내에서도 2021년 2월 현재, 84,000명 이상의 확진자와 1,500명 이상의 사망자가 발생하였다[1]. 코로나19의 장기적인 대유행은 경제, 정치, 사회, 문화 등 다양한 범위를 망라한 우리의 삶에 광대한 영향을 끼치고 있으며, 사회적, 경제적 비용을 최소화하기 위한 다방면의 노력이 절실히 필요하다.

다양한 분야의 수많은 연구진들이 코로나19 관련 연구들을 수행하고 있으며 그 결과물들이 쏟아져 나오고 있다. Stephany 외5[2]에서는 코로나19 팬데믹의 경제적인 영향을 검토하고 있는데, 산업별로 위험지수를 개발하여 제공하고 있다. 이를 위해 기업들이 미국 증권 거래위원회에 제출한 서류를 활용하여 잠재 디리슐레 할당(Latent Dirichlet Allocation, LDA) 토픽 모델링, 감성분석, 데이터 마이닝 기법을 적용하고 있다. 반면 del Rio-Chanona 외4[3]은 독일에서의 산업별 노동시장에 미치는 영향을 추정하고 있는데, 가장 타격이 큰 산업은 여행업, 음식업이며 상대적으로 보건의료업은 전면 가동중인 것으로 나타났다. Ramelli와 Wagner[4]는 구글 검색 결과 및 주식시장 데이터로 팬데믹 동안의 각 산업별 영향을 측정하는 방법을 제시하고 있다.

반면 1,472개의 진료 노트를 활용하여 Lybarger 외3[5]에서는 코로나19 관련 증상을 식별하는 모형을 제시하고 있는데, F1 점수가 0.9 가량인 것으로 나타났다. F1 점수는 분류 성능을 평가하는 지표 중 하나이다. 한편 주로 의학, 생물 관련 논문들이 등록되어 있는 PubMed Central, bioRxiv, medRxiv의 약 57,000편의 코로나19 관련 논문들을 바탕으로 Cheng 외2[6]에서는 LDA를 적용하여 6개의 연구 토픽-확진자수, 검출, 공공보건의료, 바이러스의 유전자 구조, 환자 간호, 임상증상 및 치료-을 발견하고 있다.

소셜미디어나 인터넷의 텍스트 데이터를 기반으로 한 연구들도 많이 이루어지고 있다. Bettencourt-Silva 외6[7]은 Google Trends를 이용해서 건강의 사회적 결정요인(social determinants of health)을 검토하고 있는데, 팬데믹 이후 실업과 식량부족이 가장 큰 요인인 것으로 나타났다. Walker 외2[8]에서는 Google Trends를 이용해 여러 나라의 사람들이 검색한 코로나19 관련 증상에 관해 연구하고 있다. 한편 Garcia와 Berton[9]는 2020년 4월부터 8월까지 미국과 브라질의 트위터 사용자들이 작성한 코로나19 관련 텍스트를 분석하고 있다. 미국의 경우 영어로 작성된 3,332,565개의 트윗, 브라질의 경우 포르투갈어로 작성된 3,155,277개의 트윗을 분석 대상으로 삼고 있으며, 각각에 대해서 10개의 토픽을 발견하고 있다. 미국 트윗은 크게 확진자 통계, 예방 관리, 치료, 정치, 경제적 영향, 오락, 스포츠, 온라인 이벤트, 구호 자선, 반인종차별 시위로 토픽이 나뉘며, 반면 브라질의 트윗은 7개의 겹치는 토픽을 포함하여 확진자 통계, 예방 관리, 치료, 정치, 경제적 영향, 오락, 스포츠, 교육 및 문화, 건강 및 미용, 일상생활로 나뉜다. 감성분석 결과 대부분의 토픽에서 부정적인 어조가 강한 것으로 나타났다. 흥미로운 점은 SNS 자료를 이용한 분석의 많은 경우, 코로나 초기에는 상대적으로 긍정적인 어조가 강했다면 점차 시간이 지날수록 그리고 리트윗되는 트윗일수록 부정적인 어조가 강하게 나타나 팬데믹 상황이 오래 지속됨에 따라 사람들이 심리적으로 어려움을 겪고 있다는 것을 보여준다[10,11]. Brooks 외6[12]는 팬데믹 이후 심리적인 영향을 분석하고 있다. 스트레스의 요인으로는 장기 격리, 감염 위험, 좌절감, 무료함, 공급 부족, 허위 정보, 재정적 손실 등이었다. 확진자 통계, 치료, 경제적 영향 관련하여 가장 큰 부정적인 감정은 공포로 나타났으며, 예방 관리, 정치와 관련한 가장 큰 부정적인 감정은 분노로 나타났다.

텍스트 데이터의 토픽을 분류하는 데 있어서도 잠재 디리슐레 할당(LDA), 구조적 토픽 모델링(Structural Topic Model, STM), 비음수 행렬 분해(Non-Negative Matrix Factorization, NMF) 등 다양한 모형이 사용되었다[13-15].

국내에서도 여러 분야에서 다양한 연구들이 진행되고 있다. 이동훈 외5[16]은 실험을 통하여 코로나19 감염에 대한 우울과 불안에 미치는 요인들을 밝히고 있다. 성별, 경제 수준, 평소의 심리 상태, 가족과의 관계 등이 유의한 변인으로 나타났다. 김은정 외3[17]에서는 2020년 1월부터 8월까지 언론 기사를 이용하여 코로나19 이슈의 흐름을 살펴보고 있으며, 김용회[18]은 신문 기사를 바탕으로 코로나19 이후 사회갈등 이슈를 검토하고 전략적 방안을 논의하고 있다.

특정 대학에서의 설문조사를 통해 비대면 교육의 만족도와 수강지속 의사에 영향을 미치는 요인을 분석한 송수연과 김한경[19]는 언택트 시대의 교육환경 및 학생들의 수요를 충족시킬 수 있는 대학 교육에 대한 제언을 제시하고 있다. 정치적인 측면에서의 연구로서 김상배[20]은 팬데믹 발생 후 세계정치의 복합지정학적 동학을 분

석하고 이에 신흥안보에 대해 논의하고 있다. 사회경제적인 영향을 검토한 이민우와 유지은[21]은 소비지출과 노동시장을 중심으로 코로나19의 영향을 분석하고 산업정책에 대한 함의를 모색하고 있다. Kim 외4[22]는 코로나19 바이러스의 유전체 구조, 유전자 발현, 진단 기술에 대한 정보를 제공하고 있으며, 오형근[23]은 팬데믹이 가져온 주요 정보보안 이슈를 분석하고 안전한 디지털 시스템을 구축하기 위한 방안을 제시하고 있다.

텍스트 마이닝을 활용한 연구도 많이 이루어지고 있다. 이새미 외2[24]에서는 뉴스 기사와 SNS 텍스트를 이용해 마스크 5부제에 대한 언론과 대중들의 반응을 비교분석하고 있다. Heo와 Yang[15] 역시 텍스트 마이닝 기법을 활용하고 있으며 코로나19 관련 논문을 대상으로 7개의 연구 토픽을 도출하고 각 토픽별 추세를 살펴보고 있다.

다양한 분야에서 수많은 연구들이 진행, 발표되고 있는 상황에서 각 논문들의 내용을 일일이 검토하는 것은 어려운 일이다. 이에 본 연구에서는 텍스트 마이닝 기법을 이용해 한 눈에 관련 연구의 동향과 추이를 제시하고자 한다. 본 연구는 Heo와 Yang[15]를 확장, 강화한 연구라 할 수 있는데, 먼저 2020년 10월 10일까지의 총 290편의 논문을 분석대상으로 삼고 있는 Heo와 Yang[15]에 비해 본 연구는 12월 30일까지의 최근 논문을 포함한 총 571편의 연구 논문을 대상으로 한다. 또한 Heo와 Yang[15]에서는 LDA를 이용한 토픽모델링이 사용되었으나 비지도 학습에 기반을 두기 때문에 검증과 해석이 어렵다는 한계가 있다. 이에 본 연구에서는 토픽모델링에서 가장 많이 쓰이는 LDA와 STM의 결과를 비교하고 일치성을 확인함으로써 검증 단계를 추가하였다. 그리고 본 연구에서는 토픽별로 세부적인 연구 주제를 추가 분류하고 감성 분석을 시행함으로써, 해당 분야에 새로이 진입하는 연구자들에게는 연구 방향성을 제시하고 정책결정자들에게는 연구지원 및 인력확보 정책의 가이드라인을 제시할 수 있으리라 기대한다.

2. 이론적 배경

2.1 토픽모델링

잠재 디리슐레 할당(LDA)은 토픽모델링의 대표적인 방법 중 하나로, 최근 텍스트 데이터를 분석하는 많은 연구논문에서 사용되고 있다. LDA는 토픽의 단어 비중과 문서의 토픽 비중의 결합 확률분포를 가정하여, 깁스 샘플링(Gibbs sampling)을 통해 반복적으로 토픽을 추출하고 업데이트하는 과정을 거친다[25,26]. 좀 더 자세한 내용은 Heo와 Yang[15]를 참조할 수 있다. 반면 구조적 토픽 모델링(STM)은 문서에 대한 메타데이터를 활용하여 각 문서 안에 존재하는 단어들의 빈도수를 기반으로 토픽들을 추출한다[27,28]. STM은 외부 변수의 영향을 문서별 토픽 분포와 토픽별 단어 분포에 반영해 모델의 토픽 및 단어 분포를 추정할 수 있다는 장점이 있다.

토픽 모델링을 적용하기 위해서는 사전에 토픽의 개수를 정할 필요가 있는데, 이를 위해 토픽들 간의 코사인 거리[29], 토픽-단어 행렬을 이용해 구한 KL-거리[30], 로그우도[31], 토픽간 젠슨-샤논 거리[32]를 사용할 수 있을 것이다. 코사인 거리 및 KL-거리는 작을수록, 로그우도 및 젠슨-샤논 거리는 클수록 최적의 값이다.

2.2 코더간 일치성

크리펜돌프의 알파(Krippendorff's alpha)는 코더간 일치 정도를 측정하는데, 본 연구에서는 이를 이용하여 LDA와 STM, 두 토픽 알고리즘 코더의 판단이 얼마나 일치하는지를 검토하고 있다. 크리펜돌프의 알파는 0과 1 사이의 값을 가지며, 1에 가까울수록 코더 간 분류결과가 일치함을 나타낸다. 카테고리의 개수, 코더의 개수, 데이터의 척도 유형에 제한을 받지 않아서 내용분석 데이터의 신뢰도 측정에 많이 쓰이고 있다[33,34].

2.3 k-평균 군집 알고리즘

k-평균 군집 알고리즘(k-means clustering algorithm)은 주어진 데이터를 군집간 분산과 군집내 분산을 고려하여 k개의 군집으로 분류하는 알고리즘으로, 이를 활용하여 문서들의 토픽을 분류할 때에는 한 문서에 한 개의 토픽이 존재한다고 가정한다[35]. 본 연구에서는 키워드를 이용하여 토픽별 세부 연구주제를 탐색할 때 k-평균 군집 알고리즘을 활용하고 있는데, 키워드는 토픽에 직접적으로 연관되는 단어들로 구성되어 있다고 판단했기 때문이다.

군집 수(k)는 한 군집 내 데이터들이 다른 군집에 비해 얼마나 비슷한지를 나타내는 측도인 실루엣(silhouette) 점수를 이용하여 결정하고, 문서 간 거리는 코사인(cosine)으로 정의할 수 있다[36]. 또한 군집을 시각적으로 나타내기 위해, 주성분 분석(principal component analysis, PCA)을 적용하여 2차원의 벡터로 압축할 수 있다[37].

2.4 감성분석

감성분석은 문서에 나타난 주관적 요소인 긍정적 또는 부정적 감정을 분석하는 과정으로, 주로 많이 사용되는 감성어 사전으로 Bing과 AFINN이 있다. Bing은 총 6,786개의 단어 각각에 대해서 긍정적, 부정적인 단어로 구분하고 있다[38]. 긍정적 단어에 +1, 부정적 단어에 -1을 부여한 후, 이들을 모두 더한 값을 감성점수로 정의한다. 즉 감성점수는 긍정적인 단어와 부정적인 단어의 개수의 차이로, 0에 가까우면 중립적인 문서로 볼 수 있다.

반면 AFINN은 총 2,477개의 단어에 대해 -5~5 사이의 정수를 척도로 사용해 긍정적, 부정적인 강도를 부여한다. 5에 가까울수록 긍정의 강도가 강하며, -5에 가까울수록 부정의 강도가 강하다[39]. 감성점수는 이 척도들을 합한 값이며 양수이면서 값이 클수록 긍정적인 어조가 강하며, 음수이면서 값이 작을수록 부정적인 어조가 강하다고 볼 수 있다.

3. 연구 방법

3.1 연구자료 수집 및 전처리 과정

디비피아(DBpia, http://www.dbpia.co.kr/)에서 “코로나19”, “covid19”를 키워드로 하여 2020년 10월 10일자로 검색된 논문은 총 543개였다. 이 후 12월을 기점으로 코로나가 다시 유행하면서 논문의 개수도 급증하여, 추가적으로 논문의 제목, 초록, 키워드 등을 텍스트화 하여 데이터를 수집하였다. 2020년 12월 30일자로 논문은 총 930개가 검색되었고, 세부적으로는 사회과학 분야에서 129개, 인문학 분야에서 49개, 공학 22개, 복합학 18개, 예술체육학 17개, 의약학 15개, 농수해양학 7개, 자연과학 6개의 자료가 10월 10일 이후로 출판되었다. 주된 분석 대상을 영문 초록으로 선정하였기 때문에, 초록을 한글로 작성한 논문, 초록이 없는 논문 등은 제외하고 571개의 논문을 대상으로 분석을 진행하였다. 전처리 과정으로 소문자 변환, 품사의 통일화, 특수 문자 제거, 조사와 빈도수가 10 이하의 단어들을 제거하여 총 1,147개의 단어들을 추출하였다.

3.2 자료 분석 방법

코로나19 관련 연구논문에 발현된 토픽을 찾기 위해, 가장 많이 쓰이는 잠재 디리슐레 할당(LDA)을 사용하고 있다. 하지만 비지도 학습(unsupervised learning)에 기반하기 때문에 분류가 얼마나 잘 되었는지 검증하기가 어려우며, 이에 STM을 추가 적용해서 LDA와의 결과와 비교한 후 크리펜돌프의 알파를 계산하여 분류 성능을 평가한다.

한편 해석이 용이하면서 타당한 잠재 토픽의 개수를 사전에 결정하기 위해서 네 가지 기준 값(토픽들 간의 코사인 거리, 토픽-단어 행렬을 이용해 구한 KL-거리, 로그우도, 토픽간 젠슨-샤논 거리)을 이용한다.

영문 초록을 이용해 논문에 발현된 잠재토픽을 찾은 후에 각 토픽별로 세부 연구토픽을 찾고 있다. 단 세부 토픽의 경우 키워드를 이용해서 k-평균 군집 알고리즘으로 분류하고 있다. LDA를 비롯한 토픽 모델링은 한 문서에 여러 토픽이 존재한다고 가정하지만, k-means 알고리즘은 한 문서에 한 개의 토픽이 존재한다고 가정한다[35]. 일반적으로 키워드는 토픽에 직접적으로 연관되는 단어들로 이루어지기 때문에, 키워드를 사용해 세부 연구토픽을 찾을 때에는 k-means 알고리즘이 적절하다고 판단했다.

또한 감성분석을 통해 각 토픽의 긍정적, 부정적 단어를 분기별로 살펴보고, 감성점수를 계산하여 주된 어조를 파악하고 있다.

4. 연구 결과

4.1 연구주제 식별 및 동향

우선 잠재토픽의 개수를 결정하기 위하여, 앞에서 설명한 네 가지 지표의 값을 검토하였다. Cao et al.과 Arun et al.이 사용한 기준 값은 작을 때, Griffiths et al.과 Deveaud et al.이 사용한 기준 값은 클 때 최적의 토픽의 개수가 정해진다. Fig. 1을 보면 네 가지 기준 값이 최적이 되는 토픽의 개수는 없지만, 토픽의 개수가 8일 때 Cao et al.의 기준 값은 현저히 작아지고 Deveaud et al.의 기준 값은 상대적으로 커지는 것을 볼 수 있다. 이러한 점과 해석의 용이성을 감안하여 토픽의 개수를 8로 정하였다.

LDA를 통해 발견한 8개의 토픽들에 대해서 상위 출현 단어들을 검토하여 다음과 같이 토픽의 이름을 명명하였다. (1) 경제적 영향 (2) 생물 의학 관련 (3) 사회적 보호 및 복지 (4) 국제적 역학관계 (5) 종교 관련 (6) 정보통신기술과 보안 (7) 심리적 영향 (8) 온라인 교육이다.

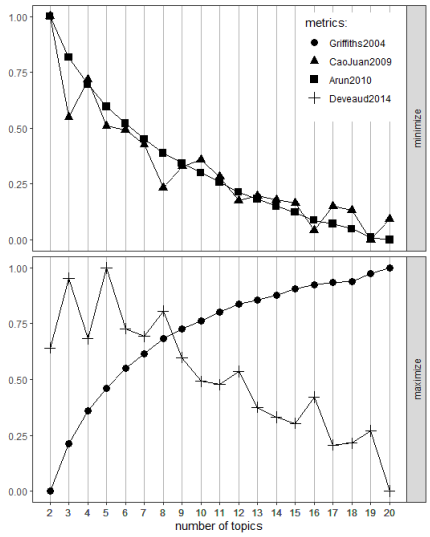

Fig. 1. Selection of the number of topics.

토픽별 주요 단어를 살펴보면, 첫 번째 토픽인 경제적 영향에는 industry, market, tourism, service, company, strategy, consume, management, economy의 단어가 자주 등장하였다. 두 번째 토픽에는 sars, patient, respiratory, cause, china, diagnostic, clinical, syndrome, mers, transmission, symptom의 단어가 많이 등장하여 토픽 이름을 생물 의학 관련이라 하였다. 세 번째 사회적 보호 및 복지 관련 주요 단어는 elderly, city, environment, support, care, necessity, space, family, management, government이다. 네 번째 토픽의 경우, china, international, nation, government, global, crisis, country, history, eu, cooperation, political, issue 등 국제적 역학 관계를 나타내는 단어가 많이 등장하였다. 다섯 번째 종교 관련 토픽의 주요 단어는 church, community, life, worship, crisis, faith, god, religious, situation, spread, change 이다. 여섯 번째 토픽에는 information, security, law, personal, management, prevent, legal, protect, regulation, datum, technology 등의 단어가 많이 나타나 정보통신기술과 보안이라 명명하였다. 일곱 번째 토픽은 perceive, factor, positive, behavior, self, conduct, stress, affect, survey, psychological 등 심리적 영향에 대한 단어가 주를 이루었다. 마지막 토픽인 온라인 교육의 주요 단어는 education, online, e-learning, student, class, teach, school, content, experience, university, program이다.

Table 1. The trend of topic appearance in percentages. Q1, Q2, Q3 and Q4 are the 1st, 2nd, 3rd and 4th quarter of 2020, respectively. T1: Economic Impact, T2: Biomedical Related, T3: Social Care & Welfare, T4: International Dynamics, T5: Religion Related, T6: Technology & Security, T7: Psychological Impact, T8: Online Education.

	Q1	Q2	Q3	Q4	Total
T1	0.00	5.63	4.36	4.17	14.16
T2	1.27	7.08	2.72	1.63	12.70
T3	0.36	1.63	2.54	4.36	8.89
T4	0.91	6.17	4.36	4.72	16.15
T5	0.18	3.81	2.72	4.54	11.25
T6	0.18	2.72	4.17	2.72	9.80
T7	0.18	2.90	4.72	6.35	14.16
T8	0.18	1.45	5.63	5.63	12.89
Total	3.27	31.40	31.22	34.12	100.00

기존 연구 Heo와 Yang[15]와 비교해서, 사회적 보호 및 복지 토픽이 추가되었다. 해당 토픽에 포함되는 논문들이 이전에는 많은 경우 정보통신기술과 보안 토픽에 포함되었다. Table 1은 2020년 분기별 각 토픽의 비율을 보여준다. 전체적으로 국제적 역학관계(16.15%), 경제적 영향(14.16%), 심리적 영향(14.16%) 순으로 많은 연구가 이루어졌음을 알 수 있다. 1분기에는 3.27%의 상대적으로 적은 논문이 게재되었는데, 코로나19 유행이후 연구를 진행하고 논문을 게재하는 데까지의 시간을 반영하는 것으로 판단된다.

주목할 만한 점은 생물 의학 관련 연구가 2분기(7.08%) 크게 증가하였으나 3분기(2.72%), 4분기(1.63%)에 크게 감소하여 관련 연구진들이 쉽게 접근할만한 국내 연구자료가 부족할 수 있음을 시사한다. 해당 분야의 국내 연구 활성화를 위한 정책적 지원이 필요할 것으로 생각된다. 반면 사회적 보호 및 복지 또는 심리적 영향 관련 연구의 비중은 증가하고 있는데, 이는 취약계층에 대한 지원, 심리적 지원의 필요성이 증대하고 있음을 시사하며 이에 대한 사회적, 정책적 관심과 노력이 요구된다. 국제적 역학관계, 경제적 영향 관련 연구는 2분기 이후 꾸준히 지속되고 있음을 확인할 수 있다.

LDA 결과의 타당성을 검토하기 위해, STM 결과와 비교하여 얼마나 일치하는지를 확인하였다. 크리펜돌프의 알파를 구한 결과, 0.71로 나타났다. 아주 높지는 않지만 어느 정도의 안정적인 수준을 확보했으며, 결과분석에 큰 무리가 없다고 판단된다[40,41]. Table 2는

LDA와 STM 결과를 비교하고 있는데, 많은 연구논문이 대각 방향에 나타나고 있어 대부분의 논문이 두 토픽 모델링에 의해서 동일한 토픽으로 분류되고 있음을 확인할 수 있다. 하지만 일부 논문에 대해서는 토픽이 일치하지 않는데, 이는 한 논문에 여러 내용이 혼재되어 있어 나타난 결과로 보인다.

Table 2. The number of research papers induced by LDA and STM. T1: Economic Impact, T2: Biomedical Related, T3: Social Care & Welfare, T4: International Dynamics, T5: Religion Related, T6: Technology & Security, T7: Psychological Impact, T8: Online Education.

		STM							
		T1	T2	T3	T4	T5	T6	T7	T8
LDA	T1	63	1	2	4	1	2	3	2
	T2	2	57	3	0	0	0	7	1
	T3	6	0	26	2	0	14	0	1
	T4	7	3	2	55	15	5	0	2
	T5	0	0	1	0	55	3	1	2
	T6	2	0	2	9	3	37	0	1
	T7	2	3	6	1	3	1	59	3
	T8	0	1	3	0	4	1	3	59

특히 노동자 및 취약 계층의 보호 정책 및 법적 이슈 관련 논문은 사회적 보호 및 복지 토픽에 분류되기도 하고 정보통신기술과 보안 토픽에 분류되기도 하였다. 국제적 역학관계 토픽과 경제적 영향 토픽도 한 논문에 혼재되어 있는 경우도 많았는데, 이는 코로나19로 인한 경제적인 효과가 국제적 통상 및 수출 전략 등에 복합적으로 연결되기 때문으로 보인다. 보안과 관련해서도 국가 안보 뿐 아니라 개인정보의 국외 이전 이슈가 많이 혼재되어 있어 정보통신기술과 보안 토픽과 국제적 역학관계 토픽이 한 논문에 복합적으로 등장하기도 하였다. 심리적인 영향과 관련해서는 취약계층에 대한 보호, 온라인 수업으로 인한 학생들의 심리적 어려움이 같이 논의되기도 하였다. 또한 감염병 확산과 종교집단의 대응, 특정 종교에 대한 국제사회에서의 포비아도 복합적으로 논의되기도 하였다.

4.2 세부 연구주제

영문 키워드를 이용하여 세부 연구주제를 k-means 알고리즘을 통해 분류하였다. 이에 대해 PCA를 적용하여 2차원으로 압축한 후 시각적으로 나타낸 결과를 Fig 2에서 확인할 수 있다.

먼저 경제적 영향 토픽은 크게 4개의 세부 연구주제로 나뉘는데, 관광, 교통, 항공, 음식 등 각 산업에 미치는 영향, 기업별 혁신 전략, 의미론적 네트워크 분석(semantic network analysis), 가치사슬(value chain) 등으로 분류된다. 생물 의학 관련해서는 역학 및 임상적 특징, 메타 분석, 급성호흡기증후군(acute respiratory syndrome), 유전자 분석 등의 세부 연구가 이루어지고 있다. 사회적 보호 및 복지 관련해서는 취약계층(노동자, 노인, 청소년 등)의 보호 및 복지, 지역사회의 건강도시 조성 정책, 장애인 지원 등의 세부 연구가 이루어지는 것으로 보인다. 국제적 역학관계 토픽은 국제정세, 국제정치 전략, 동아시아의 역사로 세부 분류될 수 있을 것이다. 반면 종교 관련해서는 크게 신학적 성찰 및 리더십의 방향성, 포스트 코로나 시대의 종교적 대응, 질병 대응으로 분류될 수 있다. 정보통신기술과 보안 토픽에서는 디지털 기술과 법적 책임, 환자 정보 보호, 의료자원의 분배(health-care resource allocation) 등이 논의되고 있다. 심리와 관련해서는 크게 정신 건강에 미치는 영향, 행동에 미치는 영향, 언론보도의 역할로 세부 분류될 수 있을 것이다. 마지막으로 온라인 교육 토픽은 비대면 수업의 환경 및 만족도, 비대면 토론수업, 혼합형 학습에 관해 세부적으로 논의되고 있다.

기본적으로 영문 키워드를 중심으로 k-means 알고리즘을 활용하여 분류한 것이지만, 세부 연구주제의 주된 내용을 파악하기 위해서 연구 논문의 제목, 초록도 추가적으로 같이 검토하였다. 좀 더 많은 분석자료(연구논문)가 축적된다면 보다 정확하고 자세한 세부 주제를 파악할 수 있으리라 판단된다. 세부 연구주제는 해당 분야에 새로이 진입하는 연구자들이 연구의 방향성을 탐색하고 정책결정자들이 연구지원 세부사업을 결정하는 데 참조할 수 있는 기초자료로 활용될 수 있을 것이다. 또한 DBpia를 비롯한 연구논문 검색 사이트에서 이러한 세부 연구주제의 결과를 활용하여 관련된 주제의 논문의 목록을 같이 제공한다면 검색의 효율성 및 정확성을 높일 수 있으리라 사료된다.

4.3 감성분석 결과

감성분석을 통해 긍정적, 부정적인 단어를 판단한 후 각 토픽의 분기별 감성점수를 계산하였다. 감성점수가 양수이고 값이 클수록 긍정적인 어조가 강하고, 음수이면서 값이 작을수록 부정적인 어조가 강하다는 것을 의미한다.

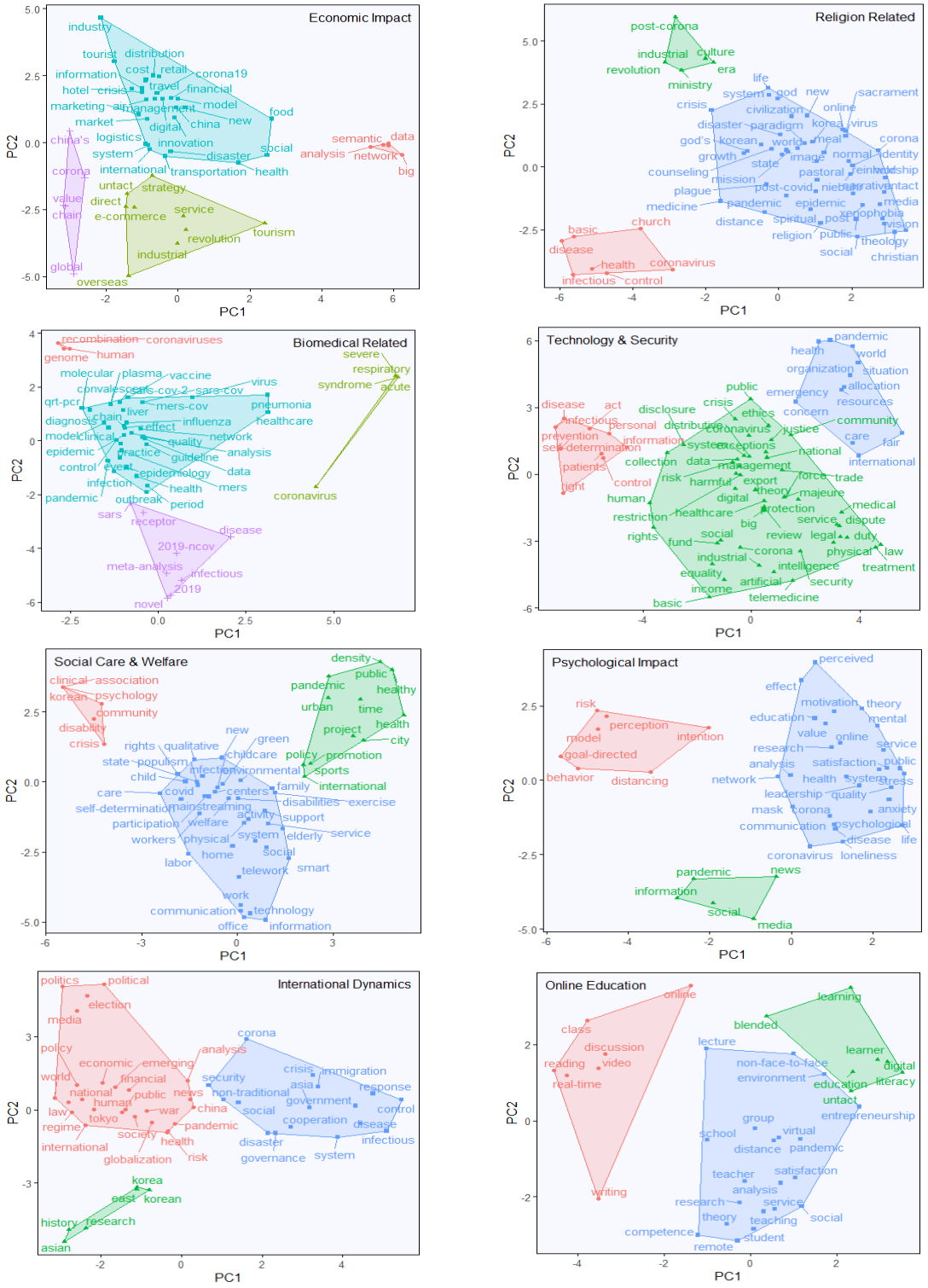

Fig. 2. Identification of subtopics.

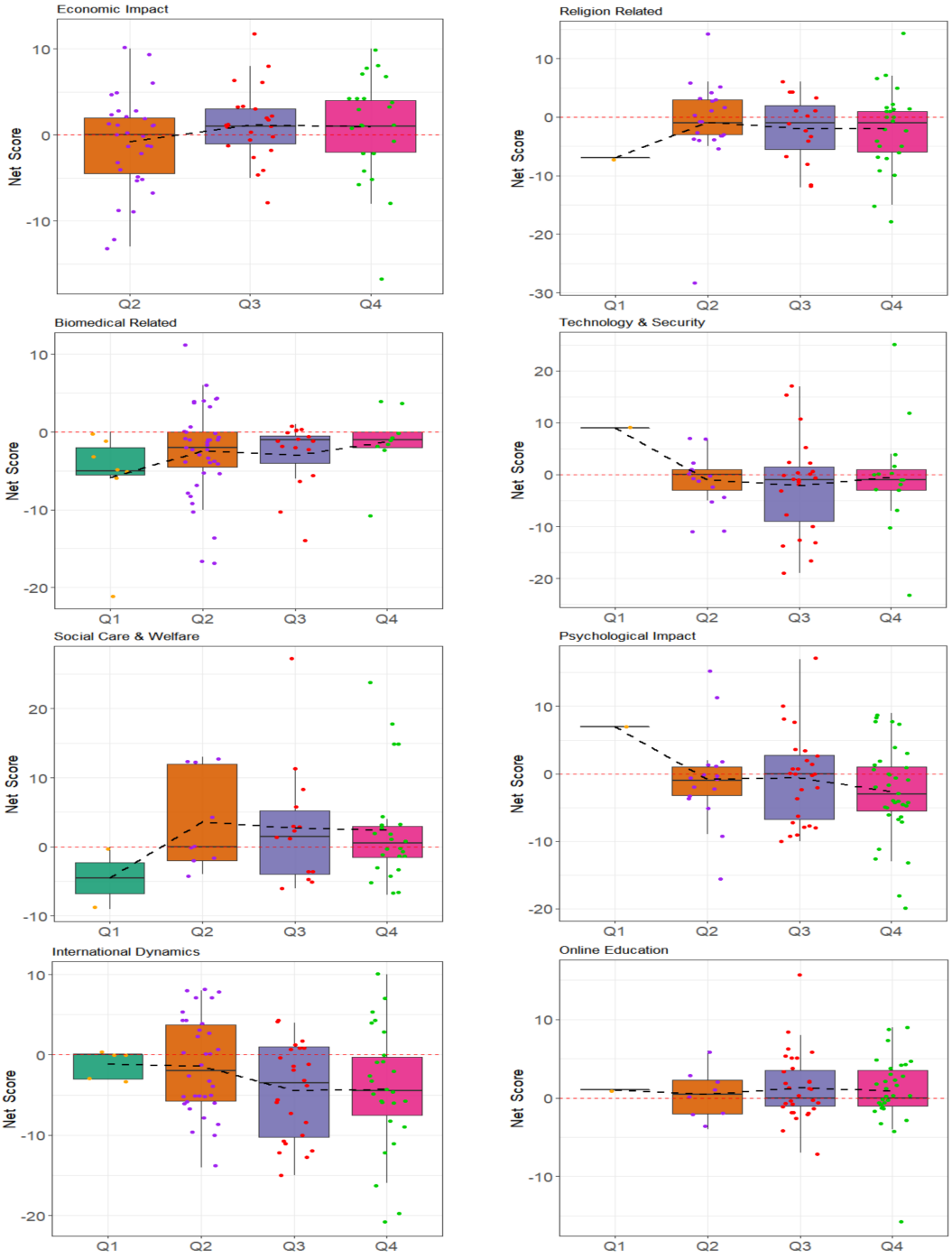

Fig. 3. Sentiment analysis using Bing lexicon. A net score is computed as the difference between positive and negative word counts, with positive(negative) values implying positive(negative) stances and values close to zero implying neutral stances. The score averages of the four quarters of 2020 are connected with lines.

Fig. 3에서 보듯이, 경제적 영향의 경우 2분기에 부정적인 어조였지만 3, 4분기에 와서는 전반적으로 긍정적인 어조로 기술되고 있었다. 하지만 통계적 유의성은 나타나지 않았다. 1분기에 경제적 영향 토픽으로 분류된 연구논문은 없었다. 많이 등장하는 긍정적인 단어로는 support, rapid, improve, sustainability, free, innovation, intelligence, recovery, revitalize, survival, cooperative 등이며, 부정적인 단어는 burden, liability, loss, passive, recession, restriction, shortage, backward, blow, decline, depression 등이 많이 나타났다.

생물 의학 관련 분야에서는 모든 분기에서 부정적인 어조가 강하였다. 4분기에 들어 부정적인 어조가 완화되고는 있지만, 그 전인 1, 2, 3분기에는 통계적으로 유의한 음의 값을 보였다. fast, protect, protective, timely, confidence, consistent, leading 등의 긍정적인 단어, severe, outbreak, syndrome, symptom, emergency, death, illness, infected, fever, isolation, complication, devastating 등의 부정적인 단어가 많이 사용되었다.

사회적 보호 및 복지 토픽은 1분기 이후에 긍정적인 어조로 바뀌었는데, 통계적 유의성은 없는 것으로 나타났다. support, suitable, protect, facilitate, guarantee, reward, cure, enhance 등의 긍정적인 단어와 disabled, difficulty, isolation, concern, emergency, limit, discrimination, restriction, lack, alienated 등의 부정적인 단어가 많이 나타났다.

국제적 역학관계 관련해서는 모든 분기에서 부정적인 어조가 강하였는데, 특히 3, 4분기에는 통계적으로 유의한 음의 값을 보였다. 3, 4분기에 연구결과의 부정적인 측면이 크게 반영되었음을 알 수 있다. 긍정적인 단어로는 diplomatic, achievement, golden, progressive, protect, dynamic, variety 등이 쓰였으며, 부정적인 단어로는 threat, terrorism, complex, conservative, dispute, attack, conflict, difficulty, cold, racism 등이 많이 쓰였음을 확인할 수 있다.

반면 종교 분야에서는 모든 분기에서 부정적인 어조였으나 통계적으로 유의하지는 않았다. 긍정적인 단어로는 faith, revival, spiritual, maturity, unity, innovation, leading, dignity, grace, pride, reform 등이 쓰였으며, 부정적인 단어로는 plague, limit, anxiety, conflict, deaf, weakening, blame, catastrophe, restriction, selfishness, fake 등이 자주 쓰였다.

정보통신기술과 보안 토픽 관련해서는 전반적으로 부정적인 어조로 기술되었으나 역시 통계적으로 유의하지는 않았다. protect, intelligence, improve, autonomous, effective, rapid, ethical, improving, advanced 등의 긍정적인 단어와 limit, restriction, concern, illegal, violation, conflict, excessive, failure, infringement 등의 부정적인 단어가 많이 사용되었다.

심리적 영향과 관련해서는 시간이 갈수록 부정적인 어조가 강해지고 있으며, 특히 4분기에는 감성지수의 값이 통계적으로 유의한 음수로 나타났다. positively, improve, recommendation, protective, optimism, commitment, affirmative, comfort, confidence, contribution 등의 긍정적인 단어와 stress, anxiety, loneliness, fear, depression, shock, poverty, discomfort, helplessness, poor, vulnerable 등의 부정적인 단어가 많이 나타났음을 확인할 수 있다.

마지막으로 온라인 교육 분야에서는 긍정적인 어조가 더 강하지만 통계적인 유의성은 없었다. 긍정적인 단어로는 effective, realistic, effectiveness, smart, convenience, flexible, complement, integrated 등이 많이 쓰였으며, 부정적인 단어로는 difficulty, hard, neglected, unprepared, confused, delayed, failure, negligence 등이 많이 쓰였다.

Fig. 3은 감성사전으로 Bing을 사용한 결과이지만 AFINN을 사용했을 경우에도 유사한 결과를 보였다. 지면 관계상 AFINN의 결과는 제시하지 않는다.

5. 논의 및 결론

코로나19로 전 세계가 유례없는 상황을 헤쳐 나가고 있다. 불안과 두려움 속에서 잘못된 정보와 루머는 사회의 혼란과 분열을 가중시킬 수 있다. 이러한 위기 상황을 극복하기 위해 각 분야의 연구자들의 지속적인 노력이 필요한 상황이다. 정확하고 신뢰할 수 있는 결과를 신속하게 공개하여 우리 사회가 객관적이고 현실적인 시야를 유지할 수 있도록 기여해야 한다.

현재까지 다양한 분야의 수많은 국내 연구진들이 코로나19 관련 연구논문을 발표하고 있다. 본 연구에서는 이러한 연구 결과물을 대상으로 잠재 디리슐레 할당(LDA)을 이용해 연구주제를 탐색하였다. 이전 분석결과([15])와 비교해 사회적 보호 및 복지 토픽이 추가되어 다음의

총 8개의 토픽을 발견하였다. (1) 경제적 영향 (2) 생물의학 관련 (3) 사회적 보호 및 복지 (4) 국제적 역학관계 (5) 종교 관련 (6) 정보통신기술과 보안 (7) 심리적 영향 (8) 온라인 교육. 이러한 결과는 구조적 토픽 모델링(STM)의 결과와 비교해서 비교적 안정적이고 용인 가능한 수준인 것으로 나타났다. 또한 k-means 알고리즘을 통해 각 토픽별로 세부 연구주제를 발견하였고 주성분 분석(PCA)을 이용하여 시각적으로 표현하였다. 감성분석을 통해 각 토픽별 긍정적, 부정적인 단어들을 살펴보고 감성점수를 계산하여 연구논문의 주된 어조를 파악하고 있다. 특히 생물 의학 관련, 국제적 역학관계, 심리적 영향과 관련된 연구에서 부정적인 어조가 강한 것으로 나타나 해당 부문에 대해서 정책결정자 및 전문가들의 주의와 관심이 요구된다.

본 연구의 학술적 시사점은 비정형 데이터인 코로나19 관련 연구논문을 텍스트 마이닝 기법인 토픽 모델링과 감성 분석 기법을 통해 분석하고, 특별한 관심을 요하는 연구주제를 탐색하였다는 데에 의의가 있다. 많은 양의 문헌들을 일일이 검토하는 것은 어려운 일이다. 하지만 텍스트 마이닝 기법을 사용할 경우 연구주제의 토픽을 한눈에 파악하고 동향을 살펴보며 경고가 되는 용어들을 추출할 수 있는 이점이 있다. 이는 연구자들이 연구의 방향성을 탐색하고 정책결정자들이 연구지원 사업을 결정하는 데 활용될 수 있을 것이다. 또한 관련 논문의 결과들의 유사점과 차이점을 비교하고 검증하는 데 유용하게 활용될 수 있으며, 연구들이 상반된 결과를 제시할 때 이를 통계적으로 통합하여 좀 더 일반적인 결과를 도출할 수 있는 가능성을 제시한다는 데 학술적인 의의를 찾을 수 있다.

본 연구에서는 DBpia에 등록된 연구논문을 대상으로 하고 있지만, 향후 다른 연구논문 검색 사이트를 이용해 분석 자료를 확대할 수 있을 것이다. 또한 연구논문 검색 사이트에 토픽별 논문 리스트를 추가 제공할 수 있다면 좀 더 빠르고 정확하게 관련 문헌을 검색할 수 있으리라 생각된다. 본 연구는 분석대상을 국내 연구에 한정하였다는 한계가 있으며, 방대한 양이겠지만 국외 논문으로까지 분석대상을 확대시킨다면 국가별 시계열 분석을 통한 흥미로운 연구를 수행할 수 있으리라 기대한다.

REFERENCES

[1] Ministry of Health and Welfare, http://ncov.mohw.go.kr/

[2] F. Stephany, N. Stoehr, P. Darius, L. Neuhäuser, O. Teutloff & F. Braesemann. (2020). The CoRisk-Index: A data-mining approach to identify industry-specific risk assessments related to COVID-19 in real-time. *arXiv preprint arXiv:*2003.12432.

[3] R. M. del Rio-Chanona, P. Mealy, A. Pichler, F. Lafond & J. D. Farmer. (2020). Supply and demand shocks in the COVID-19 pandemic: An industry and occupation perspective. *Oxford Review of Economic Policy, 36(Supplement_1),* 94-137.

[4] S. Ramelli & A. Wagner. (2020). What the stock market tells us about the consequences of COVID-19. *Mitigating the COVID Economic Crisis: Act Fast and Do Whatever,* 63-70.

[5] K. Lybarger, M. Ostendorf, M. Thompson & M. Yetisgen. (2020). Extracting covid-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. *arXiv preprint arXiv:*2012.00974.

[6] X. Cheng, Q. Cao & S. S. Liao. (2020). An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation. *Journal of Information Science,* 1-17. DOI : 10.1177/0165551520954674

[7] J. H. Bettencourt-Silva et al. (2020). Exploring the Social Drivers of Health During a Pandemic: Leveraging Knowledge Graphs and Population Trends in COVID-19. *Studies in Health Technology and Informatics, 275,* 6-11. DOI : 10.3233/SHTI200684

[8] A. Walker, C. Hopkins & P. Surda. (2020). Use of Google Trends to investigate loss-of-smell–related searches during the COVID-19 outbreak. *In International forum of allergy & rhinology, 10(7),* 839-847. DOI : 10.1002/alr.22580

[9] K. Garcia & L. Berton. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing, 101.* DOI : 10.1016/j.asoc.2020.107057

[10] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi & Z. Shah. (2020). Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. *Journal of medical Internet research, 22(4).* DOI : 10.2196/19016

[11] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag & A. E. Hassanien. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing, 97.* DOI : 10.1016/j.asoc.2020.106754

[12] S. K. Brooks et al. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The lancet, 395(10227),* 912-920. DOI : 10.1016/S0140-6736(20)30460-8

[13] A. Küsters & E. Garrido. (2020). Mining PIGS. A structural topic model analysis of Southern Europe

based on the German newspaper Die Zeit (1946-2009). *Journal of Contemporary European Studies, 28(4),* 477-493.
DOI : 10.1080/14782804.2020.1784112

[14] B. M'sik & B. M. Casablanca. (2020). Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. *International Journal, 9(4).* DOI : 10.30534/ijatcse/2020/231942020

[15] S. M. Heo & J. Y. Yang. (2020). Analysis of Research Topics and Trends on COVID-19 in Korea Using Latent Dirichlet Allocation (LDA). *Journal of The Korea Society of Computer and Information, 25(12),* 83-91. DOI : 10.9708/jksci.2020.25.12.083

[16] D. H. Lee, Y. J. Kim, D. H. Lee, H. H. Hwang, S. K. Nam & J. Y. Kim. (2020). The Influence of Public Fear, and Psycho-social Experiences during the Coronavirus Disease 2019(COVID-19) Pandemic on Depression and Anxiety in South Korea. *The Korean Journal of Counseling and Psychotherapy, 32(4),* 2119-2156. DOI : 10.23844/kjcp.2020.11.32.4.2119

[17] E. J. Kim, H. M. Sim, J. W. Won & B. J. Kang. (2020). Mapping the COVID-19 Issues from an Urban Perspective in South Korea - Text Mining Analysis Focused on Newspaper Articles. *Journal of the Urban Design Institute of Korea Urban Design, 21(6),* 163-179. DOI : 10.38195/judik.2020.12.21.6.163

[18] Y. H. Kim. (2020). Exploration of social conflict issues and future signals since the outbreak of COVID-19 in Korea: Using the keywords of news articles. *In conference of Korean Academy of Social Welfare,* 565-589.

[19] S. Y. Song & H. K. Kim. (2020). Exploring Factors Influencing College Students' Satisfaction and Persistent Intention to Take Non-Face-to-Face Courses during the COVID-19 Pandemic. *Asian Journal of Education, 21(4),* 1099-1126.
DOI : 10.15753/aje.2020.12.21.4.1099

[20] S. B. Kim. (2020). COVID-19 and the Complex Geopolitics of Emerging Security : The Emergence of Pandemic and the Transformation of World Politics. *Korean Political Science Review, 54(4),* 53-81.
DOI : 10.18854/kpsr.2020.54.4.003

[21] M. W. Lee & J. E. You. (2020). The Socio-Economic Effects of COVID-19: Focusing on Consumer Expenditure and Labor Market. *Asia-Pacific Journal of Business & Commerce, 12(3),* 121-141.
DOI : 10.35183/ajbc.2020.11.12.3.121

[22] J. S. Kim, N. K. Kang, S. M. Park, E. J. Lee & K. T. Chung. (2020). Diagnostic Techniques for SARS-CoV-2 Detection. *Journal of Life Science, 30(8),* 731-741. DOI : 10.5352/JLS.2020.30.8.731

[23] H. G. Oh. (2020). Analysis of major social changes and information security issues after COVID-19. *Communications of the Korean Institute of Information Scientists and Engineers, 38(9),* 48-56.

[24] S. M. Lee, S. E. Ryu. & S. J. Ahn. (2020). Mass Media and Social Media Agenda Analysis Using Text Mining : focused on '5-day Rotation Mask Distribution System'. *JOURNAL OF THE KOREA CONTENTS ASSOCIATION. 20(6),* 460-469.
DOI : 10.5392/JKCA.2020.20.06.460

[25] D. M. Blei, A. Y. Ng & M. I. Jordan. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3,* 993-1022.
DOI : 10.1162/jmlr.2003.3.4-5.993

[26] J. Y. Yang. (2019). Convergence Study on Research Topics for Thyroid Cancer in Korea. *Journal of the Korea Convergence Society, 10(2),* 75-81.
DOI : 10.15207/JKCS.2019.10.2.075

[27] M. E. Roberts, B. M., Stewart & E. M. Airoldi. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association, 111(515),* 988-1003.
DOI : 10.1080/01621459.2016.1141684

[28] M. E. Roberts, B. M. Stewart & D. Tingley. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software, 91(1),* 1-40.
DOI : 10.18637/jss.v091.i02

[29] J. Cao, T. Xia, J. Li, Y. Zhang & S. Tang. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72(7-9),* 1775-1781.
DOI : 10.1016/j.neucom.2008.06.011

[30] R. Arun, V. Suresh, C. V. Madhavan & M. N. Murthy. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. *In Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Berlin, Heidelberg. : Springer. DOI : 10.1007/978-3-642-13657-3_43

[31] T. L. Griffiths & M. Steyvers. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences, 101*(suppl 1), 5228-5235.
DOI: 10.1073/pnas.0307752101

[32] R. Deveaud, E. SanJuan & P. Bellot. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique, 17(1),* 61-84. DOI : 10.3166/DN.17.1.61-84

[33] K. Krippendorff. (2018). *Content analysis: An introduction to its methodology.* Los Angeles : Sage publications.

[34] A. F. Hayes & K. Krippendorff. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures, 1(1),* 77-89.
DOI : 10.1080/19312450709336664

[35] C. Buchta, M. Kober, I. Feinerer & K. Hornik. (2012). Spherical k-means clustering. *Journal of Statistical Software, 50(10),* 1-22.

[36] P. J. Rousseeuw. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20,* 53-65. DOI : 10.1016/0377-0427(87)90125-7

[37] I. T. Jolliffe. (2002). *Principal Component Analysis.* New York : Springer-Verlag

[38] M. Hu & B. Liu. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). Seattle : KDD'04

[39] F. Å. Nielsen. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv*:1103.2903.

[40] H. M. Salihu, A. A. Salinas-Miranda, L. Hill & K. Chandler. (2013). Survival of pre-viable preterm infants in the United States: a systematic review and meta-analysis. *In Seminars in perinatology, 37(6),* 389-400. DOI : 10.1053/j.semperi.2013.06.021

[41] H. J. Song. et al. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication, 37(4),* 550-572. DOI : 10.1080/10584609.2020.1723752

허 성 민(Seong-Min Heo) [정회원]

· 2014년 3월 ~ 2021년 2월 : 금오공과대학교 응용수학과 학부생
· 2017년 3월 ~ 현재 : 금오공과대학교 Applied Statistics Laboratory의 학부연구원
· 관심분야 : text mining, data mining and machine learning
· E-Mail : cjsm03@kumoh.ac.kr

양 지 연(Ji-Yeon Yang) [정회원]

· 2010년 7월 : University of Illinois Urbana-Champaign 통계학 박사
· 2010년 7월 ~ 2011년 6월 : Claremont McKenna College 방문조교수
· 2011년 7월 ~ 2014년 2월 : MD Anderson Cancer Center 연구원
· 2014년 3월 ~ 현재 : 금오공과대학교 응용수학과 조교수, 부교수
· 관심분야 : Bayesian analysis, big data analytics and computational statistics.
· E-Mail : jyang@kumoh.ac.kr