EMJ THE EWHA MEDICAL JOURNAL

Original Article

Ewha Med J 2022;45(2):46-54
https://doi.org/10.12771/emj.2022.45.2.46
eISSN 2234-2591

텍스트 마이닝을 활용한 COVID-19에 대한 대중의 관심 주제와 정서 분석

권나현*, 오종민*, 하은희
이화여자대학교 의과대학 환경의학교실

Topic and Trends of Public Perception and Sentiments of COVID-19 Pandemic in South Korea: A Text Mining Approach

Nahyun Kwon*, Jongmin Oh*, Eunhee Ha
Department of Environmental Medicine, Ewha Womans University College of Medicine, Seoul, Korea

Objectives: Public health risks and anxiety have been increasing since the outbreak of Coronavirus disease 19 (COVID-19). The public expresses questions related to the COVID-19 issue through the web base. The aim of this study was to analyze public perception and sentiments of COVID-19 Pandemic in South Korea.
Methods: We collected the text data (questions: 252,181) related to COVID-19 from Naver Knowledge-iN during January 1, 2020 to December 31, 2020. The search keywords included related to COVID-19 using Korean words for "SARS-Cov-2", "COVID19", "COVID-19", "Wuhan pneumonia", "Coronavirus", "Corona". A topic modeling analysis was used to investigate and search trends of public perception. The sentiment analysis was conducted to analyze of public emotions in the questions related to COVID-19. We performed the Pearson's correlation analysis between daily number of COVID-19 cases and daily proportion of negative sentiment in documents related to COVID-19 by COVID-19 outbreak period.
Results: A total of 241,776 documents used in this study. The most frequent words in the documents to appear cough, symptoms, tests, confirmed patients, mask and etc. Twenty topics (COVID-test, Economy, School, Hospital/Diagnose, Travel/Overseas, Health, Social issue, Symptom 1 (respiratory), Relationships, Symptom 2 (e.g., fever), Workplace, Mask/Social distancing, infection/Vaccine, Stimulus Package, Family, Delivery Service, Unclassified, Region, Study/Exam, Worry, Anxiety) were extracted using the topic modeling. There was a positive association between the daily counts of COVID-19 patients and proportion of negative sentiment. By COVID-19 period, Stage 4 had the highest correlation.
Conclusion: This study identified the South Korean public's interest and emotions about COVID-19 during the prolonged pandemic crisis. **(Ewha Med J 2022;45(2):46-54)**

Received January 24, 2022
Revised April 15, 2022
Accepted April 22, 2022

Corresponding author
Eunhee Ha
Department of Environmental Medicine,
Ewha Womans University College of Medicine,
260, Gonghang-daero, Gangseo-gu, Seoul
07804, Korea
Tel: 82-2-6986-6234, Fax: 82-2-6986-7022
E-mail: eunheeha@ewha.ac.kr

*These authors contributed equally to this work.

Key Words
COVID-19; Data mining; Sentiment analysis; Korea

서 론

2021년 11월 기준 전세계 코로나바이러스감염증-19 (Coronavirus disease 19, COVID-19, 코로나19) 누적 확진자는 2억 5천만명을 돌파하였다. 해외 유입 국내 COVID-19 감염자 발생은 2020년 1월 20일 처음 발생하였으며, 해외 유입이 아닌 국내 첫 환자 발생은 1월 30일 발생하였다. 이후 기하급수적으로 COVID-19 누적 확진자가 증가하였다.

국민들은 다양한 언론 매체, SNS, 포털 사이트 등을 통해 COVID-19의 발병 상황에 대해 언급하고 있다. 포털 사이트의 지식 응답은 사회 현상에 대해 즉각적인 커뮤니케이션을 요구하며 이러한 자료들은 주로 텍스트 형태의 빅데이터 자료로 존재한다[1].

텍스트 데이터는 구조화된 정형데이터와 달리 많은 양의 데이터를 가지고 있으며, 구조화되지 않은 형태의 자연어로 쓰여진 형태이다. 최근, 이러한 비정형 데이터인 텍스트 자료에 대해 텍스트 마이닝(text mining) 분석하여 여러 사회현상을 설명하려는 시도가 활발하게 이루어지고 있다.

이전의 몇몇 국외 연구들은 레딧(Reddit) 또는 트위터(Twitter)와 같이 소셜 미디어 자료를 기반으로 COVID-19 관련 텍스트 마이닝 분석을 수행한 바 있다[2,3]. 국내 트위터 및 네이버 지식인 자료를 이용하여 COVID-19과 대중의 인식, 불안 확산에 관해 수행한 연구들이 있다[4-6]. 이전 국내 연구들은 최소 1개월-최대 3개월의 비교적 짧은 기간 동안 수행되었다.

본 연구는 2020년 1월부터 2020년 12월까지 네이버 지식인에 등록된 COVID-19 관련 질의에 대하여 텍스트 마이닝 기법을 이용하여 국민들의 COVID-19과 관련된 주요 질문 주제와 핵심어를 도출하고자 하였다.

방 법

1. 자료 수집

본 자료는 2020년 1월 1일부터 2020년 12월 31일 동안 네이버 지식인에 COVID-19 관련 질의에 해당하는 일일 자료를 크롤링하였다. COVID-19 관련 검색 키워드는 다음과 같다: "SARS-Cov-2", "COVID19", "COVID-19", "우한폐렴", "코로나바이러스", "코로나". 네이버에서 제공하는 검색 결과는 최대 1,000건의 검색 결과만 제공하고 있다. 관련 검색 키워드를 네이버 지식인에 검색하면 한 페이지당 최대 10건의 질의를 제공하고 있다. 따라서 1일 최대 100페이지에 해당하는 질문 링크를 얻을 수 있다. 이중 중복되는 URL은 제거하였으며, CSV 파일 형태로 저장하였다.

2. 텍스트 자료 전처리

수집된 자료는 비정형 한글 텍스트 자료이기 때문에 분석 가용한 자료 형태로 변경하기 위해 텍스트 전처리를 수행하였다. (1) 결측 제거, (2) 숫자 제거, (3) 영문 제거, (4) 공백/띄어쓰기 제거, (5) 한글 불용어(stop words) 제거, (6) 구두점 제거 등을 수행하였다. (7) 한글 특성을 고려하여 '확진자가', '확진자랑', '확진자와'처럼 조사가 붙는 단어의 경우 '확진자'와 같이 하나의 명사로 처리하였다. 이후 통계적 기법을 적용할 수 있도록 질의 당 단어들에 대해 문서-단어 행렬(document-term matrix, DTM) 형태로 변환하였다. 단순히 단어의 출현 빈도가 아니라 특정 문헌에서 특정 단어의 출현 빈도의 비중을 고려한 term frequency-inverse document frequency (TF-IDF)를 가중치로 사용하였다. TF-IDF는 전체 문서들에서 빈번하게 출현하는 단어는 특정 문서의 구분에 잘 활용되지 않으므로 중요도를 낮게, 특정 문서에서 빈번하게 출현하는 단어의 중요도를 높게 고려하는 것으로 문서 빈도의 역수인 역문서 빈도와 단어 출현 빈도를 고려하여 계산한다[7].

3. 월별 단어 추이 분석

국민들의 COVID-19 관련 주제에 대한 추이를 살펴보기 위하여 단어별 빈도를 집계하였다. 전체 문서에서 출현 빈도가 높은 상위 20개 단어들을 추출하여, 상위 단어들의 월별 추이를 살펴보았다.

4. 토픽 모델링

본 연구에서는 토픽 모델링(topic modeling)을 수행하기 위해 주로 이용하는 비지도 학습 기법인 latent dirichlet allocation (LDA)를 이용하였다[8]. 토픽 모델링은 문서나 문헌 속의 텍스트에서 확률적으로 주제들을 추출하는 텍스트 마이닝 기법의 하나이다. LDA는 연속 확률 분포이며 다항분포인 디리클레 분포(Dirichlet distribution)를 허용하고 각각의 단어를 특정 주제에 할당하여 문장 내에 잠재된 의미(토픽)을 발견하는 방법이다[8].

LDA는 토픽별 단어의 분포와 문서별 토픽의 분포를 모두 추정하며 생성 과정은 다음과 같다.

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

여기서 K는 토픽 수, D는 문서 수, n은 d번째 문서의 총 단어 개수를 의미한다. 토픽 분포 $Z_{d,n}$은 문서당 토픽의 분포 θ_d에 의존한다. $\beta_{1:K}$는 각 토픽을 의미하며 β_K는 주제별 어휘 분포

를 의미한다. d번째 문서의 n번째 단어인 단어 W_d는 유일하게 관찰 가능한 변수이며, 나머지 변수는 추정해서 얻어낼 수 있다.

토픽의 수는 토픽 분류의 정확성과 타당성에 영향을 미치기 때문에 토픽 결정 수에 따라 토픽 주제는 조금씩 달라질 수 있다. 따라서, 적절한 토픽의 수를 정하기 위해 모형의 복잡도(perplexity)와 로그 가능도(log-likelihood)를 계산하였다(Fig. 1)[9]. 전체 자료를 3:1로 훈련 자료(train set)와 검증 자료(validation set)로 분할하여 평가하였다. 복잡도를 평가하기 위해 토픽의 수는 2부터 100까지 그리드 서치(Grid search)하였다. 토픽의 수는 일반적으로 복잡도의 감소 폭이 완만해지는 지점으로 결정한다. 본 연구에서는 너무 많은 토픽의 수를 결정하는 것은 대중의 실제 관심사/이슈가 무엇인지 파악하기 어렵다고 판단하였고, 너무 적은 토픽의 수는 실제 관심사를 적게 반영한다고 고려하였다. 따라서 적정 토픽의 수를 10-30개 사이로 정하기로 고려하였고 최종적으로 토픽의 수를 20으로 결정하였다(Fig. 1).

각 추출된 토픽에 대해 연구진 협의를 거쳐 적정 토픽 레이블을 명명하였다(Table 1). 각 정의한 토픽 주제 20개의 주별 등장 확률을 계산하여 국내 유행 시기별 추이를 살펴보았다.

5. 감성 분석

연구 기간 동안 COVID-19 관련 주제에 대한 국민들의 인식을 살펴보기 위해 감성 분석(sentiment analysis)을 수행하였다. 감성 분석이란 감성 어휘(sentiment terms)와 어휘의 긍정 및 부정의 정도를 나타내는 극성(polarity)으로 구성된 감성 사전(sentiment dictionary or sentiment lexicon)을 활용하여 감성을 정량화 하는 것을 의미한다[10,11]. 대표적인 한국어 감성 사전으로는 서울대학교에서 개발한 한국어 감성분석 코퍼스(KOSAC)과 군산대학교에서 구축한 KNU 감성 사전이 있다. KNU 감성 사전에서 정의되는 감정은 인간의 보편적인 감정 표현을 나타내는 표현을 위주로 사용하고 있어 본 연구에서 활용하였다. KNU 감성 사전은 다양한 분야에서 사용될 수 있는 범용 감성 사전으로 국립국어원 표준국어대사전의 뜻풀이 분석을 통한 긍/부정 어휘를 추출하여 총 14,843개의 어휘(관용구, 문형, 축약어, 이모티콘)에 대한 극성 값이 계산되어 있다[12,13] .

각 일별 수집 문서 중 명사 단어와 KNU감성사전에서 정의된 부정어/긍정어를 강한 긍정(+2), 약한 긍정(+1), 중간(+0), 약한 부정(-1), 강한 부정(-2)의 점수를 연계하여 총 점수를 합산하였다. 문서에 대하여 합산된 점수가 양수이면 긍정(positive) 문서, 음수이면 부정(negative) 문서로 간주하여 일일 긍정 질의와 부정 질의를 분류 하였다. 수집된 일별 전체 문서를 분모로 하고, 부정 문서를 분자로 하여 일일 부정 정서 비율(proportion of negative sentiment)을 계산하였다.

본 연구에서는 산출한 일일 부정 정서 비율과 일일 확진자 수 간의 상관성을 파악하고자 하였다. 한국의 코로나19 확진자

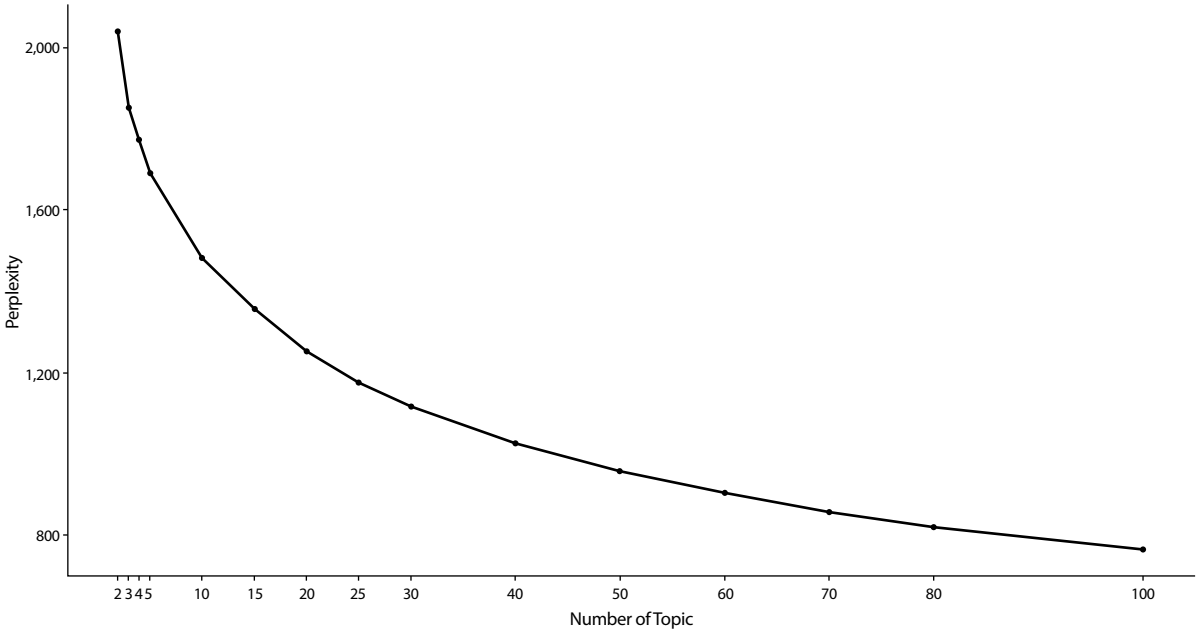

Fig. 1. The Perplexity of topic modeling related to COVID-19. The x-axis indicates the number of topic. The y-axis indicates perplexity of latent Dirichlet allocation (LDA) models.

Table 1. Topics related to COVID-19

OBS	Topic	Top keyword
1	COVID-test	'test', 'confirmed cases', 'sequester', 'self', 'contact', 'public health center', 'visit', 'result', 'negative', 'call'
2	Economy	'loans', 'contract', 'personal', 'progress', 'bank', 'situation', 'operation', 'monthly rent', 'card'
3	School	'school', 'academy', 'student', 'class', 'start of school', 'online', 'postponement', 'teacher', 'attending school', 'lecture'
4	Hospital/Diagnose	'hospital', 'pain', 'treatment', 'left', 'right', 'prescription', 'leg', 'doctor', 'clinic', 'surgery'
5	Travel/Oversea	'travel', 'situation', 'cancel', 'entry', 'oversea', 'refund', 'schedule', 'reservation', 'visa', 'period'
6	Life/Health	'exercise', 'method', 'start', 'day', 'life', 'support', 'health', 'diet', 'dog', 'health'
7	Social issue	'reason', 'thinking', 'government', 'Korea', 'church', 'nation', 'situation', 'country', 'economy', 'a new world'
8	Symptom 1 (respiratory)	'cough', 'symptom', 'runny nose', 'cold', 'sputum', 'fever', 'sneezing', 'rhinitis', 'dry cough', 'tickling'
9	Relationship	'friend', 'contact', 'idea', 'story', 'female', 'male', 'heart', 'worry', 'game', 'present'
10	Symptom 2 (headache, fever)	'head', 'symptom', 'headache', 'muscle pain', 'slight fever', 'body temperature', 'body ache', 'normal', 'suspicion', 'vomiting'
11	Workplace	'company', 'salary', 'work', 'unemployment', 'monthly salary', 'staff', 'resignation', 'job', 'commute', 'period'
12	Mask/social distance	'mask', 'stage', 'distance', 'food', 'cafe', 'bus', 'song', 'wearing', 'disinfection', 'utilization'
13	Infection/Vaccine	'person', 'infection', 'likelihood', 'pneumonia', 'dangerous', 'vaccine', 'Wuhan', 'probability', 'worry', 'Chinese'
14	Stimulus package	'application', 'subsidy' 'support', 'document', 'report', 'income', 'confirm', 'disaster', 'emergency', 'standard'
15	Family	'mother', 'thinking', 'father', 'parents', 'family', 'brother', 'stress', 'heart', 'grandmother', 'person'
16	Delivery service	'use', 'parcel service', 'delivery', 'purchase', 'call', 'internet', 'arrival', 'price', 'number', 'thing'
17	Unclassified	'problem', 'photo', 'partly', 'face', 'skin', 'situation', 'first', 'start', 'washroom', 'total'
18	Region	'region', 'Daegu', 'Busan', 'Gyeonggi', 'week', 'schedule', 'near', 'vacation', 'this year', 'Seoul'
19	Study/Exam	'study', 'examination', 'ready', 'school year', 'English', 'university', 'math', 'term', 'high school', 'middle term'
20	Worry/Anxiety	'worry', 'anxiety', 'stuffy', 'chest', 'symptom', 'usual', 'situation', 'uncomfortable', 'day', 'menstruation'

수 추이는 해외 유입, 신천지 관련, 집단 발병, 확진자 접촉, 개별 사례 등의 이유의 원인으로 시기별로 구분할 수 있다. 다만 코로나19 확진자 수의 증감 시기를 연구자 임의로 구분하기 어렵기 때문에 중앙대책본부 역학조사분석단에서 발표한 사회적 거리두기 지침에 따라 시기를 구분하였다. 정의한 코로나19 발생 양상의 기간(1–5기)에 따라 일일확진자수와 부정 정서 비율에 대해 피어슨 상관분석하였다[14].

모든 분석은 R 통계 소프트웨어(version 3.5.1; R Foundation for Statistical Computing, Vienna, Austria)를 이용하여 수행하였다. 텍스트 마이닝을 분석을 위해 "*KoNLP*" "*tidyr*", "*tidytext*", "*topicmodels*" 등 R packages을 이용하였다.

결 과

1. COVID-19 관련 검색 키워드 빈도

총 252,181개의 문서 중 241,776개의 문서와 614,682개 단어를 이용하였다. 전체 개별 단어 중에서 가장 빈도가 높은 상위 20개 검색 키워드는 Fig. 2, 월별 추이 결과는 Fig. 3에 제시하였다. 가장 빈도가 높은 단어는 기침이었으며, 증상, 검사, 확진자, 마스크 순이었다(Fig. 2). 20개 키워드 중 12개 키워드('가래', '감기', '감염', '걱정', '기침', '두통', '마스크', '머리', '병원', '사람', '증상', '콧물')는 2월에 급증 이후 감소하는 패턴을 보였다. '신청' 키워드는 3–4월에 증가 추세 이후 감소하는 패턴을 나타냈다(Fig. 3). 5개 키워드('검사', '격리', '친구', '학교', '학원')는 연간 증가 추세를 보였다. '단계' 키워드는 7월 이후 12월까지 급증하였다. '확진자'는 2월 이후 증감을 반복하는 패턴을 보였다.

2. 20개 COVID-19 관련 주제

토픽 모델링으로 결정된 COVID-19 관련 질의에 대한 주제 20개는 Table 1에 제시되어 있다. 토픽 주제는 Topic 1: 검사, Topic 2는 자영업/경제, Topic 3: 학교/등교, Topic 4: 병원/진료, Topic 5: 여행/해외, Topic 6: 일상/운동, Topic 7: 사회현상, Topic 8: 증상 1(호흡기), Topic 9: 대인관계, Topic 10: 증

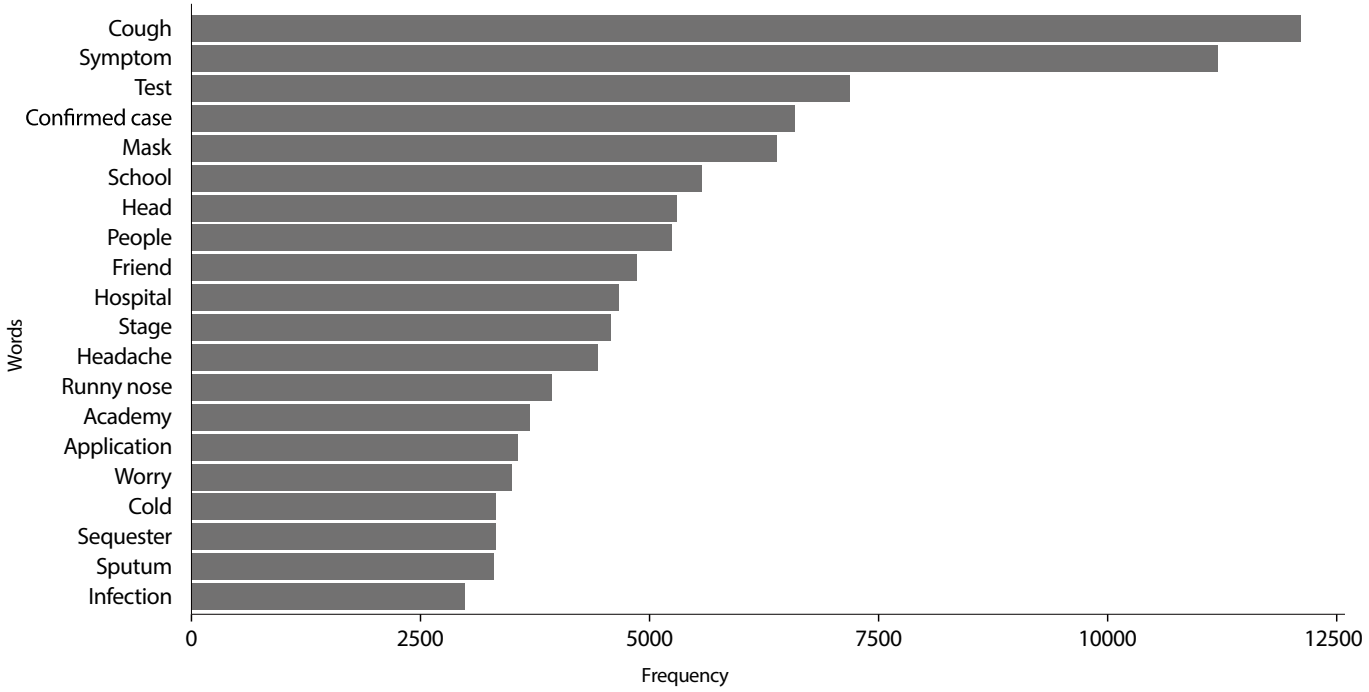

Fig. 2. Top 20 frequent words related to COVID-19 documents.

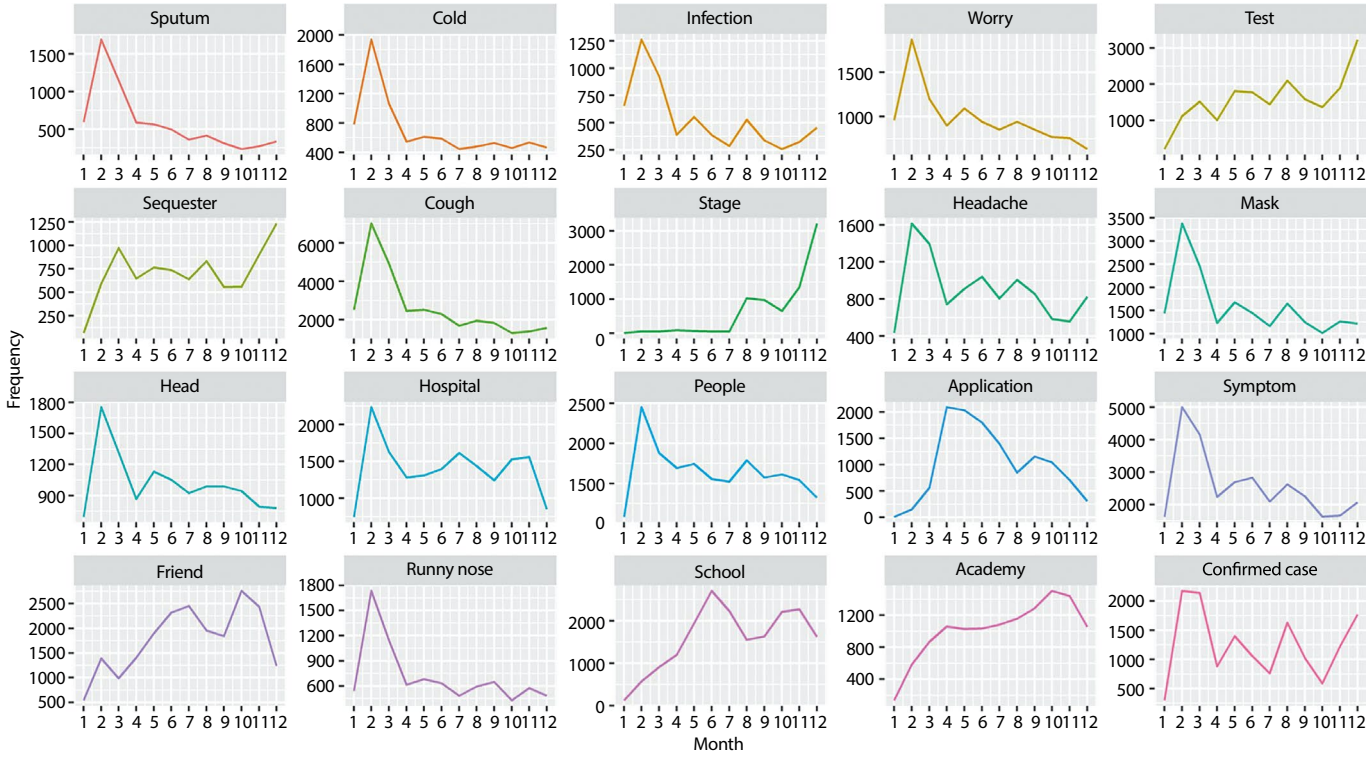

Fig. 3. Top 20 frequent words' monthly trend from January 1, 2020 to December 31, 2020. The x-axis indicates months. The y-axis indicates the frequency of words.

상 2(두통발열), Topic 11: 직장/회사, Topic 12: 마스크/단계, Topic 13: 감염/백신, Topic 14: 재난 지원금, Topic 15: 가족, Topic 16: 택배/구매, Topic 17: 미정, Topic 18: 지역, Topic 19: 공부/시험, Topic 20: 걱정/불안으로 정하였다. Fig. 4는 유행 시기별 토픽 주제의 일별 추이 변화를 보여준다. Topic 1: 검사, Topic 8: 증상 1(호흡기), Topic 12: 마스크/단계, Topic 14: 재난 지원금은 상반기에 증가 추세를 보이다 하반기에 점차 감소하였다. Topic 2: 자영업/경제, Topic 3: 학교/등교, 병원/진료 등은 코로나19 초기에 비해 증가 추세를 나타냈다.

3. 감성 분석 수행 결과

2020년 1월 20일부터 2020년 12월 31일까지 사회적 거리두기 정책변화와 코로나19 일일확진자수 발생 양상에 따른 표는 Table 2 [14]에 제시하였다. 연구 기간 동안 일일 확진자 수와 부정 정서 비율은 Fig. 5에서 볼 수 있다. 부정 정서 비율의 경우 2020년 2월 23일 0.920으로 연구 기간 동안 가장 높은 부정 정서 비율을 보였다. 연구 기간 전체 동안 일일 확진자 수와 부정 정서 비율의 상관계수는 0.170(P=0.001)으로 양의 상관성을 보였다. 제2기(2020.2.8–5.5)의 경우 r=0.659, 제4기(8.12–11.12)의 경우, r=0.739로 가장 높게 나타났으며, 제5기(11.13–12.31)의 경우 r=0.516이었다. 이에 비해 제1기(1.20–2.17)의 경우 r=0.281로 낮은 양의 상관관계를 보였다. 또, 제3기(5.6–8.11)는 가장 낮은 관련성을 보였다(r=–0.164).

고 찰

본 연구는 대규모의 비정형자료에서 토픽모델링을 통해 주요 키워드를 추출하고, 시기별/유행양상별 변화를 파악하며, 감성 분석을 통해 부정 정서 비율과의 연관성을 살펴봄으로써 COVID-19에 대한 대중의 관심 주제와 정서를 도출해낼 수 있음에 그 의의가 있다.

코로나19 발생 직후인 1–2월에는 코로나19의 일반적 증상에 대한 대중들의 궁금증이 높게 나타났음을 알 수 있다. 또, 이 시기는 공적 마스크 제도가 2020년 3월 9일 시행되기 전으로, 당시의 마스크 품귀 대란에 대한 대중의 관심과 초기 코로나19 감염 확산에 대한 대중들의 우려가 반영되어 '마스크'와 '감염' 키워드가 피크를 나타낸 것으로 해석될 수 있다. '단계'의 경우 질병관리청에 의해 위기 단계 조정이 이루어진 8월

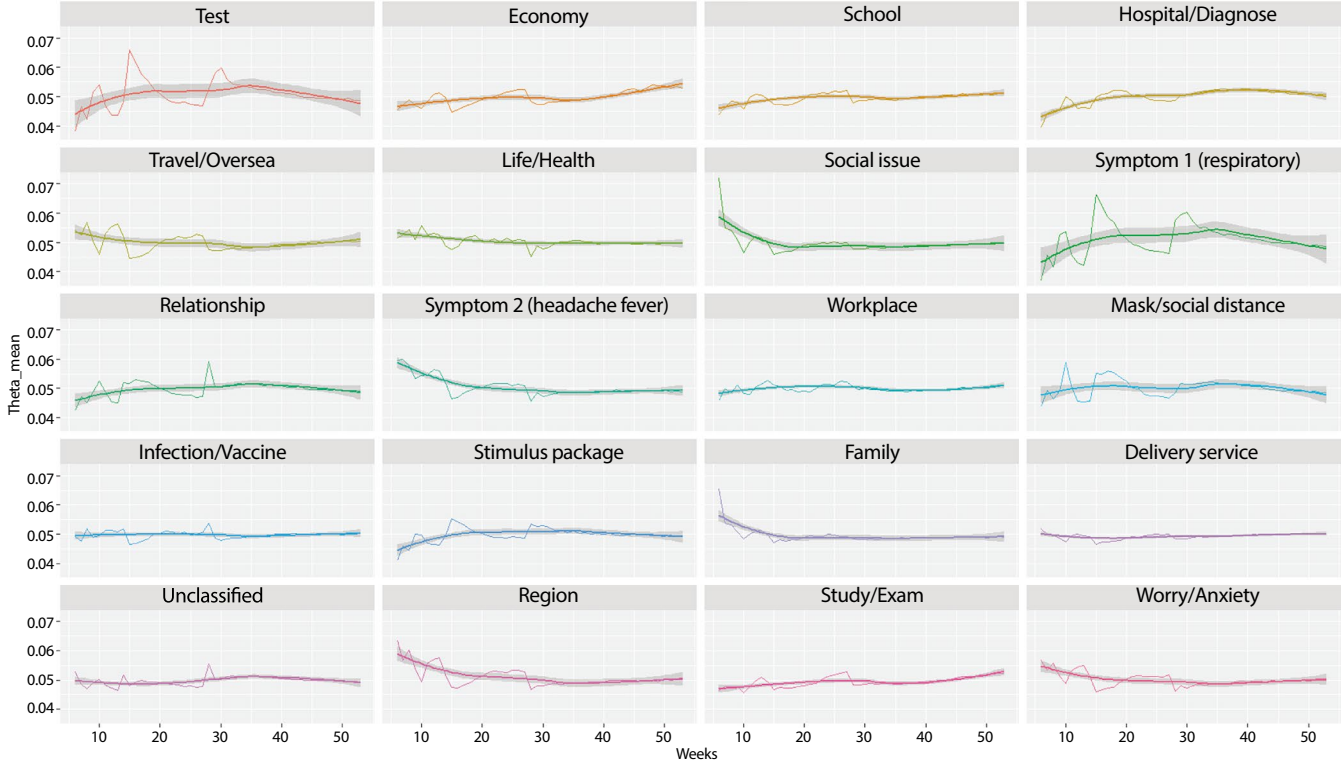

Fig. 4. The time-plot of 20 topics related to COVID-19 documents from January 1, 2020 to December 31, 2020. The x-axis indicates weeks. The y-axis indicates the probability that each topics appear.

Table 2. COVID-19 outbreak period

Categories	Stage 1	Stage 2 (1st Outbreaks)	Stage 3	Stage 4 (2nd Outbreaks)	Stage 5 (3rd Outbreaks)
Date	(2020.1.20–2.17)	(2.18–5.5)	(5.6–8.11)	(8.12–11.12)	(11.13–12.31)
Classification	Imported cases	Large-scale clusters	Local clusters, sporadic cases	Small and medium-scale clusters many outbreaks	A large outbreak nationwide
Confirmed cases	n=30	n=10,774	n=3,856	n=13,282	n=45,173

Adapted from Korea Disease Control and Prevention Agency [14] with CC-BY.

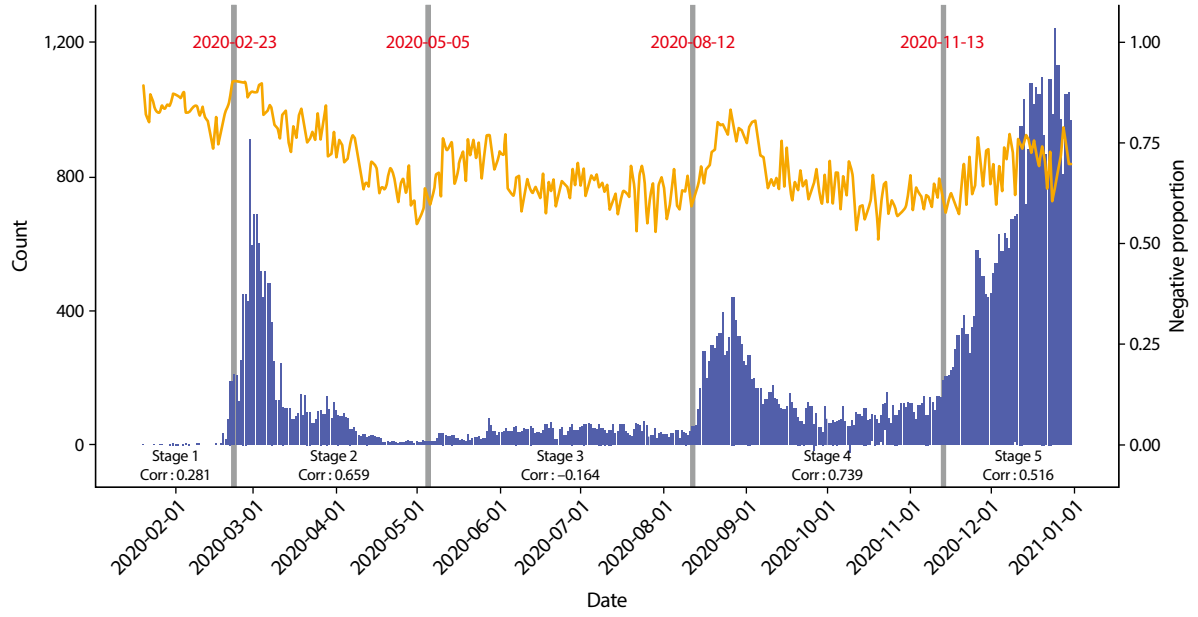

Fig. 5. The correlation of between daily COVID-19 cases and proportion of negative sentiment of documents from January 1 2020 to December 31 2020. The x-axis indicates date. The y-axis indicates the daily number of confirmed COVID-19 case (counts) in Korea. The auxiliary axis indicates the proportion of negative sentiment of documents. The blue bar indicates the daily number of confirmed COVID-19 cases. The orange line indicates the proportion of negative sentiment. Five stages are defined by Table 2. COVID-19 outbreak period (Stage 1: 2020.1.20–2.17, Stage 2: 2.18–5.5, Stage 3: 5.6–8.11, Stage 4: 8.12–11.12, Stage 5: 11.13–12.31).

과 12월에 특히 관심이 고조되었으며, 점점 정책의 변화에 대해 관심도가 높아진 것에 비해, '확진자'의 경우 2월의 대구, 경북 지역 중심의 1차 대유행, 5월의 소규모 집단발생, 8월의 2차 대유행시작 등 실제 확진자 수의 규모에 따라 영향을 받았음을 알 수 있다. '신청'의 경우 5월 긴급 재난 지원금 신청 및 지급 절차에 대한 정부 논의가 구체화되기 시작하여 5–6월에 대중의 관심이 높았음을 유추할 수 있다.

본 연구에서 토픽모델링을 통해 도출된 20개의 대중들의 코로나-19 관련 주제와 발견된 토픽의 1년간 추이를 파악할 수 있었다. 토픽들의 추이 변화는 코로나19 유행 양상, 정부의 지침, 개인의 사회적·경제적 지위, 국내외 코로나19 확진자 수 등에 따라 달라질 수 있으며, 향후 연구에서는 이러한 요인들을 고려할 필요가 있다.

코로나19 기간 동안 대중들의 정서와 일별 확진자 수간 관련성을 파악하기 위해 수행한 감성 분석을 토대로, 부정 정서 비율이 유행 양상과 상관성이 있음을 보인다. 2월 23일(0.920)은 지역사회가 감염되면서 정부에서 감염병 위기 단계를 '경계'에서 '심각' 단계로 격상한 시기와 일치하며 사회적거리두기 실시가 언급되기 직전으로 볼 수 있다. 5월 11일(0.761)은 5월 6일 생활 속 거리두기로 전환된 시점이지만, 5월초 수도권 유흥시설에서 시작된 집단감염 사례가 인근지역으로 확산된 시기와 밀접한 연관이 있다고 볼 수 있다. 8월 21일(0.800)과 8월 26일(0.830)의 경우, 수도권 일부교회 등 종교시설과 대규모 도심 집회 등 집단발생이 다수 일어난 시기와 일치한다. 12월 23일

(0.726)과 12월 28일(0.788)의 경우 코로나19 대유행 이후 최초 1,000명대 진입(12월 13일 이후)으로 2020년 최대 규모의 4차 대유행을 기록하며 고강도 사회적 거리두기가 이루어지던 시기와 연관성이 높다고 볼 수 있다.

부정 정서 비율을 코로나19 유행 양상에 따라 분류한 제5기와 상관분석을 진행한 결과, 제2, 4, 5기에서 부정 정서 비율과 일일 확진자수와의 상관관계가 높게 나타났다. 제2기(2020.2.8–5.5)의 경우 신천지 대규모 유행을 시작으로 의료기관, 종교기관, 다중시설의 집단발생이 전국적으로 발생, 확산되었으며 이로 인해 r=0.659의 높은 상관관계를 나타냄을 유추할 수 있다. 가장 높은 값(r=0.739)을 기록한 제4기(8.12–11.12)의 경우 수도권 종교시설, 대규모 집회 등 집단 발생과 더불어 위중증 환자도 다수 발생한 시기로 이와 관련성이 높다고 생각된다. 제5기(11.13–12.31)의 경우 수도권에서 전국적으로 확산되는 4차대유행과 고강도 거리두기로 인해 r=0.516)로 나타남을 알 수 있다. 이에 비해 제1기(1.20–2.17)의 경우, 중국 등 해외유입위주의 개별적 산발사례 발생 시기로, 감염병 위기단계가 '주의', '경계' 등의 초기 단계로 r=0.281(p=0.140)로 낮은 양의 상관관계를 보였음을 알 수 있다. 또, 제3기(5.6–8.11)는 5월 6일 '생활 속 거리두기' 전환 이후 소규모 집단에서 산발적으로 발생하였으며, 주로 200명 미만의 낮은 확진자 수를 기록하였기 때문에 이 시기의 상관계수가 r=–0.164로 가장 낮게 나타남을 유추할 수 있다.

본 연구의 강점은 다음과 같다. 첫째, 본 연구는 네이버 지식인에 등록된 COVID-19 관련 질의에 대하여 텍스트 마이닝 기법을 이용하여 국민들의 COVID-19과 관련된 주요 질문 주제와 핵심어를 도출하였다. 둘째, 본 연구는 토픽 모델링과 감성 분석을 수행함으로써 중심키워드와 관심의 경향, 유행 시기별 대중의 감정 변화를 유추하였다. 국내의 경우 감성 분석을 진행한 연구가 드물기 때문에 본 연구는 1년 동안 확진자 수, 유행 양상에 따라 국민들의 감성 수준을 확인하였다는 점에서 의의를 가진다.

셋째, 본 연구는 이전 코로나19 관련 초기 국내 연구들과 달리 상대적으로 연구 기간이 길다. 이전의 SNS, 뉴스 기사 등을 활용한 연구들은 연구 기간이 1–3달 내외로 수행되었다. 본 연구는 2020년 한 해 동안 COVID-19 유행양상시기와 코로나19 확진자 수 증감에 따른 질의 변화 추이를 파악할 수 있었다.

본 연구에서는 존재하는 한계점과 향후 연구 방향은 다음과 같다. 첫째, 우리는 질의를 올린 사용자의 인구통계학적 특징(성별, 연령 등)과 사회·경제학적 특징(지역, 소득수준, 직업 등)을 고려하지 못하였다. 또한 한 사람이 여러 질문을 할 수 있기 때문에 고유 대상자는 문서 수보다 적을 수 있다. 하지만 우리는 고유 대상자가 몇 명인지는 수집된 자료에서 알 수 없었다. 향후 사용자에 대해 이러한 특징을 파악할 수 있다면, 연령이나 성별, 지역에 따른 더 구체적인 분석 결과를 제시할 수 있을 것이다.

둘째, 우리는 연구 기간 동안 네이버 지식인에 COVID-19 관련 질의를 모두 모은 것이 아니라, 전체 자료 중 일부만 크롤링하였다. 네이버는 한국에서 가장 자주 사용하는 포털 사이트 중 하나임에도 불구하고 모든 사람들이 온라인 공간에 개인의 정보와 의견을 공유하지 않아 전체 인구 집단을 대표하기에 적절하지 않으며 연구 결과를 일반화하기 어렵다[4]. 연구 기간 동안 질의 전체를 추출한 것이 아니기 때문에 우리 연구에 샘플로서 포함되지 않은 질의들이 많다면, 우리 연구에서 산출한 토픽의 주제와 비율이 다를 수 있다. 또한 백신이나 재난 지원금 등 세부적인 검색 키워드로 주제들을 다루지 않아 도출된 단어나 토픽들로 대중의 관심과 이슈를 구체화하지 못하였다. 셋째, 한글은 영어와 달리 자연어에 영향을 많이 받으며 전처리 과정이 까다롭다. 향후 연구는 한글에 대한 정제된 자연어 처리 과정이 보완될 필요가 있다. 또한 문서 내의 단어의 출현 순서를 고려하지 않고 출현 빈도를 고려한 방법론을 고려했다[8,15]. 향후 연구에는 문서 내 단어 순서의 의미를 고려한 분석 방법론을 적용할 필요가 있다.

본 연구는 2020년 한 해 동안 네이버 지식인에 COVID-19 관련 검색어 관한 질의에 대한 토픽 모델을 수행하여 대중들의 관심사를 파악하고자 하였다. 이 연구는 대규모 한글 텍스트 자료를 처리하여 코로나19 기간 동안 대중들의 코로나19 관련 질의 주제, 주제 동향, 정서를 파악한 점에서 의의가 있다. 추후 연구에서는 코로나19 변이 시기에 따른 변화 양상과 다양한 매체와 자료원(뉴스, 역학조사, 설문자료, 인터뷰 자료 등)을 이용하여 분석 및 결과를 비교할 필요가 있다.

References

1. Lee SM, Ryu SE, Ahn S. Mass media and social media agenda analysis using text mining: focused on '5-day rotation mask distribution system.' *J Korea Content Assoc* 2020;20:460-469.
2. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on twitter: sentiment analysis and topic modeling study. *JMIR Public Health Surveil* 2020;6:e21978 .
3. Naseem SS, Kumar D, Parsa MS, Golab L. Text mining of COVID-19 discussions on reddit. In: He J, Purohit H, Huang G, Gao X, Deng K, editors. Proceedings of the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT); 2020 Dec 14-17; Melbourne. Piscataway (NJ): IEEE; 2020. p. 687-691.
4. Jo W, Lee J, Park J, Kim Y. Online information exchange and anxiety spread in the early stage of the novel coronavirus (COVID-19) out-

break in South Korea: structural topic model and network analysis. *J Med Internet Res* 2020;22:e19455.
5. Shim JG, Ryu KH, Lee SH, Cho EA, Lee YJ, Ahn JH. Text mining approaches to analyze public sentiment changes regarding COVID-19 vaccines on social media in Korea. *Int J Environ Res Public Health* 2021;18:6549 .
6. Jo W, Chang D. Political consequences of COVID-19 and media framing in South Korea. *Front Public Health* 2020;8:425.
7. Ramos J. Using TF-IDF to determine word relevance in document queries [Internet]. State College (PA): Citeseer; 2013 [cited 2022 Jan 10]. Available from: https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-in-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c.
8. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003;3:993-1022.
9. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, et al. Applying LDA topic modeling in communication research: toward a valid and reliable methodology. *Commun Methods Meas* 2018;12:93-118.
10. Dubey AD. Twitter sentiment analysis during COVID-19 Outbreak [Internet]. Amsterdam: Social Science Research Network; 2020 [cited 2022 Jan 5]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3572023.
11. Li N, Wu DD. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decis Support Syst* 2010;48:354-368.
12. Suh H, So J. A study on the topic and sentiment of national petition data using text analysis. *Korean Data Anal Soc* 2020;22:999-1011.
13. Park SM, Na CW, Choi MS, Lee DH, On BW. KNU Korean sentiment lexicon: Bi-LSTM-based method for building a Korean sentiment lexicon. *J Intell Inf Syst* 2018;24:219-240.
14. Kim Y, Kim YY, Yeom H, Jang J, Hwang I, Park K, et al. COVID-19 1-year outbreak report as of January 19, 2021 in the Republic of Korea. *Public Health Wkly Rep* 2021;14:478-481.
15. Tang J, Meng Z, Nguyen X, Mei Q, Zhang M. Understanding the limiting factors of topic modeling via posterior contraction analysis. In: Xing EP, Jebara T, editors. Proceedings of the 31st International Conference on Machine Learning; 2014 Jun 21-26; Stroudsburg (PA) : International Machine Learning Society; 2014. p. 190-198.