경영전문석사학위 논문

토픽 모델링을 활용한 코로나19 팬데믹에 대한 한국인들의 인식 분석

2022년 8월

서울과학종합대학원대학교

김 완 수

경영전문석사학위 논문

토픽 모델링을 활용한 코로나19 팬데믹에 대한 한국인들의 인식 분석

2022년 8월

서울과학종합대학원대학교

김 완 수

토픽 모델링을 활용한 코로나19 팬데믹에 대한 한국인들의 인식 분석

지도교수 박 정 열

이 논문을 경영학 석사 학위논문으로 제출함
2022 년 8 월
서울과학종합대학원대학교
김 완 수

김완수의 석사 학위논문을 인준함
2022 년 7 월

위 원 장 김보영 (인)

위 원 임효숙 (인)

위 원 박정열 (인)

초 록

2019년에 시작된 코로나19 팬데믹은 2021년 5월 기준 전 세계에서 1억5천800만여 명의 발병자 및 329만여 명의 사망자가 발생하였다. 한국인에서 코로나19 팬데믹에 대한 인식을 분석하기 위해 소셜 빅데이터로 LDA 토픽 모델링과 감성 분석을 수행하였다.

2019년 2월부터 2021년 12월까지 코로나19 팬데믹 관련 네이버 뉴스 116개를 수집하고, 각 기사에 달린 댓글 총 4837개를 분석하였다. 뉴스 기사 수집에 이용한 검색어는 “속보”, “코로나19”, “신규확진 ” 이었다. 세 개 검색어가 모두 포함된 기사만을 수집하였다. 전날 확진자가 다음날 반영되는 것을 고려하여 검사 수가 많은 월요일 확진자 수가 반영된 화요일 뉴스 기사를 수집하였다. 코로나19 신규확진 속보 뉴스 기사에 달린 댓글에는 ‘코로나’, ‘확진’, ‘백신’, ‘국민’, ‘방역’ 등의 명사가 순서대로 상위를 차지하고 있었다.

토픽 모델링 결과, “해외”, “정부”, “검사” 세 개의 토픽이 도출되었으며, 각 토픽별 주요 키워드를 기반으로 단어 네트워크를 생성하였다. “해외” 토픽에서 가장 중요한 키워드는 ‘해외’ 였고, “정부” 토픽에서는 ‘백신, “검사” 토픽에서는 ’ 확진 ‘이 가장 중요한 키워드였다. 각 토픽별로 주요 키워드와 댓글을 구체적으로 살펴보면 부정적인 단어가 많았다.

감성 분석은 군산대학교 감성사전을 이용하였다. 그 결과, 댓글의 부정 비율이 긍정 비율보다 약 2.6배 많았다. 중립으로 분류된 댓글을 제외하면, 긍정 비율은 27.8%, 부정 비율은 72.2%였다. 코로나19 팬데믹이 2년 이상 지속되면서 한국의 방역과 의료 체계에 대한 불만 여론이 본 연구를 통해서도 드러났다.

본 연구의 한계로는 뉴스 기사와 댓글의 수가 적은 편이었고, 감성 분석 방법이 반정량적이었다는 점이 있다. 더 포괄적으로 뉴스 기사를 수집하고, 댓글 수가 더 많았다면 시기별 여론 경향 변화 분석도 가능했을 것이다. 본 연구에서 사용한 감성 분석 방법은 문장 전체에 점수를 매겨서 양수면 긍정, 음수면 부정으로 이분화하였으나, 소수점까지 수치화할 수 있는 감성 분석 방법을 사용하였다면 더 구체적

으로 여론 인식을 알 수 있었을 것이다.

본 연구는 코로나19 팬데믹과 같은 위기상황에 온라인 뉴스 기사 댓글 빅데이터 분석을 통해 여론 경향을 알 수 있음을 보여주었다. 이러한 방법론은 점차 보편화, 자동화되고 있으며 산업계에서 비즈니스와 마케팅에 적극 활용할 수 있겠다.

목 차

표 목 차

그 림 목 차

제 Ⅰ 장 서론

제 1 절 연구 배경

코로나바이러스 감염증-19(코로나19)는 중국 우한에서 2019년 12월 첫 감염자가 발견되었으며, 한국에서는 2020년 1월 20일 첫 확진자가 발생하였다. 이후 2022년 6월 30일 기준 누적 확진 환자는 18,359,309명에 이른다.[1] 전 세계적으로 코로나19가 빠르게 확산하자, 2020년 3월 11일 역사상 세 번째로 세계보건기구(World Health Organization, WHO)는 가장 높은 전염병 경보 등급에 해당하는 '팬데믹(Pandemic)'을 선언하였다.

세계보건기구에 따르면, 전 세계적으로 주간 확진자 수는 2022년 3월 마지막 정점 이후 감소세를 보인 후 2022년 6월에는 3주 연속 증가세를 보였다. 2022년 6월 20일에서 26일 사이에 410만 명이 넘는 새로운 확진자가 보고되었으며, 이는 이전 주와 비교하면 18% 증가한 수치였다. 새로운 주간 사망자 수는 8500명 이상의 사망자가 보고된 이전 주와 유사하였다.[2] 동남아시아 지역은 6월 한 달 동안 131,000건 이상의 새로운 확진자가 보고되어 이전 주에 비해 새로운 확진자 수가 32% 증가하였다.

코로나19는 중증급성호흡기증후군 코로나바이러스 2 (severe acute respiratory syndrome coronavirus-2, SARS-CoV-2)에 의해 유발되는 전염성 질병이다. 코로나19의 증상은 다양하지만 열, 기침, 두통, 피로, 호흡곤란, 후각 상실 및 미각 상실을 포함하는 경우가 많다.[3] 같이 감염되었다해도 환자마다 다른 증상을 보일 수 있고, 시간이 지남에 따라 증상이 변할 수 있다. 세 가지 일반적인 증상 군집이 확인되었다: 기침, 가래, 호흡곤란 및 발열을 동반한 호흡기 증상 군집; 근육 및 관절 통증, 두통 및 피로를 동반한 근골격계 증상 군집; 복통, 구토, 설사를 동반한

소화기 증상 군집으로 크게 나뉜다. 이전에 이비인후과적 장애가 없는 환자들의 경우 후각 상실과 함께 미각 상실이 코로나19 감염과 관련이 있으며 88%의 환자에서 후각 및 미각 상실이 보고되었다.[4] 코로나19의 전염 경로는 바이러스를 포함하는 비말로 오염된 공기를 흡입하여 감염된다.[5] 팬데믹을 극복하기 위하여 여러 코로나19 백신이 승인되어 다수의 국가에서 코로나19 백신 접종을 시행하였다. 현재까지의 연구에 따르면 코로나19 백신은 코로나19로 인한 중증도와 사망률을 줄이는 데 기여하였다.[6]

한국에서도 백신 접종률이 80%를 넘기고, 항체 보유자가 95%를 넘겼으나[7] 변종 바이러스의 등장으로 팬데믹이 종식될 기미는 좀처럼 보이지 않는다. 이러한 코로나19의 장기적인 대유행은 경제, 정치, 사회, 문화 등 모든 영역에서 사람들의 삶에 영향을 끼쳤다. 팬데믹으로 인한 경제적, 사회적 혼란은 파괴적이었다. 세계보건기구를 비롯한 ILO (International Labour Organization), FAO (Food and Agriculture Organization of the United Nations), IFAD (International Fund for Agricultural Development)의 공동 성명에 따르면 수천만 명의 사람들이 극심한 빈곤에 빠질 위험에 처했으며, 2021년 기준 거의 6억9천만 명으로 추산되는 영양실조 인구가 2022년에는 최대 1억3,200만 명까지 증가할 것으로 예상된다.[8]

현재 국내 뉴스 이용자의 대부분이 네이버(www.naver.com)나 다음(www.daum.net)과 같은 인터넷 포털 사이트(portal website)를 통해 뉴스를 접한다. 이에 따라 온라인 뉴스가 사람들의 삶에 끼치는 파급효과가 증가하고 있다. 온라인 뉴스에 대한 접근성이 쉬워지고 댓글 서비스가 등장하면서, 온라인상에서 댓글이 여론 형성에 미치는 영향도 커지고 있다. 댓글 서비스 시행 전의 뉴스는 구독자들이 수동적으로 수용했다면, 댓글 서비스 시행 후에는 뉴스에 대한 비판과 의견 개진까지 가능하게 되었다.[9] 게다가 개인의 태도에는 기사의 논조보다 댓글이 더 큰 영향력을 미치는 것으로 확인되었다.[10] 양혜승(2008)에 의하면 이제 댓글은 “특

정 이슈를 바라보고 해석하는 사회 구성원들의 생생한 목소리가 직접 기록되는 공간”이자 “여론의 동향을 손쉽게 감지할 수 있는 공간”으로 기능한다. 그러므로 코로나19 관련 기사에 달린 댓글은 코로나19를 바라보는 일반인의 인식과 태도를 파악하는 데 중요한 자료이다.

소셜 미디어 플랫폼은 다양한 이슈 논의가 쉽고 실시간으로 이루어져 공론화가 용이하다.[11] 결국 개인들의 아이디어, 태도, 감정 등이 어우러져 여론을 형성하게 된다.[12] 여론은 사회적으로 구성된 대중의 표현이다.[13] 소셜 미디어의 여론은 비합리성, 강한 감염성 및 적합성의 특성을 가진다.[14] 사회적 이슈에 대한 여론 파악을 위해서 이러한 비정형 빅데이터를 분석하는 것은 정책 발전에 큰 도움이 될 수 있다. 기계학습(머신러닝, machine learning) 및 텍스트 마이닝의 급속한 발전은 긴급 상황에서 소셜 미디어 데이터로부터 인간 행동, 대중의 반응, 잠재적 행동 과정 및 여론에 대한 분석 및 이해를 용이하게 하였다.[15]

제 2 절 연구 목적

본 연구의 목적은 코로나바이러스감염증-19 (이하 ‘코로나19’) 관련 네이버 뉴스 댓글을 LDA 토픽 모델링 분석방법을 활용하여 분석함으로써, 코로나19 팬데믹에 대한 한국 인터넷 여론의 주요 토픽을 도출하고 각 토픽의 양상을 파악하는 데에 있었다. 뉴스 수집 기간은 코로나19 팬데믹 초기였던 2020년 2월부터 2022년 6월까지로 설정하였다. 궁극적으로 한국인들이 코로나19 팬데믹에 대하여 어떠한 인식을 가지고 있는지 파악하고자 하였다.

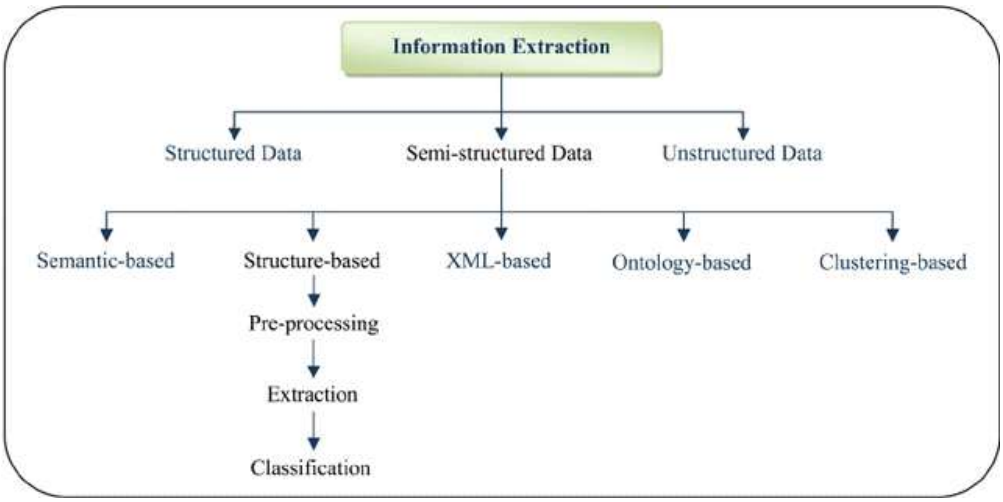
제 II 장 이론적 배경

제 1 절 빅데이터

빅데이터는 기존의 데이터 처리 응용 소프트웨어로 처리하기에는 너무 크거나 복잡한 데이터 세트를 의미한다.[16] 빅데이터라는 용어는 1990년대부터 사용되었다.[17] 빅데이터 철학은 비정형, 반정형 및 정형 데이터를 포괄하지만 주요 초점은 비정형 데이터이다.[18] 고정된 형식으로 저장, 액세스 및 처리할 수 있는 모든 데이터를 '정형' 데이터라고 한다. 현재의 컴퓨터 과학 기술 수준은 정형 데이터로 작업하고 이로부터 가치를 도출하는 기술이 잘 발달되어 있다. 그러나 이미 이러한 데이터의 크기가 엄청나게 커지리라고 예상되며, 일반적인 크기는 제타바이트에 이르고 있다.[19]

형태나 구조를 알 수 없는 모든 데이터는 비정형 데이터로 분류된다. 비정형 데이터는 엄청난 크기 외에도 가치를 끌어내기 위한 처리 측면에서 여러 문제를 가지고 있다. 비정형 데이터의 일반적인 예는 간단한 텍스트 파일, 이미지, 비디오 등의 조합을 포함하는 이질적 데이터이다. 현대에는 가용한 데이터가 풍부하지만 이러한 데이터들이 비정형 형식이기 때문에 이로부터 가치를 도출하는 방법이 명확하지 않다.[20]

반정형 데이터는 정형과 비정형 데이터를 모두 포함한다. 반정형 데이터는 형태는 구조화된 것으로 볼 수 있으나 실제로는 관계형 DBMS (Database Management System)의 표와 같이 정의되어 있지 않다. 대표적인 반정형 데이터의 예는 XML (eXtensible Markup Language) 파일로 표현되는 데이터이다. <그림 1>에 빅데이터 분석에 사용되는 데이터가 형식으로 분류되어있다.[21]



<그림 1> 데이터의 종류(출처: Shaker et al., 2008[21])

빅데이터의 속성 중 가장 중요한 것을 3V라고도 하는데, 이는 Volume (고용량), Variety (다양성), Velocity (고속)이다.[22] 이러한 빅데이터의 속성을 바탕으로 빅데이터 분석 결과로 경향을 예측하여 마케팅에 적극적으로 활용하려는 시도가 늘고 있다.[23] 소셜 미디어가 널리 사용되면서 빅데이터 분석을 이용해 사회 전 분야에서 활용이 가능해진 것이다.[24,25]

대표적인 빅데이터 분석방법에는 텍스트 마이닝(text mining), 의미연결망 분석(semantic network analysis), CONCOR (convergence of iteration correlation)분석 등이 있다. 빅데이터 분석은 여러 이점이 있다. 기업은 의사 결정에 외부 정보를 활용할 수 있다. 검색 엔진, 페이스북 및 트위터와 같은 인터넷 사이트에서 소셜 데이터에 접근함으로써 비즈니스 전략을 미세하게 조정할 수 있다. 기존의 고객 피드백 시스템 또한 빅데이터 기술로 설계된 새로운 시스템으로 대체되고 있다.[26] 이러한 새로운 시스템에서는 빅데이터 및 자연어 처리 기술을 사용하여 소비자 응답을 읽고 평가하고 있다. 빅데이터 기술은 데이터 웨어하우스(data warehouse)로 이동해야 하는 데이터를 식별하기 전에 새 데이터의 스테이징 영역(staging area) 또는 랜딩 영역(landing zone)을 만드는 데 사용할 수도 있다. 빅데이터 기술과 데이터 웨어하우스의 이러한 통합은 조직이 자주 접근하지 않는 데

이터를 처리(offload)하는 데 도움이 된다.[27)]

제 2 절 텍스트 마이닝

텍스트 데이터 마이닝이라고도 하는 텍스트 마이닝은 텍스트에서 고품질 정보를 추출하는 과정이다. 조금 더 구체적으로 말하자면, 비정형 텍스트 데이터에서 자연어 처리 후 유의미한 정보를 추출하는 것이다.[28)] 전체 텍스트 데이터의 80%가 비정형이므로 검색과 관리가 용이하지 않아 그 자체만으로는 유용한 데이터가 아니다. 수동으로 처리 시 시간 소모가 크고, 비용 과다, 부정확성과 비확장성의 문제가 발생한다. 텍스트 마이닝은 정보 검색, 자연어 처리, 정보 추출 및 데이터 마이님/지식 발견 등을 자동으로 처리하는 기술이다. 정확성, 확장성이 있으며 응답이 빨라 안정적이고 비용 효율적이다. 확장성이란 텍스트 마이닝으로 단 몇 초만에 대량의 데이터를 분석할 수 있음을 말한다. 기업의 경우 실시간 분석이 가능해져 잠재적 위기를 감지하고 제품 결함이나 부정적인 리뷰를 실시간으로 발견하여 긴급한 문제의 우선순위 지정 등 의사 결정이 쉬워질 수 있다. 또한 자동화로 시간 절약과 정확한 결과 획득, 균일한 기준의 적용이 가능하다.

텍스트 마이닝의 장점에는 다음과 같은 것들이 있다:[29)]

1) 시간과 자원을 절약하고 사람의 두뇌보다 효율적이다.
2) 시간의 흐름과 함께 의견을 추적하는 데 도움이 된다.
3) 문서 요약에 도움이 된다.
4) 더 간단한 방법으로 텍스트에서 개념을 뽑아내고 보여줄 수 있다.
5) 텍스트 마이닝으로 인덱싱된 텍스트는 예측 분석에 사용할 수 있다.
6) 관심 분야의 용어를 사용하기 위해 어떠한 단어도 연결할 수 있다.

텍스트 마이닝 절차는 크게 세 단계-데이터 수집, 전처리, 분석-로 이루어진다. 데이터 분석은 분류와 추출로 나뉠 수 있다. 데이터 수집은 데이터가 포함된 문서를 생성하는 것에서 출발한다. 데이터 전처리는 토큰화, 파싱, lemmatization, 형태 소거 및 중지 제거와 같은 몇 가지 자연어 처리 기술을 사용하여 기계 학습 모델의 입력 데이터를 생성한다. 데이터 분석 중 분류는 내용에 따라 태그 또는 범

주를 텍스트에 할당하는 과정이다. 분류 시스템은 크게 두 가지로 나눌 수 있는데, 규칙 기반 시스템과 기계 학습 기반 시스템으로 나눌 수 있다. 규칙 기반 시스템에서 규칙은 일반적으로 구문, 형태 및 어휘 패턴에 대한 참조로 구성된다. 또한 의미론 또는 음운론 측면과 관련될 수도 있다. 기계 학습 기반 시스템은 훈련 데이터로부터 학습한다. 하이브리드 시스템은 규칙 기반 시스템과 기계 학습 기반 시스템을 결합한 시스템이ᄃ. 분류에 대한 평가는 교차 검증 통계량을 활용한다. 데이터 추출은 구조화되지 않은 데이터에서 특정 정보를 얻는 과정이다. 정규식과 조건부 임의 필드(Conditional Random Fields)를 활용할 수 있는데, 정규식은 태그와 연관될 수 있는 일련의 문자를 정의하는 것이고 조건부 임의 필드는 기계 학습을 통한 텍스트 추출에 사용할 수 있는 통계적 접근 방식이다. 추출에 대한 평가는 텍스트 분류와 동일한 성능평가 행렬을 사용하여 텍스트 추출기를 평가함으로써 이루어진다.[30)]

텍스트 마이닝을 통한 분석방법으로는 TF-IDF (Term Frequency-Inverse Document Frequency)와 N-gram이 많이 사용된다.

TF-IDF는 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다.[31)] 이는 단어 빈도와 역문서 빈도의 곱으로 구한다. TF값을 산출하는 방식에는 불린 빈도, 로그 스케일 빈도, 증가 빈도 등이 있으며, IDF값을 산출하는 방식은 아래 식과 같다.[32)]

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

분자: 전체 문서의 수, 분모: 단어 t가 포함된 문서의 수

TF-IDF는 다음과 같이 표현한다.

$$tfidf(t,d,D) = tf(t,d) * idf(t,D)$$

TF-IDF 분석은 특정 키워드의 빈도수에 따라 가중치 값을 계산하므로 단순 빈도수로 판단할 때보다 정확하게 중요한 단어 파악이 가능하다.

N-gram은 (n-1)차수 Markov 모델의 형태로 특정 시퀀스에서 다음 항목을 예측하기 위한 확률적 언어 모델이다. 이 모델은 통계적 자연어 처리에 널리 사용된다.[33] N-gram 분석은 알고리즘이 단순하고 검색 누락이 발생하지 않는다는 이점이 있으나, 검색 노이즈가 크고 데이터베이스 용량(인덱스 크기)이 커진다는 단점도 있다.

제 3 절 토픽 모델링

자연어 처리에서 토픽 모델(topic model)은 문서에서 발생하는 추상적인 "토픽(주제)"을 발견하기 위한 일종의 통계 모델이다. 토픽 모델링은 텍스트 본문에서 숨겨진 의미 구조를 발견하기 위해 자주 사용되는 텍스트 마이닝 도구이다.[34]

초기 토픽 모델은 1998년 Papadimitriou, Raghavan, Tamaki 및 Vempala에 의해 제시되었다.[35] PLSA (Probabilistic latent semantic analysis)라는 또 다른 방법은 Thomas Hofmann이 1999년에 만들었다.[36] 현재 가장 일반적으로 사용되는 LDA (Latent Dirichlet Allocation)는 PLSA의 일반화된 방법이다. 2002년 David Blei, Andrew Ng 및 Michael I. Jordan이 개발한 LDA는 문서 토픽 및 토픽-단어 분포에 대한 희소 Dirichlet 사전 분포(sparse Dirichlet prior distribution)를 도입하여 문서가 적은 수의 토픽을 다루고, 토픽은 주로 적은 수의 단어를 사용한다는 직관을 인코딩한다.[37] 디리클레 분포는 베타 분포의 다변량 일반화 분포이기도 해서, 다변량 베타 분포(MBD, Multivariate beta distribution)이라고도 한다.[38] 쉽게 말하면, 단어들의 집합이 어떤 토픽들로 묶인다고 가정하고, 이 단어들이 각각의 토픽에 구성될 확률을 계산하여 결괏값을 토픽에 해당할 가능성이 큰 단어들의 집합으로 추출하는 방식이다. 깁스 샘플링을 활용하는데, 우선 나머지 단어는 고정한 채 한 단어만을 빼고 분포를 추론한다. 이렇게 단어를 하나씩 제외하고 추론하면 제외된 단어에 대해 전체 분포가 추정된다. 마지막으로 모든 단어에 대하여 두 가지 분포를 업데이트한다. LDA 분석 방법의 한계에는 이렇게 샘플링을 이용하므로 실행 시마다 결과가 달라질 수 있다는 점이 있다. 문서 집합

이 작을수록 실행 결과가 많이 달라지 확률이 커진다. 또한, 단어의 분포만을 가지고 토픽을 묶으므로 실제 토픽과는 다를 수 있고, 한 문서 내에서 각 토픽 간의 연관성은 찾을 수 없다는 점도 한계가 있다.[39)]

이러한 한계점을 극복하기 위하여 단어의 상대적인 중요성을 고려한 LDA 토픽 모델링이 있는데, 이는 단어마다 가중치를 부여하여 확률분포를 계산한다. 중요하지 않은 단어에는 가중치를 덜 부여하여 토픽을 더욱 더 잘 찾을 수 있도록 유도한다. 토픽의 레이블링(labeling)에 대한 한계는 아직 연구가 많지 않아 딥러닝을 활용한 토픽 모델링에 대한 기법으로 해결을 시도하고자 하는 노력이 있다.

제 4 절 감성 분석

감성 분석(감정 분석, 오피니언 마이닝 또는 감정 AI라고도 한다)은 자연어 처리, 텍스트 분석, 컴퓨터 언어학 및 생체 인식을 사용하여 정서적 상태와 주관적 정보를 체계적으로 식별, 추출, 수량화 및 연구한다. 이는 리뷰 및 설문 조사 응답, 온라인 및 소셜 미디어와 같은 고객의 소리 자료, 마케팅의 고객 서비스, 임상 의학에 이르기까지 다양한 자료에 널리 적용된다. RoBERTa (A Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach)와 같은 심층 언어 모델의 등장으로 뉴스와 같이 일반적으로 글쓴이의 의견이나 감정을 명시적으로 표현하지 않는 텍스트도 분석할 수 있게 되었다.[40)]

감성 분석에 대한 기존 접근 방식은 지식 기반 기술, 통계적 방법 및 하이브리드 접근의 세 가지 주요 범주로 나눌 수 있다.[41)] 지식 기반 기술은 행복, 슬픔, 두려움, 지루함과 같은 명확한 감정 단어의 존재를 기반으로 감정 범주별로 텍스트를 분류한다.[42)] 통계적 방법은 잠재 의미론적 분석, 지원 벡터 머신, 의미론적 방향을 위한 “포인트별 상호 정보(Pointwise Mutual Information)”, 의미론적 공간 모델 또는 단어 임베딩 모델, 딥러닝과 같은 기계 학습의 요소를 활용한다.[43,44)] 보다 정교한 방법은 감정의 소유자(그 감정 상태를 유지하는 사람)와 대상(감정이 느껴지는 개체)을 감지하려고 시도한다.[45)] 하이브리드 접근은 기계 학습과 의미 네트워

크와 같은 지식 표현의 요소를 활용하여 미묘한 방식으로 표현되는 의미를 감지한다.[46)]

제 III 장 연구방법

제 1 절 데이터 수집

분석 대상 온라인 뉴스는 네이버 뉴스(www.naver.com)를 이용하였다. 분석 기간은 2020년 2월 25일부터 2022년 6월 20일까지 총 약 27개월이다. Python을 이용하여 116개의 뉴스 기사로부터 댓글 6860개를 추출하였다. 이 중 "작성자에 의해 삭제된 댓글입니다."를 제외하고 분석 가능한 댓글은 모두 4837개였다.

뉴스 기사 수집을 위한 검색어는 '속보', '코로나19', '신규확진'이었다. 세 개의 검색어가 동시에 출현하는 뉴스 기사들만을 추출하였다. 전날 확진자가 다음날 반영되는 것을 고려하여 검사 수가 많은 월요일 확진자 수가 반영된 화요일 뉴스 기사를 수집하였다.

뉴스 기사 및 댓글은 Python의 셀레니움(Selenium) 패키지를 이용하여 크롤링(crawling) 기법을 통해 수집하였다.

제 2 절 데이터 분석

본 연구에서는 코로나19 신규확진 속보 뉴스 댓글 텍스트에 내재된 토픽을 추출하기 위해, LDA 기법을 사용하였다. NetMiner4 프로그램을 활용하여 추출한 댓글을 텍스트 파일로 변환시켜 LDA 분석(alpha=0.5, beta=0.01)을 실시하였다.

토픽의 수(kappa)를 결정하기 위해, 샘플링 반복횟수를 1,000회로 설정하여 토픽의 수를 3개에서 10개로 변경하면서 토픽 모델링을 실시한 결과, 3개의 토픽을 추출했을 경우 범주별로 해석이 가장 합리적이고 용이하였다. 그리하여 최종 토픽 수는 3개로 설정하였다. 토픽 모델링 결과, 토픽별로 출현 빈도가 높은 순으로 키워드 단어가 도출되었다. 이후 도출된 주요 키워드에 맞게 토픽명을 부여하고, 토픽별 단어 네트워크를 분석하였다.

감성 분석에는 군산대학교 소프트웨어융합공학과 감성사전(작성자: 온병원, 박상민, 나철원)을 이용하였다. (군산대학교 소프트웨어융합공학과 Data Intelligence Lab 홈페이지: http://dilab.kunsan.ac.kr/) 긍부정어를 모두 1점으로 하고, 각 문장에 대하여 긍정에서 부정을 빼는 방식으로 감성 분석을 하였다. 어떤 문장에서 긍정어 1개와 부정어 2개가 있다면, 1(긍정어) − 2(부정어) = −1이 나와 부정이 되는 방식이다. 본 한국어 감성사전은 국립국어원 표준국어대사전의 뜻풀이(glosses) 분석을 통한 긍부정 추출, 김은영(2004)[47]의 긍부정어 목록, SentiWordNet 및 SenticNet−5.0에서 주로 사용되는 긍부정어 번역, 최근 온라인에서 많이 사용되는 축약어 및 긍부정 이모티콘 목록으로부터 통합되어 개발되었다. 표준국어대사전을 구성하는 형용사, 부사, 동사, 명사의 모든 뜻풀이에 대해 긍정, 중립, 부정으로 분류하기 위해 Bi−LSTM (Bidirectional−Long Short−Term Memory) 딥러닝 모델을 사용하였다. Bi−LSTM 모델은 각 뜻풀이의 확률값을 계산하여 최종적으로 300,000개에 달하는 뜻풀이를 긍정, 중립, 부정으로 분류하였다. 긍정으로 분류된 뜻풀이 그룹에서 상위 2,500개 긍정어를 추출하였고, 유사한 방식으로 상위 2,500개 부정어를 추출하였다. 상위 2,500개만을 추출한 이유는 2,500개 초과 넘어가면 이미 추출된 긍부정어들이 반복적으로 추출되기 때문이었다. 최소 3명의 평가자가 각 단어의 긍정, 중립, 부정을 판별하고, 이의가 있을 경우 토론을 통해 합의를 이루는 방식을 사용하였다. 각 단어의 긍부정 판별은 매우 부정, 부정, 중립, 긍정, 매우 긍정 등 리커트 척도를 이용하여 평가자들의 합의를 통해 선택하였다.

감성 분석은 R 소프트웨어(version 4.2.0)를 이용하였다. R 스크립트에서 주요 함수 부분은 아래와 같다.

```r
sentimental = function(sentences, positive, negative){

  scores = laply(sentences, function(sentence, positive, negative) {

    sentence = gsub('[[:punct:]]', '', sentence)
    sentence = gsub('[[:cntrl:]]', '', sentence)
    sentence = gsub('\\d+', '', sentence)

    word.list = str_split(sentence, '\\s+')
    words = unlist(word.list)

    pos.matches = match(words, positive)
    neg.matches = match(words, negative)

    pos.matches = !is.na(pos.matches)
    neg.matches = !is.na(neg.matches)

    score = sum(pos.matches) - sum(neg.matches)
    return(score)
  }, positive, negative)

  scores.df = data.frame(score=scores, text=sentences)
  return(scores.df)
}
```

IV 장 연구결과

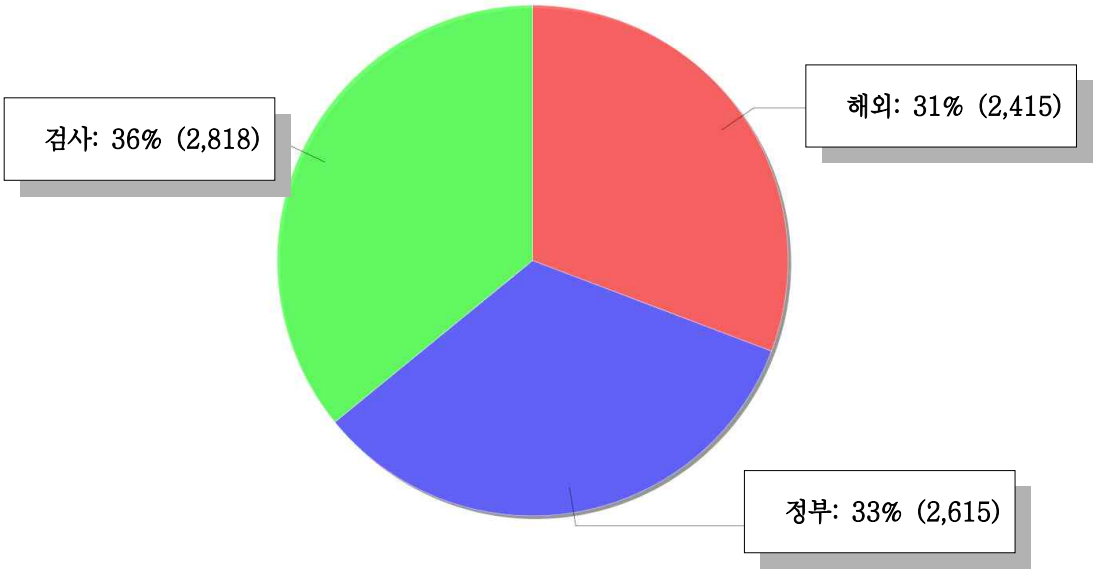
제 1 절 토픽 모델링 결과

본 연구에서는 명사를 기준으로 댓글의 주요 키워드를 살펴보았다. 전체 댓글에서 20개의 주요 키워드를 <표 1>에 제시하였다. 전반적으로 코로나19 신규확진 속보 뉴스 기사에 달린 댓글에는 '코로나(2.45%)', '확진(2.21%)', '백신(1.76%)', '국민(1.56%)', '방역(1.53%)' 등의 명사가 상위를 차지하고 있었다. 나머지 상위 키워드에는 '사람', '검사', '정부', '해외', '나라', '댓글', '표현', '유입', '감지', '클린', '문재인', '숫자', '단계', '마스크', '거리'가 있었다. 주요 키워드 20개 중 고유명사는 전 대통령의 이름인 '문재인' 뿐이었다.

<표 1> 전체 댓글에 나타난 상위 명사 빈도 수

No.	단어	N	%
1	코로나	735	2.45
2	확진	663	2.21
3	백신	528	1.76
4	국민	467	1.56
5	방역	460	1.53
6	사람	435	1.45
7	검사	426	1.42
8	정부	375	1.25
9	해외	336	1.12
10	나라	271	0.90
11	댓글	261	0.87
12	표현	211	0.70
13	유입	210	0.70
14	감지	207	0.69
15	클린	207	0.69
16	문재인	187	0.62
17	숫자	187	0.62
18	단계	185	0.62
19	마스크	179	0.60
20	거리	174	0.58

<표 2>에는 기사에 달린 댓글의 토픽 유형을 제시하였다. 최종적으로 선택된 LDA 모델의 토픽 개수는 3개였다. 전체 댓글을 토픽별로 분류한 파이 그래프는 <그림 2>에 나타내었다. 세 개의 토픽 모두 비슷한 비율로 나타났다. “해외” 토픽이 31%, “정부” 토픽은 33%, “검사” 토픽이 36%를 차지하였다. “해외” 토픽에 해당하는 댓글 수는 2,415개였고, “정부” 토픽에는 2,615개의 댓글이 해당하였다. “검사”토픽에는 2,818개 댓글이 포함되었다.



<그림 2> 코로나19 신규확진 속보 기사 댓글에서 각 토픽이 차지하는 비율

첫 번째 토픽은 "해외"이다. 해당 토픽에 속한 댓글들에 나타난 주요 키워드를 살펴본 결과, '해외', '사람', '유입', '단계', '마스크', '입국' 등 해외로부터 한국으로 입국하여 유입되는 코로나19 확진 환자들에 대한 표현이 많았다. 아래 제시한 문장들은 첫 번째 토픽에 해당하는 댓글 중 일부이다.

"중국인 입국 통제 안한 정부 때문이냐"
"결국 해외 유입자 안 막아서 이 꼴 난거임"
"앞으룬 해외입국 향후 5년간 막거나 강화 하고 一입출국사람은 자가격리 무조건 3주 一느슨해져서는 절대 안된다"
"지금도 시진핑 방한에 눈이멀어 중국인 입국제한을 안한다"
"청와대가 컨트롤 타워라며"
"해외유입때문에 300명대인데 해외유입 그대로 둘거냐"
"태극기부대빤스들 또 시작 났구나 이런 놈들만 없으면 코로나 청정 국가인데 이간질 지역감정 선동 없에야할 집단들"
"앞으룬 해외입국 향후 5년간 막거나 강화 하고 一입출국사람은 자가격리 무

조건 3주 一느슨해져서는 절대 안된다”

두 번째 토픽은 “정부”이다. 해당 토픽에 속한 댓글들에 나타난 주요 키워드에는 ‘신뢰’, ‘조선’, ‘지지율’, ‘문죄’ 등 정부를 평가하는 표현이 많았다. 아래 제시한 문장들은 두 번째 토픽에 해당하는 댓글 중 일부이다.

“문재인폐렴”
“대깨문들이 근원지다”
“지금 무능한 문재앙 정부는 애시당초 능력도 안돼 컨트롤 못하고 확산만 시키는 무능함에 극치를 보았다”
“세계 많은국가에서 한국인 입국을 거부하니 이게 어찌된일인가 본원지인 중국은 중국이지만 한국에서 확진자가 많이나오는것은 중국에 빌붙은 친 공산사회 친중정책으로인한 우리국민의 중국출입이 많은결과이며 정부의 질병관리 대책이 부족한결과라고본다”
“지금 중국 전역에서 이제 한국인들 역병 취급 당하고 있다이게 말이나 되는 소리인가 그렇게 초기에 중국인 입국금지 하라고 원성이 하늘을 찌르는 소리가 전국 방방곡곡에서 일어나도 눈하나 꿈쩍도 안하고 이제 국민들이 되레 중국에서 전염병자로 찍혀 생업이고 삶이고 엉망진창으로 만들어 놓은 문재인 정부에 묻고 싶다”
“그것은 국가간 외교문제・공산품 수출하고 농산물・ 원자재 수입하는 나라에서 쇄국정책을 펼수는 없지요 ・ 그래서 세계가 우리 K바멱을 칭송하는 것입니다”
“이제 개인방역수칙을 모르는 분들은 없을것이고 중요한것은 생활속 실천과 함께 느슨한 방심을 막는것이 최고의 예방일것입니다”
“"문통이 입만열면 자화자찬하던 세계가 알아주는 K-방역 여름휴가갔냐, 왜 확진자가 이렇게 많이 나오냐”

세 번째 토픽은 “검사”이다. 해당 토픽에 속한 댓글들에 나타난 주요 키워드

에는 ‘확진’, ‘검사’, ‘코로나’, ‘숫자’ 등 코로나 검사 및 확진자 수에 대한 표현이 많았다. 아래 제시한 문장들은 세 번째 토픽에 해당하는 댓글 중 일부이다.

“확진자랑 접촉한적 있다고 어제 전화했음 그랬더니증상이 어떠냐묻길래 아무 증상 없는데 검사 받아야 하는거 아닌가요”

“국회, 공무원말고 일반인이 전화해서 대구, 은평성모 다녀왔고 열은 없는데 기침, 가슴에 통증이있다고하면 일단 더 지켜보자고 검사도 안해준다”

“검사대기자는 만3천인데 꼴랑 60명”

“바이러스는 겨울에 더 극성인데 겨울이라 모기없다고 한 작자도 정신상태가 의심스러우니 진단받고 폐쇄병동 입원 고려해야함""

“자칭 재림예수 이며 육신을 입은 성령이라는 그라면 지금 그가 있을 곳은 역병에 신음하는 환자들의곁이어야하며 그들이 낫기 위해 손을 얹어야 하는 것이 아닌가”

“검사중인 사람이 13,000 명이 넘으니 엉터리 키트로 음성나올 때까지 계속 다시 검사하는 중”

“환자가 병원 못 찾아서 돌아다니다가 구급차 안에서 사망하는 지경인데 안정적으로 관리 하고 있다는 해괴한 말을 하는 문재인”

“매일매일 속보로 숫자 적어주는거 정보도아니고 아무의미없는짖 작년부터 2년동안 코로나로 죽은사람이 2천명 조금넘냐”

<표 2> 코로나19 신규확진 속보 뉴스 기사에 달린 댓글의 토픽 유형

No.	토픽 유형	단어	비율
1	해외	해외/사람/유입/단계/마스크/거리/입국/코로나/확진/격리	31%
2	정부	신뢰/조선/지지율/문죄/정은경/공무원/정신/정상/부정/신경	33%
3	검사	확진/검사/코로나/댓글/표현/감지/숫자/자수/감기/환자	36%

<그림 3>, <그림 4>, <그림 5>, <그림 6>은 전체 토픽과 각 토픽별로 단어 클라우드를 보여준다. 글씨 크기는 단어의 출현 빈도에 비례하여 증가하였다.

<그림 3> 전체 토픽(전체 댓글)에 대한 단어 클라우드

전체 토픽 또는 전체 댓글에 대한 단어 클라우드에서 눈에 띄는 단어는 '코로나', '확진', '사람', '국민', '백신', '방역', '정부', '검사', '문재인' 등이 있었다. 코로나19 팬데믹이 지속되고, 정부의 방역 지침이 지속적으로 업데이트되면서 관련 단어가 댓글에서 많이 등장하였다.

<그림 4> 첫 번째 토픽(해외)에 대한 단어 클라우드

"해외" 토픽에 대한 단어 클라우드에서 눈에 띄는 단어는 '해외', '사람', '입국', '유입', '대구', '서울', '단계', '마스크' 등이었다. 코

로나19 팬데믹이 전 세계적이었기 때문에 해외로부터 유입되는 확진자라도 차단하고 싶어 하는 것이 여론이었다.

<그림 5> 두 번째 토픽(정부)에 대한 단어 클라우드

“정부” 토픽에 대한 단어 클라우드에서 눈에 띄는 단어는 ‘백신’, ‘정부’, ‘국민’, ‘방역’, ‘나라’, ‘문재인’, ‘재앙’, ‘의료진’ 등이었다. 한국 정부에서 코로나19 백신 접종을 의무화하였고, 방역 지침은 국민들의 일상생활 피로도를 증가시켜 관련 단어가 댓글에서 많이 등장하였다.

<그림 6> 세 번째 토픽(검사)에 대한 단어 클라우드

“검사” 토픽에 대한 단어 클라우드에서 눈에 띄는 단어는 ‘확진’, ‘검사’, ‘코로나’, ‘클린’, ‘감지’, ‘자수’, ‘숫자’ 등이 있었다. 코로나19 검사를 받는 사람들의 수도 증가하고, 코로나19 검사에 소요되는 시간과 비용이 증가하자 관련 단어가 댓글에 많이 등장하였다.

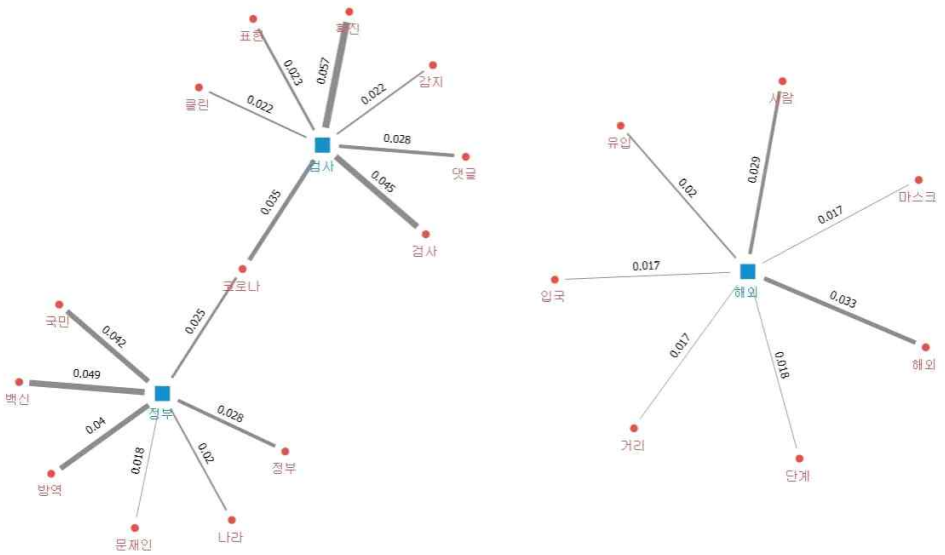
제 2 절 단어 네트워크 결과

<그림 7>은 전체 토픽에 대한 단어 네트워크를 보여준다. ‘코로나’라는 단어는 “정부”와 “검사” 토픽 모두와 밀접한 관련이 있었다. “해외” 토픽과 가장 관련이 높은 단어는 ‘해외’였으며(가중치 0.033), “정부” 토픽과 가장 관련이 높은 단어는 ‘백신’(가중치 0.049), “검사” 토픽과 가장 관련이 높은 단어는 ‘확진’이었다(가중치 0.057).

“해외” 토픽에 대한 단어 네트워크를 자세히 살펴보면, ‘해외’, ‘사람’, ‘유입’, ‘단계’, ‘마스크’와 ‘거리’와 ‘입국’ 순으로 가중치가 높았다. 가중치는 순서대로 0.033, 0.029, 0.02, 0.018, 0.017였다.

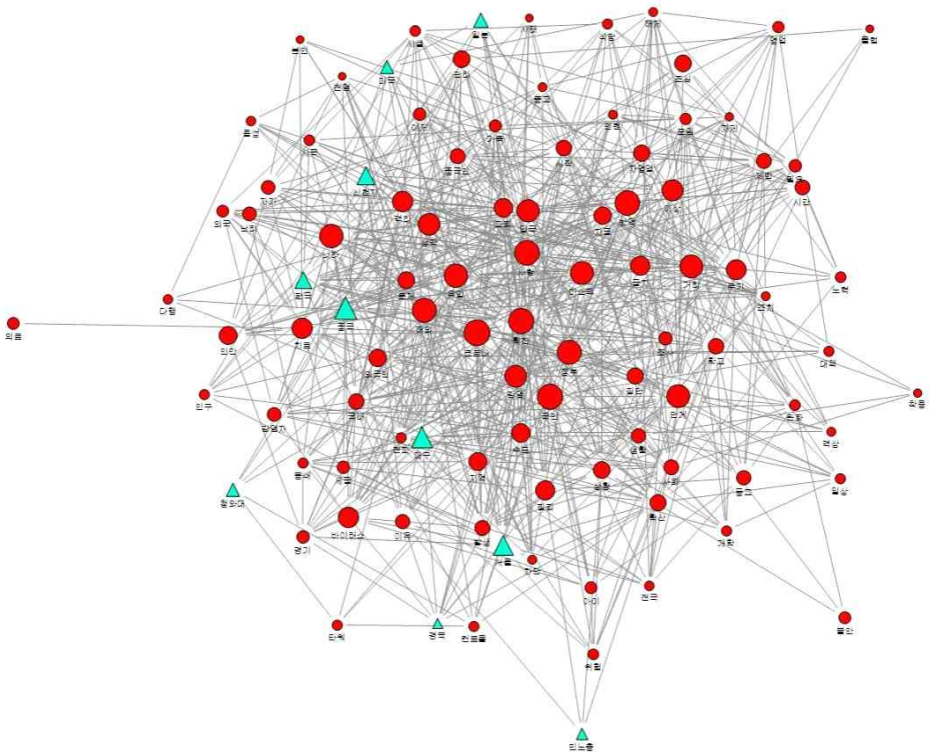
“정부” 토픽에 대한 단어 네트워크에서는, ‘백신’, ‘국민’, ‘방역’, ‘정부’, ‘코로나’, ‘나라’, ‘문재인’ 순으로 가중치가 높았다. 가중치는 순서대로 0.049, 0.042, 0.04, 0.028, 0.025, 0.02, 0.018이었다.

“검사” 토픽에 대한 단어 네트워크에서는, ‘확진’, ‘검사’, ‘코로나’, ‘댓글’, ‘표현’, ‘감지’와 ‘클린’ 순으로 가중치가 높았다. 가중치는 순서대로 0.057, 0.045, 0.035, 0.028, 0.023, 0.022였다.



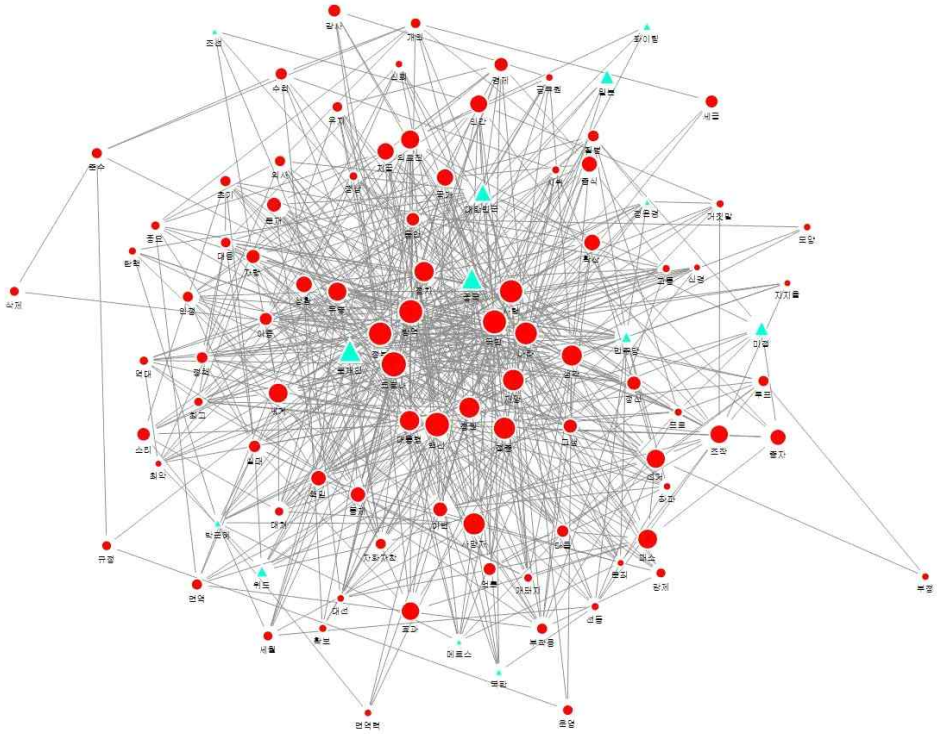
<그림 7> 전체 토픽에 대한 단어 네트워크(숫자: 가중치, 선굵기: 가중치에 비례)

<그림 8>, <그림 9>, <그림 10>는 토픽별 단어 네트워크를 나타낸다. 일반명사는 빨간 원, 고유명사는 하늘색 삼각형으로 표현하였으며, 도형의 크기는 단어 출현 빈도 수와 비례한다.



<그림 8> 첫 번째 토픽(해외)에 대한 단어 네트워크

“해외” 토픽에 대한 단어 네트워크에 출현하는 주된 고유명사에는 ‘중국’, ‘신천지’, ‘대구’, ‘서울’, ‘일본’, ‘미국’, ‘청와대’, ‘민노총’ 등이 있다. ‘민노총’과 연관된 단어에는 ‘이유’, ‘집회’, ‘확산’ 등이 있었다. ‘청와대’와 연관된 단어에는 ‘컨트롤’, ‘타워’ 등이 있었다. 단어 네트워크의 중심부에 있는 명사에는 ‘확진’, ‘정부’, ‘국민’, ‘유입’, ‘해외’, ‘수도’ 등이 있었다.



<그림 9> 두 번째 토픽(정부)에 대한 단어 네트워크

“정부” 토픽에 대한 단어 네트워크에 출현하는 주된 고유명사에는 ‘대한민국’, ‘문재인’, ‘중국’, ‘민주당’, ‘정은경’, ‘메르스’ 등이 있다. ‘정은경’과 관련된 단어에는 ‘민주당’, ‘정치’, ‘재앙’, ‘거짓말’ 등이 있었다. ‘메르스’와 관련된 단어에는 ‘박근혜’, ‘사망자’, ‘언론’, ‘다음’ 등이 있었다. 코로나19 팬데믹이 시작된 시점의 정권에 영향을 받았다고 볼 수 있다. 단어 네트워크의 중심부에 있는 명사에는 ‘대통령’, ‘정권’, ‘재앙’, ‘정치’, ‘무능’, ‘생각’, ‘사람’, 등이 있었다.

<그림 10> 세 번째 토픽(검사)에 대한 단어 네트워크

"검사" 토픽에 대한 단어 네트워크에 출현하는 주된 고유명사는 '조선일보' 하나뿐이다. 아래 제시한 문장들은 '조선일보'가 들어가는 댓글 중 일부이다.

"조선일보는 박근혜털었듯이 신천지 좀털어봐"

"조선일보 진짜 꿋꿋하다"

"무슨 우한코로나예요 조선일보 똑바로 기사쓰세요"

"ㅎㅎ 조선일보는 지금 망국을 불러오고 있다"

"기자양반,,,, 우한코로나라는 용어는 코로나19로 통일해서 사용하기로 한건데 아직도 우한코로나야,,,,,,, 이런 우라질,,,, 그러니 당신같은 기자들이 대구코로나 말을 만들어 내지,,, 조선일보 수준이하구만,,,, 똥인지 된장인지 구분좀 하고 기사 씁시다,,,"

"조선일보는 어떻게 생각하세요"

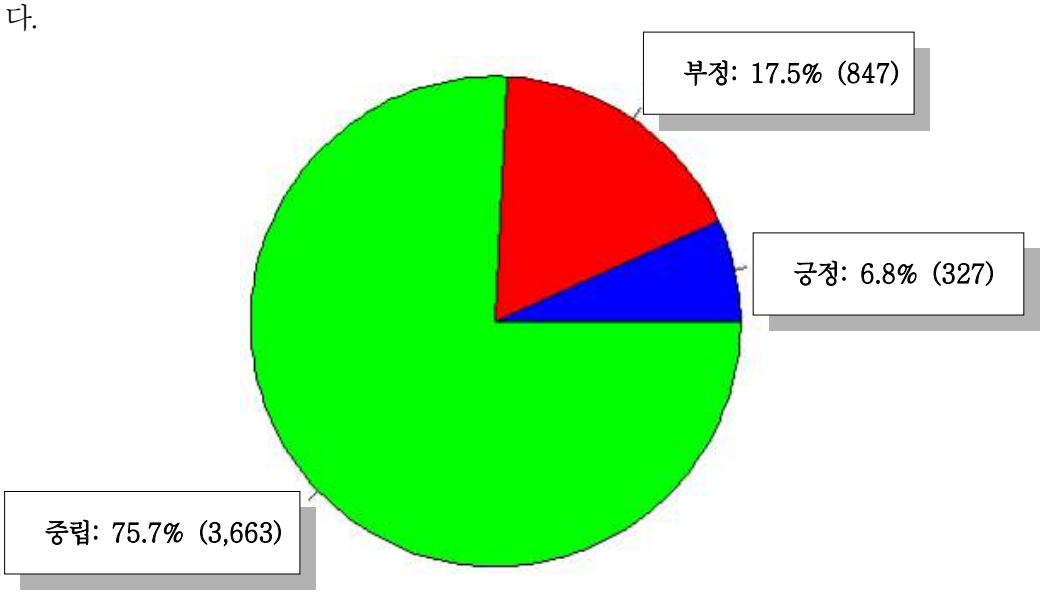
<그림 11>의 가장 오른쪽에 있는 세 개 단어, '댓글', '표현', '감지'는 아래와 같은 댓글 때문에 나타난 것으로 생각된다.

"클린봇이 부적절한 표현을 감지한 댓글입니다"

위 댓글은 네이버 뉴스 기사에 댓글을 등록했을 때 적절하지 않다고 판단된 경우 원래 댓글은 가려지면서, 자동으로 보여지는 댓글 문구이다.

제 3 절 감성 분석 결과

전체 댓글 4,837개에 대하여 군산대학교 감성사전으로 감성 분석을 한 결과, 긍정 비율은 6.8%(327개 댓글), 부정 비율은 17.5%(847개 댓글), 중립 비율은 75.7%(3,663개 댓글)였다. 본 연구에서 수행한 감성 분석의 방법론상, 한 문장 내에 긍정 단어와 부정 단어 개수가 같다면 중립으로 평가되었다. 그러므로 중립으로 분류된 댓글을 제외하고 결과를 분석하면, 전체 1,174개 댓글 중 긍정 비율은 27.8%, 부정 비율은 72.2%였다. 부정 비율이 긍정 비율보다 약 2.6배 많았다. <그림 11>는 코로나19 신규확진 속보 기사 댓글의 감성 분석 결과를 도식화하였다.



<그림 11> 코로나19 신규확진 속보 기사 댓글 감성 분석 결과

제 Ⅴ 장 결론

제 1 절 요약 및 결론

본 연구는 2020년 2월부터 코로나19 팬데믹이 선언된 이후 약 27개월 동안 온라인 뉴스 기사 댓글로 LDA 토픽 모델링을 수행하여 한국인들의 인식을 파악하고자 하였다. 116개의 뉴스 기사에 달린 댓글 약 6000여 개 중 분석 가능한 댓글 총 4837개를 분석에 이용하였다. 전체 텍스트 중 명사만 추출하여 사용하였다. 사용한 소프트웨어는 Python과 NetMiner4, R이다. Python의 셀레니움 패키지를 이용하여 웹 크롤링을 하였고, NetMiner4를 이용하여 텍스트 마이닝과 토픽 모델링 후, R을 이용하여 군산대학교 감성사전으로 감성 분석을 진행하였다.

코로나19 신규확진 속보 뉴스 기사에 달린 댓글에는 '코로나', '확진', '백신', '국민', '방역' 등의 명사가 순서대로 상위를 차지하고 있었다.

총 세 개의 토픽이 선정되었으며, "해외", "정부", "검사" 토픽이 거의 유사한 비율로 도출되었다. "해외" 토픽에서 가장 중요한 키워드는 '해외'였고, "정부" 토픽에서는 '백신, "검사" 토픽에서는 ' 확진 '이 가장 중요한 키워드였다. 각 토픽별로 주요 키워드와 댓글을 구체적으로 살펴보면 부정적인 단어가 많았다.

감성 분석 결과, 부정 비율이 긍정 비율보다 약 2.6배 많았다. 중립으로 분류된 댓글을 제외하면, 긍정 비율은 27.8%, 부정 비율은 72.2%였다.

코로나19 팬데믹이 2년 이상 지속되면서 한국의 방역과 의료 체계에 대한 불만 여론이 본 연구를 통해서도 드러났다.

제 2 절 시사점

본 연구와 같이 약 120여 개의 뉴스 기사 댓글의 텍스트 마이닝과 토픽 모델링 분석만 해도 여론 파악이 어느 정도 가능하다. 코로나19 팬데믹과 관련하여 한국

인들이 정부의 대응과 코로나19의 심각성을 어떻게 인식하고 있는지 알 수 있었다. 그러므로 앞으로도 여론 파악에 이러한 빅데이터 분석 기술을 적극적으로 활용할 수 있겠다.

제 3 절 연구의 한계 및 향후 연구 과제

본 연구의 한계점에는 뉴스 기사와 댓글의 수가 상대적으로 적었다는 점이 있다. 조금 더 포괄적으로 뉴스 기사를 수집하고, 댓글 수가 더 많았다면 시기별 여론 경향 분석도 가능했을 것이다.

또한 본 연구에서 택한 감성 분석 방법 외에 더 섬세한 방법을 사용했다면, 긍정의 정도와 부정의 정도를 수치화된 척도로 나타냈을 수 있겠다. 본 연구에서는 문장 전체에 점수를 매겨서 양수면 긍정, 음수면 부정으로 이분화하였으나, 소수점까지 수치화할 수 있는 방법을 사용했다면 더 구체적인 여론 인식을 알 수 있었을 것으로 예상된다.

웹 크롤링을 통하여 수집한 데이터로 텍스트 마이닝, 토픽 모델링 후 감성 분석까지 하는 일련의 연구방법으로 많은 연구가 이미 진행되었다. 학계에서 이러한 방법론의 보편화가 이루어지고 있으므로 앞으로는 산업계에서 실제 비즈니스와 실생활에 활발하게 응용할 수 있겠다.

참고문헌

1) 질병관리청 홈페이지(http://ncov.mohw.go.kr/bdBoardList_Real.do), Last access date: 2022.07.04.

2) World Health Organization (WHO), "COVID-19 Weekly Epidemiological Update", Edition 98, published 29 June 2022

3) Centers for Disease Control and Prevention (CDC) 홈페이지(https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html), Last access date: 2022.07.06.

4) Paderno, A., Mattavelli, D., Rampinelli, V., Grammatica, A., Raffetti, E., Tomasoni, M., ... & Schreiber, A. (2020), "Olfactory and gustatory outcomes in COVID-19: a prospective evaluation in nonhospitalized subjects", *Otolaryngology-Head and Neck Surgery,* 163(6), 1144-1149

5) Wang, C. C., Prather, K. A., Sznitman, J., Jimenez, J. L., Lakdawala, S. S., Tufekci, Z., & Marr, L. C. (2021), "Airborne transmission of respiratory viruses", *Science,* 373(6558), eabd9149

6) Mallapaty, S., Callaway, E., Kozlov, M., Ledford, H., Pickrell, J., & Van Noorden, R. (2021), "How COVID vaccines shaped 2021 in eight powerful charts", *Nature,* 600(7890), 580-583

7) 질병관리청 홈페이지(https://ncv.kdca.go.kr/mainStatus.es?mid=a11702000000), Last access date: 2022.07.06.

8) Word Health Organization 홈페이지(https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems), "Impact of COVID-19 on people's livelihoods, their

health and our food systems “, Last access date: 2022.07.06.
9) 김혜미, 이준웅(2011), “인터넷 뉴스와 댓글의 뉴스 프레임 융합 효과 연구”, 한국언론학보, 55(2), 33-55
10) 양혜승(2008), “인터넷 뉴스 댓글의 견해와 품질이 독자들의 이슈에 대한 태도에 미치는 영향”, 한국언론학보, 52(2), 254-280
11) 김동성(2015), “소셜 미디어 상에서의 여론 변화 추이 분석을 위한 감성사전 구축 방안 연구: 원자력 관련 트윗을 중심으로”, 한양대학교 경영학과 박사학위논문
12) 홍순구, 유승의, 김나랑, 이태헌, 이새미, 안순재(2020), “스마트 거버넌스 정책과정의 혁신”, 유원북스
13) McGregor, S. C. (2019), “Social media as public opinion: How journalists use social media to represent public opinion”, *Journalism,* 20(8), 1070-1086
14) Han, X., Wang, J., Zhang, M., & Wang, X. (2020), “Using social media to mine and analyze public opinion related to COVID-19 in China”, *International Journal of Environmental Research and Public Health,* 17(8), 2788
15) Steiger, E., Resch, B., & Zipf, A. (2016), “Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks”, *International Journal of Geographical Information Science,* 30(9), 1694-1716.
16) 홍민기, 조민제, 김혜정(2017), “머신러닝을 이용한 기계학습 분류법”, 대한전자공학회 학술대회 1339-1341
17) Steve Lohr(2013), “The origins of ‘Big Data’ : An etymological detective story”, (https://archive.nytimes.com/bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/), Last access date: 2022.07.04.)

18) Dedić, N., Stanier, C.(2017). "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery". *Innovations in Enterprise Information Systems Management and Engineering,* 285, 114-122

19) Hormann, P., & Campbell, L. (2014), "Data storage energy efficiency in the Zettabyte Era", *Journal of Telecommunications and the Digital Economy,* 2(3), 51-1

20) Patibandla, R. L., Kurra, S. S., Prasad, A., & Veeranjaneyulu, N. (2015), "Unstructured Data: Qualitative Analysis", *Journal of Computation In Biosciences And Engineering,* 2(3), 1-4

21) Shaker, M., Ibrahim, H., Mustapha, A., & Abdullah, L. N. (2008), "A framework for extracting information from semi-structured web data sources", In *2008 Third International Conference on Convergence and Hybrid Information Technology* 1, 27-31, IEEE

22) 이서구(2015), "빅데이터 분석에 관한 마케팅적 접근", 대한경영학회지, 28(1), 21-35

23) 강지원, 남궁영(2021), "빅데이터를 활용한 식품 유통 플랫폼에 대한 소비자 인식 분석: 텍스트 마이닝과 의미연결망 분석을 중심으로", 호텔경영학연구, 30(2), 37-52

24) 최홍열, 박은경(2019), "소셜 미디어 빅데이터 분석을 이용한 나홀로 여행 트렌드 분석: 제주도를 중심으로", 관광경영학회, 87(0), 45-66

25) 박종영, 서충원(2015), "TF-IDF 가중치 모델을 이용한 주택시장의 변화특성 분석", 부동산학보, 63(63), 46-58

26) Park, E., Jang, Y., Kim, J., Jeong, N. J., Bae, K., & Del Pobil, A. P. (2019), "Determinants of customer satisfaction with airline services: An analysis of customer feedback big data", *Journal of Retailing and Consumer Services,* 51, 186-190

27) Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013), "Big data

imperatives: Enterprise 'Big Data' warehouse, 'BI' implementations and analytics" *Apress*

28) 이재문(2021), "빅데이터 분석을 활용한 홈트레이닝 시장 전망 및 발전방안에 관한 연구", 한국체육학회지, 60(1), 189-202

29) Rose, J., & Lennerholt, C. (2017), "Low cost text mining as a strategy for qualitative researchers", *Electronic Journal of Business Research Methods*, 15(1), 2-16

30) Agrawal, R., & Batra, M. (2013), "A detailed study on text mining techniques", *International Journal of Soft Computing and Engineering*, 2(6), 118-121

31) 박대서, 김화종(2018), "TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합벡터 제안", 한국정보기술학회논문지, 16(2), 1-16

32) Salton, G., & McGill, M. J. (1983), "Introduction to modern information retrieval", mcgraw-hill

33) Dunning, T. (1994), "Statistical identification of language", Las Cruces: Computing Research Laboratory, New Mexico State University

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

35) Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000), "Latent semantic indexing: A probabilistic analysis", *Journal of Computer and System Sciences*, 61(2), 217-235

36) Hofmann, T. (1999), "Probabilistic latent semantic indexing", In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50-57

37) Blei, D., Ng, A., & Jordan, M. (2001), "Latent dirichlet allocation", *Advances in neural information processing systems*, 14

38) Kotz, S., Balakrishnan, N., & Johnson, N. L. (2004), "Continuous multivariate distributions, Volume 1: Models and applications (Vol. 1)", *John Wiley & Sons*

39) Liu, Z. (2013), "High performance latent dirichlet allocation for text mining", *Doctoral dissertation, Brunel University School of Engineering and Design PhD Theses*

40) Hamborg, F., Donnay, K., & Merlo, P. (2021). "NewsMTSC: a dataset for (multi-) target-dependent sentiment classification in political news articles", *Association for Computational Linguistics (ACL)*

41) Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). "New avenues in opinion mining and sentiment analysis", *IEEE Intelligent systems* 28(2), 15-21

42) Ortony, A., Clore, G. L., & Collins, A. (1988), "The Cognitive structure of emotions cambridge". *UK: Cambridge University Press9*

43) Sahlgren, M., Karlgren, J., & Eriksson, G. (2007). "SICS: Valence annotation based on seeds in word space", *In Proceedings of the Fourth International Workshop on Semantic Evaluations,* SemEval-2007, 296-299

44) Turney, P. D. (2002), "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", *arXiv preprint cs/0212032*

45) Kim, S. M., & Hovy, E. (2006), "Identifying and analyzing judgment opinions", *In Proceedings of the human language technology conference of the NAACL, main conference,* 200-207)

46) Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., & Hussain, A. (2015). "Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach", *Cognitive Computation,*

7(4), 487-499
47) 김은영. (2004). “국어 감정동사 연구”. 전남대학교 박사학위논문.

Abstract

Analysis of Koreans' reception of the COVID-19 pandemic using topic modeling

Kim, Wansoo
Seoul School of Integrated Sciences and Technologies
Advisor: Park, Cheong-Yeul

The COVID-19 pandemic, which started in 2019, has caused more than 158 million cases and 3.29 million deaths worldwide as of May 2021. LDA topic modeling and sentiment analysis were performed with social big data to analyze the perception of the COVID-19 pandemic in Koreans.

From February 2019 to December 2021, 116 Naver news related to the COVID-19 pandemic were collected, and a total of 4837 comments on each article were analyzed. The search terms used to collect news articles were "breaking news", "corona 19", and "new confirmation". Only articles containing all three search terms were collected. Considering that the number of confirmed cases from the previous day is reflected the next day, we collected news articles on Tuesday that reflected the number of confirmed cases on Monday, which had a high number of tests. Nouns such as 'corona', 'confirmation', 'vaccine', 'people', and 'prevention' were ranked in the comments on the news article of breaking news of new corona19 confirmed in that order.

As a result of topic modeling, three topics were derived, "overseas," "government," and "inspection," and a word network was created based on the main keywords for each topic. The most important keyword in the "overseas" topic was 'overseas', the most important keyword in the "government" topic was "vaccine" and the most im-

portant keyword was “confirmation” in the “test” topic. Looking at the main keywords and comments for each topic in detail, there were many negative words.

For sentiment analysis, Kunsan University sentiment dictionary was used. As a result, the negative rate of comments was about 2.6 times higher than the positive rate. Excluding comments classified as neutral, the positive rate was 27.8% and the negative rate was 72.2%. As the COVID-19 pandemic has continued for more than two years, this study also revealed dissatisfaction with Korea's quarantine and medical system.

The limitations of this study are that the number of news articles and comments was small, and the sentiment analysis method was semi-quantitative. If news articles were collected more comprehensively and the number of comments had been greater, it would have been possible to analyze changes in public opinion trends over time. The sentiment analysis method used in this study divided the whole sentence into positive and negative by scoring the whole sentence.

This study showed that public opinion trends can be known through online news article comment big data analysis in crisis situations such as the COVID-19 pandemic. These methodologies are becoming more common and automated, and can be actively used in business and marketing in the industry.

Key words: COVID-19, pandemic, big data, comment, social media, text mining, topic modeling, word network

Student Number: 1925406001