creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

의학석사 학위논문

텍스트 분석을 이용한 코로나바이러스감염증 19 관련 이슈 분석

- 토픽 모델링, 동시 출현 단어, 감성 분석 방법을 중심으로 -

Issue analysis using text mining on COVID-19

2021년 8월

서울대학교 대학원

의학과 예방의학 전공

송 성 희

텍스트 분석을 이용한 코로나 바이러스 감염증 19 관련 이슈 분석

- 토픽 모델링, 동시 출현 단어, 감성 분석 방법을 중심으로 -

지도교수 민 경 복

이 논문을 의학석사 학위논문으로 제출함
2021년 4월

서울대학교 대학원
의학과 예방의학 전공
송 성 희

송성희의 석사 학위논문을 인준함
2021년 7월

위 원 장 _____

부위원장 _____

위 원 _____

초록

코로나바이러스감염증-19 (코로나19)는 2019년 12월 중국 우한에서 처음 확인된 호흡기 감염질환이다. 코로나 바이러스와 관련하여 보건 의료 및 우리 사회 전반의 이슈를 파악하고자 본 연구를 수행하였다.

2020년 1월부터 2021년 3월까지 1년 3개월 동안 지식 정보 공유 포털 사이트에서 검색어 '코로나'를 키워드로 하는 질문 23,463건을 수집하였다. 수집한 문서에서 최소 의미 단위인 형태소로 분해하는 전처리 과정을 거친 후, 텍스트마이닝 방법을 통해 분석을 수행하였다. 형태소 단위로 추출한 단어의 빈도를 측정하여 '코로나'와 관련된 키워드의 중요도를 살펴보고 Latent Dirichlet Allocation (LDA) 방법을 통해 문서 내 키워드의 집합을 분석하고 8개의 주제를 추출하였다. 동시 출현 단어 분석을 통해 한 문서에 동시에 등장한 단어의 빈도를 측정하여 단어와 단어 사이의 연관성을 살펴보았으며 감성 분석을 통해 감염병에 대한 대중의 의견 또는 감정을 시간의 흐름에 따라 분석하였다.

'코로나'와 관련된 주요 이슈로는 사회, 보건 의료, 법률 등으로 분류되었고 사회적 측면에서는 고용과 급여를 둘러싼 회사와 관련된 주제가 문서에서 가장 많은 비중을 차지하여 중요한 문제가 될 수 있음을 암시하였다. 보건 의료 측면에서는 코로나19의 의심 증상과 통증을 통해 감염 여부를 파악하는 것이 대중의 걱정이자 주요 관심사라고 볼 수 있다. 코로나19이 장기화되고 '감염병예방법'이 개정됨에 따라 관련 법률에 대해 대중의 관심이 있다는 것을 확인할 수 있다. 감성 분석을 통해 코로나19에 대한 대중의 전반적인 태도는 감염병 유행 초기일수록 부정적일 수 있음을 시사한다.

본 연구는 텍스트 마이닝 방법을 통한 탐색적 연구로 코로나19 데이터에 대한 신뢰성을 높이고 감염병 대응을 위한 활용 방안을 마련하기 위해서 뉴스 보도 및 논문 초록 등을 포함하여 다양한 자료원을 추가한 연구가 필요하다.

주요어 : 코로나19, 보건, 사회, 텍스트마이닝, 토픽모델링, 감성분석
학 번 : 2019 - 23275

목 차

제 1 장 서 론 ····· 1
제 1 절 연구의 배경 ····· 1
제 2 절 연구의 목적 ····· 8

제 2 장 방 법 ····· 9
제 1 절 데이터 수집 ····· 9
제 2 절 데이터 전처리 ····· 10
제 3 절 토픽 모델링 분석 ····· 11
제 4 절 동시 출현 단어 분석 ····· 15
제 5 절 감성 분석 ····· 17

제 3 장 결 과 ····· 20
제 1 절 주요 키워드의 빈도 ····· 20
제 2 절 토픽 모델링 결과 ····· 24
제 3 절 동시 출현 단어 ····· 29
제 4 절 감성 분석 결과 ····· 35
제 4 장 고 찰 ····· 44
제 5 장 결 론 ····· 54

참고문헌 ····· 56
Abstract ····· 63

표 목 차

<표 1> 감정 범주와 감정 단어 ····· 19
<표 2> 지식 포털 사이트 내 '코로나' 관련 상위 25개 단어 검색 결과 ····· 21
<표 3> 지식 포털 사이트 내 '코로나' 관련 문서의 전문가 답변 수 ····· 22
<표 4> '코로나' 관련 토픽 모델링 주제 및 키워드 ····· 27
<표 5> '코로나' 관련 토픽별 전문가 답변 비율 (%) ····· 28
<표 6> 지식 포털 사이트 내 '코로나' 관련 상위 14개 단어-단어 분석 결과 (단위 : 건) ····· 30
<표 7> 지식 포털 사이트 내 '운동' 관련 단어-단어 분석 결과 (단위 : 건) ····· 33
<표 8> 지식 포털 내 '코로나' 관련 문서에서의 감정 지수 월별 추이 ····· 36
<표 9> 지식 포털 내 '코로나' 관련 문서에서의 감정 범주 빈도 분석 결과 (2020.01 – 2021.03) ····· 38
<표 10> 지식 포털 내 '코로나' 관련 문서에서의 감정 단어 빈도 분석 결과 (2020.01 – 2021.03) ····· 39
<표 11> 지식 포털 내 '코로나' 관련 월별 감정 범주 변화 결과 (일 평균) ····· 41

그 림 목 차

[그림 1] Latent Dirichlet Allocation 모델의 구조 도식화 ····· 12
[그림 2] LDA 모델의 토픽별 Coherence 측정값 ····· 14
[그림 3] 단어 연결망 시각화 그래프 ····· 16
[그림 4] 지식 포털 사이트 내 '코로나' 관련 상위 빈도 100개 워드 클라우드 ····· 23
[그림 5] 지식 포털 내 '코로나' 관련 토픽 모델링 결과 ··· 26
[그림 6] '코로나' 관련 주요 키워드의 단어 네트워크 분석 결과 ····· 31
[그림 7] '운동' 관련 주요 키워드의 단어 네트워크 분석 결과 ····· 34
[그림 8] 지식 포털 '코로나' 관련 문서에서의 월별 감정 지수 추이 ····· 37
[그림 9] 지식 포털 '코로나' 관련 문서에서의 감정 범주 분포 ····· 38
[그림 10] 상위 4개 감정 범주(고통, 불안, 우울, 당혹)의 월별 추이 (일 평균) ····· 42
[그림 11] 하위 4개 감정 범주(만족, 화, 지루함, 혐오)의 월별 추이 (일 평균) ····· 43

제 1 장 서 론

제 1 절 연구의 배경

코로나 바이러스의 발생

2019년 말 중국 우한에서 신종 코로나 바이러스(SARS-CoV-2)에 의하여 발병된 신종폐렴이 발견되었다. 세계적으로 확인된 감염자는 2021년 3월 14일까지 1억 명을 넘어섰고, 감염으로 250만 명 이상이 목숨을 잃은 것으로 알려졌다.[1] 학자들은 전염성, 사망률, 그리고 기원을 포함한 바이러스의 특성을 설명하기 위해 빠르게 연구를 시작했다.[2]

코로나 바이러스의 특성

코로나 바이러스는 인간을 감염 시키는 양성 단일 가닥의 대형 RNA 바이러스로 둘러싸여 있다. 1966년, 티렐과 비노가 감기에 걸린 환자들로부터 바이러스를 배양해 코로나 바이러스를 처음으로 기술하였다.[3] 원형(spherical) 바이러스의 형태와 태양 상층부 대기층인 코로나와 유사한 표면을 닮은 것에 기초하여 코로나 바이러스라고 명명하였다. 코로나 바이러스는 알파, 베타, 감마, 델타 코로나 바이러스로 총 4개의 소분류로 나뉘어진다. 알파와 베타 코로나 바이러스는 포유류로부터, 감마와 델타 코로나 바이러스는 조류로부터 유래한 것으로 알려진다. 인간을 감염시킬 수 있는 7가지 종류의 코로나 바이러스 중 베타 코로나 바이러스는 심각한 질병과 사망을 일으킬 수 있는 반면 알파 코로나 바이러스는 무증상 또는 약한 증상의 감염을 일으킨다.[4] 신종 코로나 바이러스(SARS-CoV-2)는 베타 코로나 바이러스 계통에 속하며 2002년 말, 중동에서 유행했던 중증급성호흡기증후군 코로나바이러스 (SARS-CoV)

와 유사하고 박쥐 코로나 바이러스와 96% 동일한 것으로 소개된 바 있다.[5,6]

코로나 바이러스 감염증의 현황

신종 코로나 바이러스(SARS-CoV-2)로 인한 코로나바이러스감염증-19(이하, '코로나19')이 전 세계적으로 공중 보건에 큰 위협을 가하고 있다. 2021년 3월 현재, 백신이 접종이 되고 있지만 안전하고 효과적인 코로나19의 치료제는 개발되어 있지 않다.[7] 따라서, 보건 당국은 정책적인 측면에 의존해야 하는 상황 속에 개인이 필요하고 적절한 예방 조치를 취하고 각국의 정부가 이에 따른 계획을 수립하고 대응할 수 있도록 대중에게 중요한 정보를 전달하기 위한 작업에 착수했다.[7]

코로나 바이러스 감염증의 사회적 영향

코로나19의 발생은 보건 의료 분야를 넘어서 사회 전반적인 분야에 걸쳐 전 세계 사람들의 일상생활에 큰 영향을 끼쳤다. 코로나19의 급격한 확산 때문에 대부분의 국가에서 마스크 착용 등의 보건 정책을 강화하고 있으며 사회적 거리 두기 또는 감염 의심 대상자의 격리 조치와 같은 정책이 시행되면서 많은 사람들은 일상 생활에 제약을 받고 있다.

세계보건기구(World Health Organization)는 코로나19에 대하여 2020년 3월 11일 전염병 경보 최고 단계인 '세계적 대유행'이라는 의미의 '팬데믹(Pandemic)'을 선언하였다. 세계보건기구는 코로나19의 세계적 확산 사태가 단순히 공중 보건의 위기가 아닌 사회, 경제 등 모든 분야에 영향을 미치는 위기이며, 이를 극복하기 위해 전세계 국가의 진단, 방역, 전염 차단, 바이러스에 대한 정보 취득 등의 대응을 요구했다.

국내 또한 2020년 1월 20일 첫 코로나19 확진자가 발생한 후 국민의 일상생활을 포함하여 사회 전반적으로 다양한 변화를 초래했다. 휴업,

실직, 무급휴가 등으로 인해 생활 경제의 어려움으로 이어졌고 외출과 공공시설 출입이 제한되면서 홈쇼핑, 모바일 쇼핑 등 온라인 위주의 소비 문화 패턴의 변화를 가져 왔다.[8] 초, 중, 고등학교 개학 연기와 재택근무 확대 및 비대면 접촉 등으로 실내에서 보내는 시간이 증가하였으며 사회적 거리 두기와 안면 마스크 착용과 같은 방역 지침을 전 국민이 동참함으로써 방역에 대한 인식이 강화 되었다.[9] 질병관리본부에서는 해외로부터 유입되는 바이러스와 지역사회에서의 전파를 지연 하는데 주력 하였고 방역 조치를 취하는 것을 포함해 관련 정보를 공개하는 등 효과적으로 대응하고 있다. 코로나19에 더욱 신속하게 대응하기 위해 공식적으로 드러나지 않은 대중의 반응을 살펴 볼 수 있는 텍스트 데이터를 수집하여 분석할 수 있는 시스템이 필요한 실정이다.

공중 보건 연구의 한 분야는 감염병 발생에 대한 대중의 관점과 의견을 조사하는 것을 전문으로 한다. 인터뷰와 설문 조사는 정보의 유효한 출처이긴 하지만 관점과 의견이 자연스럽지 않고 질문에 대한 답변이 정형화되어 있다고 볼 수 있다. 소셜 미디어 데이터는 실용적이고 여과되지 않은 여론의 원천으로 감염병 발생에 대해 구체적으로 활용될 수 있다. Ahmed 연구팀[10]은 에볼라 바이러스, 지카 바이러스, 돼지 독감에 대한 개별적인 사례를 분석한 후, 트위터에서 일어나고 있는 토론의 유형에서 중요한 비교를 수행하여 인터뷰에서 나타나지 않을 수도 있고 설문조사에 포함되지 않을 수도 있는 전염병에 관한 논의를 식별할 수 있는 가능성을 보여주었다. 이 밖에 소셜 미디어 데이터를 활용한 연구는 재해 추적, 위험 예측에 중요한 역할을 하였고 질병 발생의 맥락에서 건강 감시 및 통신 시스템의 데이터 원천으로도 사용되어 왔다. 이와 같이 소셜 미디어 데이터는 질병 모니터링, 대중 반응, 비상 상황, 예측 등 여러 공중 보건 분야에 적용하는 데 유용한 것으로 밝혀졌다.[11]

소셜 미디어의 파급 효과

스마트 기기의 보급으로 인해 온라인 포털 사이트 및 소셜 미디어에 게시된 문서의 수가 증가하였고 이전보다 쉽게 대중의 의견을 수집할 수 있게 되었다. 이에 따라 인터넷 게시판이나 블로그, 댓글 등을 통해 게시되는 텍스트 데이터들은 대중의 생각, 태도 그리고 관심사를 쉽게 파악할 수 있게 해 주는 중요한 정보로 이용될 수 있다. 또한, 텍스트 마이닝 기술이 발전함에 따라 양질의 텍스트 데이터를 수집하여 주제를 파악하고 의미를 분석하는 연구의 중요성 또한 부각 되었다. 텍스트 마이닝 방법은 정성적인 분석에 강점을 보이지만 사망률, 유병률 측정과 같은 정량적인 측면에서 파악이 어렵다는 특성을 가지고 있다.

소셜 미디어를 활용한 코로나 바이러스 감염증의 선행 연구

코로나19에 대한 주요 주제를 파악하기 위한 여러 텍스트 마이닝 연구는 이전부터 수행되어 왔다. Amin[12] 연구팀은 페이스북 데이터를 활용하여 2020년 1월 1일부터 2020년 5월 15일까지의 기간을 1차 시기 (2020년 1월~2월), 2차 시기 (2020년 3월~4월), 3차 시기 (2020년 5월)로 총 세 구간으로 나눈 시간적 요소를 고려한 토픽 모델링 분석을 수행하였다. 1차 시기에서는 전반적으로 보건 분야에 초점을 맞춘 것이 특징이었고 2차 시기에서는 전염병에 대응하기 위한 검역 조치를 언급하였다. 마지막 3차 시기에서 모든 사람들이 정상 생활로 돌아가려는 욕구가 나타났다고 보고했다. Iyad AlAgha[13] 연구팀은 2020년 2월부터 2020년 5월까지 약 970만 건의 트위터 문서를 수집하여 국가별로 분류하고, 토픽 모델링 분석을 통해 코로나19에 대한 주제를 20가지로 나누어 설명하였다. 주제는 발병기간 동안 발생하는 중요한 사건에 의해 영향을 받는다는 점과 각국의 대비태세와 감염 영향 정도에 따라 각국의 관심사와 정서가 다르다는 것을 소개하였다.

정부의 코로나 바이러스 대응을 위한 보건 정책 속에서 공간적인 제약

의 발생으로 인해 대중은 비대면 활동이 증가하였고 인터넷 사용량이 증가함에 따라 온라인을 통해 많은 감정과 생각을 표현할 수 있게 되었다. 걱정이나 두려움, 좌절, 불안과 같은 많은 정신적 문제들을 포함한 이야기를 할 수 있으며 이는 코로나 바이러스의 유행 기간 동안 오랜 사회 활동 제한으로 인해 사람들에게 정신 건강 문제를 반영할 수 있다. Hui Yin [14] 연구진은 2020년 4월 1일부터 2주 동안 트위터에서 수집된 1,300만 개 이상의 트윗에 대해 글로벌 수준에서 감성 분석을 실시하였다. 분석 결과는 사람들이 최근 확인된 코로나 바이러스 사례, 감염 예방 조치, 대유행 사태에 대한 정부의 태도 및 구체적인 조치에 대해 우려하는 것으로 나타났다. 전체적인 감성 극성은 긍정적이었지만 주제마다 주제 감정 극성은 다양했다. 국내에서는 코로나19와 관련된 문서를 대상으로 빈도 분석, 토픽 모델링 분석, 의미 연결망 분석, 감성 분석 등 다양한 연구 분석 방법이 사용되었다.[15-18] 김태종 연구팀[15]은 2019년 12월부터 2020년 3월까지 보도된 47,816건의 코로나 19 관련 뉴스 데이터로 토픽 모델링 분석을 실시하여 20개의 토픽을 도출하였고, 주요 의제에 대해 중립적인 언론 보도의 방향성을 제안하였다. 김상미 연구팀[16]은 코로나19와 온라인 교육에 관련된 국내 언론 기사 602건을 대상으로 토픽 모델링 분석을 실시하여 온라인, 디지털, 원격 수업이 핵심 이슈로 부각되었음을 보고하였다. 유소연 연구팀[17]은 코로나 19 감정과 관련 있는 뉴스 데이터를 수집하였고 동시 출현 단어 분석과 의미 네트워크 분석을 사용한 텍스트 마이닝 연구를 실시하였다. ‘심리’와 ‘코로나19’, ‘불안’, ‘블루’라는 단어가 높은 연결 중심성을 가지는 것을 통해 심리적인 방역의 필요성을 강조하였다. 손민성 연구팀[18]은 검색 포털 (네이버, 다음)의 코로나 발생 이전 2개월과 발생 이후 5개월간 텍스트 데이터를 대상으로 텍스트 마이닝 방법 중, 키워드 빈도 분석, 감성 분석, 네트워크 분석을 사용하였다. 시기별로 코로나 발생 이전에는 중립적인 글이 대부분이었으나 코로나 발생 이후 온라인 교육에 대한 긍정적 또는 부정적 의견이 논의되기 시작하였다고 보고하였다.

선행 연구의 한계

선행 연구[13,14]의 데이터로 많이 사용되어 왔던 대표적인 소셜 미디어인 트위터는 글자 수가 140자로 제한되어 있어 충분한 감정 또는 의사 표현을 할 수 없고, 지나친 개방성으로 인한 광고성 게시물과 다수의 중복 데이터 발생으로 데이터의 신뢰성을 떨어뜨릴 수 있다는 단점이 있다. 국내 대부분의 선행 연구[15-17]는 감정 표현이 제한적인 뉴스 데이터를 자료원으로 사용하여 감성 분석이 어렵다는 단점이 있다. 또한, 1년 이하의 짧은 조사 기간으로 감염병에 대한 계절성을 반영한 추이를 파악하는 데 어려움이 있을 수 있다.[12-18]

소셜 미디어 데이터 선정

네이버 지식IN은 네이버 사용자 간 지식 교류 서비스로 사용자가 올린 질문이나 궁금한 내용과 고민에 대해 전문가를 포함한 다른 사용자들이 자발적으로 답변을 올리면서 지식을 주고받는 국내 대표적인 지식 공유 포털이다. 네이버에 가입하지 않은 사용자를 포함하여 누구나 게시글을 열람할 수 있고 질문자의 신상이 공개되지 않기 때문에 질문자는 비교적 자유롭게 민감한 정보를 포함한 내용을 게시할 수 있다는 장점이 있다. 네이버 지식IN 서비스는 전문가의 답변이 존재하는 문서를 선별하여 신뢰할 수 있는 데이터를 추출할 수 있다는 장점이 있다. 또한, 사용자가 게시하는 자유로운 형식의 질문을 글자 수의 제한 없이 수집하는 것이 가능하기 때문에 문서 내에서 충분한 분량의 감정이나 의견을 여과 없이 분석할 수 있다.

본 연구는 코로나19가 발생한 시점으로부터 1년 3개월간 대표적인 국내 지식 포털 사이트 네이버 지식IN에 게시된 텍스트 데이터를 대상으로 토픽 모델링, 동시 출현 단어 분석, 그리고 감성 분석을 모두 포함한 국내 최초의 텍스트 마이닝 연구이다. 토픽 모델링 분석

방법에서 잠재 디리클레 할당 (Latent Dirichlet Allocation, LDA) 모델의 핵심적인 요소는 하나의 문서에서 다수의 토픽이 존재한다는 가정 하에 연관된 단어들 사이에서 동시에 출현할 확률을 분석하여 잠재적인 토픽을 찾아내는 것이다. 본 연구에서 수집한 데이터는 2만 건이 넘는 텍스트 문서로 알고리즘을 통한 자동 분석 기반의 LDA 모델을 사용한 토픽 모델링 방법을 활용하고자 했다. 코로나19 관련 이슈 탐색과 함께 대중이 직접 느끼고 있는 감정이나 태도를 소셜 미디어를 통해 파악하는데 의의가 있으며, 장기간에 걸쳐 수집된 데이터를 바탕으로 도출된 연구 결과는 코로나19에 대한 보건 의료학적 문제 진단 및 대응 방안 개선에 대해 고려해야 할 시사점을 제공할 수 있을 것이다.

제 2 절 연구의 목적

본 연구의 목적은 일반 대중들이 지식 공유 포털에 게시한 ‘코로나 바이러스’를 주제로 한 질문들을 텍스트마이닝 기법을 사용하여 분석함으로써 ‘코로나 바이러스’와 관련하여 제기된 보건 의료 및 우리 사회 전반의 이슈를 파악하기 위함이다.

• 텍스트 마이닝 기법을 통한 ‘코로나 바이러스’에 대한 탐색적 연구로 단어의 빈도를 측정하여 특정 키워드의 중요도를 알아 보고자 한다.
• 토픽 모델링 분석을 통해 ‘코로나 바이러스’의 특정 키워드가 어떤 주제를 내포하고 있으며 분류된 각 주제별로 얼마나 많은 비중을 차지하고 있는 지 파악하고자 한다.
• 문서 전체의 감성 분석을 바탕으로 감염병을 향한 대중의 감정을 알아보고 시간의 흐름에 따른 코로나 바이러스에 대응하는 감정 변화를 파악하기 위함이다.

제 2 장 방 법

제 1 절 데이터 수집

웹 크롤링 기법을 사용하여 네이버 지식인에서 ‘코로나’로 검색 되어지는 질문과 답변 문서들을 수집하였다. ‘코로나19’ 키워드가 공식 명칭이지만 ‘코로나’ 키워드를 선정한 이유는 ‘코로나19’는 바이러스의 발생 초기부터 사용된 단어가 아니라서 코로나19의 발생 초기 기간은 검색이 되지 않는다. 또한, ‘코로나’는 대중이 쉽게 접할 수 있고 널리 사용되는 말이므로 검색에 쉽게 노출되며 코로나19와 관련된 광범위한 의미를 포괄하고 있기 때문에 검색어 키워드로 선정하였다. 광고성 글을 포함하거나 코로나19와 관련이 없는 문서들이 존재하므로 신뢰성이 있는 결과를 얻기 위해서 일정 자격이 요구되는 분야별 전문가(의사, 노무사, 변호사 등)의 답변이 포함된 질문을 선별하여 수집하였다. 코로나19의 발생 이후 1년 3개월 동안 (2020년 1월1일 ~ 2021년 3월 14일) 총 23,463건의 문서를 수집하였고 Python의 라이브러리인 Selenium을 사용하였다. 수집된 데이터의 속성으로는 네이버 지식인에 등록된 질문의 제목, 질문 내용, 질문 작성일과 전문가 답변 내용 및 답변 작성일이 있다.

제 2 절 데이터 전처리

분석의 정확성을 높이기 위해 수집한 텍스트 데이터에서 분석에 불필요한 특수 기호 또는 공백을 제거하여 데이터 정제를 실시하였다. 형태소 분석은 어절 또는 문장 단위의 문서에서 최소 의미 단위인 형태소로 추출하여 분석하는 방법이다. Python의 라이브러리에 등록된 KoNLPy 패키지의 대표적인 형태소 분석기인 Mecab을 사용하여 형태소 분석을 수행하였다. KoNLPy 패키지는 한국어 정보처리 패키지로 오픈소스로 공개되었으며, Mecab은 가장 최근에 개발된 형태소 분석기이다.[19] 형태소에서 의미 있는 단어와 감정을 나타내는 형용사를 추출하기 위해 Mecab 형태소 분석기 내부의 품사 태그인 일반명사(NNG), 고유명사(NNP), 동사(VV), 형용사(VA)에 해당하는 단어를 추출하였다. 단어가 한 글자일 때 의미 해석이 모호한 경우가 많아서 제외하였다. 단어 빈도 분석은 '코로나' 와 관련이 있는 키워드를 알아보기 위해 수집한 단어들의 단순 출현 빈도를 산출하는 분석 방법이다. 시간의 흐름에 따라 상위로 노출된 키워드를 파악하기 위하여 월별 단어 빈도를 산출하였다. Python을 이용하여 단어 빈도에 대한 워드클라우드 분석을 실시하였다.

제 3 절 토픽 모델링 분석

토픽 모델링은 방대한 단어 또는 문서 집합에서 숨겨진 주제를 발견하기 위한 텍스트 마이닝 방법 중 하나이다. 토픽 모델링은 문서로부터 핵심적인 이슈를 파악하고 시간의 흐름에 대한 변화를 이해하기 위해 주로 사용된다.

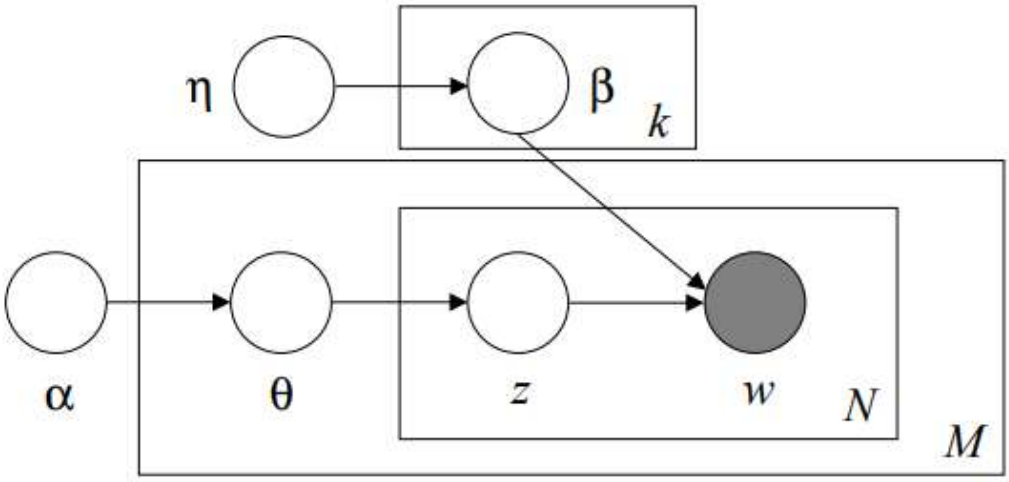
Deerwester 연구팀[20]은 문서-단어 행렬에서 문서-토픽 행렬과 토픽-단어 행렬의 개념으로 나누어 주제를 분류할 수 있는 최초의 토픽 모델링인 잠재 의미 분석(Latent semantic indexing, LSI)을 소개하였다.

Hofmann 연구팀[21]은 LSI를 발전시켜 확률 모형을 적용해 확률 잠재 의미 분석(Probabilistic latent semantic indexing, pLSI) 모델을 소개하였다. pLSI 방법은 문서 내 특정 용어가 등장할 확률을 기반으로 구축된 모델로 LSI 모델에서 진화했지만 문서 내에서 주제가 어떤 분포를 가지고 있는 지 파악할 수 없다는 한계를 가지고 있다.

이후 Blei 연구팀[22]이 LDA 모델을 발표하였는데 확률 분포로 디리클레 분포(Dirichlet Distribution)를 적용하여 pLSI 모델의 한계점을 극복하여 문서 내에 주제의 분포를 파악할 수 있으며 현재 많은 연구에 활용되고 있는 토픽 모델링 방법 중 하나이다.

본 연구의 데이터 분석은 토픽 모델링 방법 중 대표적으로 활용되고 있는 잠재 디리클레 할당 (Latent Dirichlet Allocation, LDA) 모델을 사용하였다. LDA는 말뭉치를 생성하는 확률 기반의 모델이다. 여기서 말뭉치란 연구를 위한 목적으로 표본을 추출한 텍스트의 집합을 의미하고 본 연구에서는 형태소 집합이 말뭉치로 사용되었다.

LDA 모델의 구조를 살펴 보면 [그림 1]과 같다.



[그림 1] Latent Dirichlet Allocation 모델의 구조 도식화

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \tag{1}$$

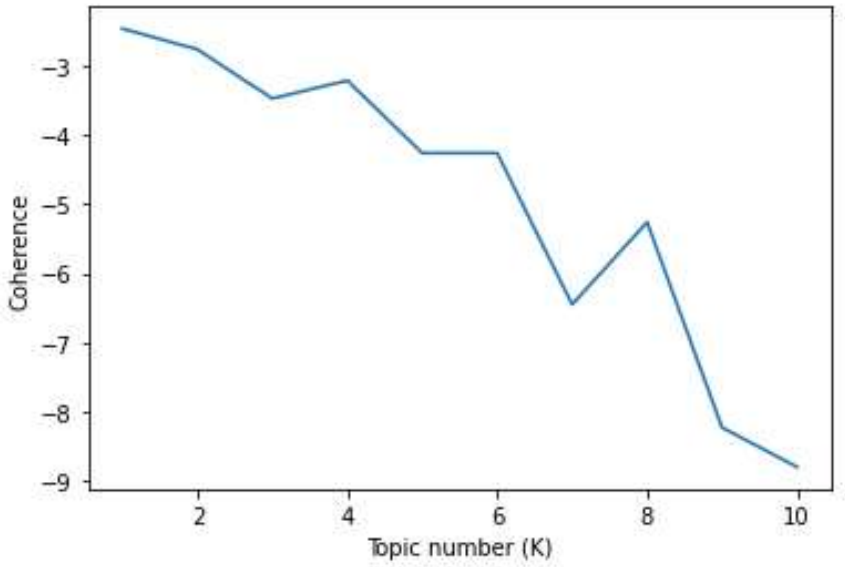
$$p(w \mid \alpha, \beta) = \frac{\Gamma(\Sigma_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \int (\prod_{i=1}^{k} \theta_i^{\alpha_i - 1})(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}) d\theta \tag{2}$$

M : 문서의 개수
N : 문서 내 단어의 개수
w : 단어
z : 특정 단어가 속한 토픽
θ : 문서 내 토픽 분포
α : 문서-토픽 θ 분포 파라미터
β : 토픽 내 단어 분포
η : 토픽-단어 β 분포 파라미터

[그림 1]의 LDA 모델에서 연구자가 직접 입력할 수 있는 값은 하이퍼 파라미터 값인 α, η 그리고 토픽의 개수 k이다. α값을 입력하여 문서 내 토픽 분포를 결정하고 η 값을 입력하여 토픽 내 단어 분포를 결정한 후 토픽의 개수 k를 입력하면 k 개의 주제로 문서가 자동으로 분류된다. 문서 내 주제 분포와 주제 내 단어 분포는 계산식(1)에 의해 문서 내 특정 단어들 w는 각각 어떠한 주제 z에 할당된다.
특정 단어가 특정 주제에 수렴할 때까지 반복 할당 되는 알고리즘(2)을 통해 최종 주제를 선정하는 구조를 가지고 있다.
LDA 모델에서 고려할 요소는 연구자가 토픽의 개수를 설정해야 하는 것이다.

$$Coherence = \sum_{i<j} score(w_i, w_j)$$

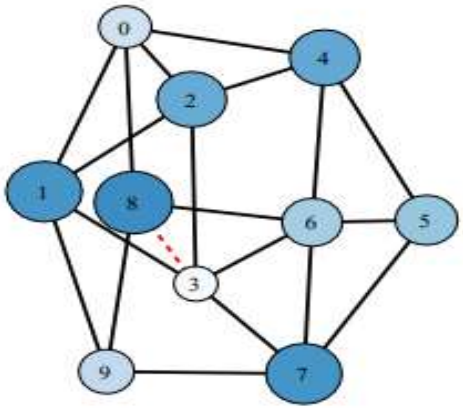
일반적으로 LDA 모델은 토픽의 개수 별로 모델의 Coherence (일관성)를 측정한다. Coherence는 토픽의 각 주제와 관련이 높은 상위 빈도에 해당하는 단어들 간 유사도를 계산하고 각각의 유사도를 모두 합하는 방식으로 측정된다.[23] Coherence가 높을수록 의미 있는 결과를 나타내는 모델이라고 판단하며 최적의 토픽 수 K를 설정한다.[24] 토픽의 개수 K를 1부터 20까지 증가 시키며 Coherence를 측정한 결과, K=8 일 때 Coherence가 높아 토픽의 개수를 8개로 선정하였다. [그림 2]



[그림 2] LDA 모델의 토픽별 Coherence 측정값

제 4 절 동시 출현 단어 분석

동시 출현 단어 분석은 두 단어 간의 연관성의 정도를 파악하는 텍스트 마이닝 분석 방법이다. 단어의 동시 출현은 한 문장에서 두 단어가 동시에 나타난 횟수를 말하며 출현한 빈도가 잦을수록 두 단어의 연관성이 깊다. 문서-단어 행렬은 다수의 문서에 등장하는 단어들의 빈도를 행렬로 표현한 것이다. 문서-단어 행렬의 전치 행렬과 문서-단어 행렬의 내적을 구하면 단어-단어 행렬을 구할 수 있는데, 단어-단어 행렬로 동시 출현 단어 빈도를 구할 수 있다. 단어 네트워크 분석은 문서에 출현하는 단어와 단어 사이의 관계를 선으로 표시함으로써 네트워크를 구축하고 시각적으로 나타내는 분석 방법이다. 코로나 관련 문서에 등장하는 중요한 단어들이 서로 어떤 관계가 있는 지 구체적으로 알아보기 위해 출현 빈도가 높은 키워드를 대상으로 단어-단어 행렬을 만들고 네트워크 분석을 실시하였다. 네트워크 분석을 위한 시각화 도구로 Python 라이브러리인 NetworkX를 활용하였다.[25] 단어의 빈도가 높을수록 원 모양의 노드(Node) 크기가 커지고 단어와 단어의 관계에서 동시 출현 빈도가 높을수록 노드(Node)를 연결하는 선(Edge)의 굵기가 굵어진다.
[그림 3]



[그림 3] 단어 연결망 시각화 그래프

제 5 절 감성 분석

감성 분석은 저자나 글쓴이가 표현한 글에서 의견이나 태도 및 감정 등을 파악하기 위한 분석 기법이다. 사전에 정의된 감성 사전을 활용하여 저자의 긍정적 또는 부정적 태도를 파악하거나 기쁨, 슬픔, 공포 등 다양한 감정의 상태를 평가하는데 주로 사용된다.

문서 내에서 긍정과 부정을 나타내는 감정을 판단하기 위해 사전에 명시된 극성 어휘들의 빈도와 강도를 연산하여 점수로 나타내는 방법이 있다.[26] Korean Sentiment Analysis Corpus (이하, 'KOSAC')[26]는 세종 구문 분석 코퍼스의 7,774문장에 대해 17,582번의 감정 표현을 주석한 한국어 감성 사전이다. KOSAC 감성 사전의 종류로 긍정, 부정의 방향성을 나타낸 정보를 포함한 어휘 사전인 polarity가 존재한다. 극성과 강도의 계산에 의해 긍정적인 문장은 0에서 1 사이 값에 해당하고 부정적인 문장은 -1에서 0 사이 값에 해당한다. KOSAC 감성 사전 중 polarity를 사용하여 긍정, 부정 감정 지수를 측정하였다. 계산 방식은 문장 내 긍정으로 분류된 단어가 등장하면 polarity 사전에 표기된 단어의 점수를 더해주고 문장 내 부정으로 분류된 단어가 등장하면 polarity 사전에 표기된 점수를 빼 주는 형식으로 진행 된다. 예를 들어, '희망' 이라는 단어는 polarity 사전에 긍정으로 등록되어 1점으로 표기되어 있고 '희생자' 라는 단어는 polarity 사전에 부정으로 등록되어 마찬가지로 1점으로 표기되어 있을 때, "나는 코로나 바이러스가 종식되기를 희망한다" 라는 문장에서 '희망' 이 포함되고 '희생자' 는 포함되지 않으므로 +1점이 되어 긍정적인 문장으로 분류된다.

손선주[27] 등은 총 504개의 한국어 어휘로 감정 표현 단어의 목록을 구축하고 문서들의 감성 분석을 시행한 바 있으며 본 연구에서도 동일한 어휘 사전을 활용하여 '코로나' 에 대한 사람들의 태도 또는 감정 상태를 분석하였다. 감정 상태가 나타난 감정 범주로 '통증', '공포', '놀람', '기쁨', '지루함', '슬픔', '분노', '혐오', '흥미', '중성',

‘기타’를 사용했지만 본 연구에서는 일반 사회적으로 대중이 쓰는 표현에 맞게 사용하기 위해서 ‘통증’을 ‘고통’으로 변경 하였다. ‘공포’는 ‘불안’으로 변경하였고 ‘놀람’은 ‘당혹’으로 ‘기쁨’은 ‘만족’으로 ‘슬픔’은 ‘우울’로 ‘분노’는 ‘화’로 ‘흥미’는 ‘관심’으로 각각 변경한 후 적용하였다. 또한, 기존의 감정 범주인 ‘중성’은 감정 표현을 나타내기 적합하지 않기 때문에 삭제하고 ‘걱정’이라는 감정 범주를 새롭게 추가하였다. <표 1> 수집한 ‘코로나’관련 문서에서 기존의 감성 사전[27]을 새롭게 변경한 감성 사전에 등장하는 어휘가 나타날 때마다 기본 감정 범주에 해당하는 감정의 빈도를 측정하였다.
예를 들어, ‘코로나 바이러스에 감염되어 고통스럽다’라는 문장이 있을 경우 ‘고통스럽다’는 기본 감정 범주인 고통에 해당한다. 이 경우 고통에 대한 빈도는 1 증가한다. 기본 감정 범주의 빈도에 대한 총합을 구하고 10개의 감정 중 어떤 감정에 해당했으며 얼마나 비중이 높았는지에 대해 관찰하였다.

<표 1> 감정 범주와 감정 단어

감정 범주	감정 단어	감정 범주	감정 단어
우울	가련하다	만족	가뿐하다
	고달프다		감동하다
	고독하다		반갑다
	낙심하다		사랑스럽다
당혹	갑작스럽다	불안	겁나다
	당혹하다		긴장되다
	이상하다		불안하다
	충격적		소름끼치다
지루함	갑갑하다	화	기분 나쁘다
	답답하다		불만스럽다
	싫증나다		어이없다
	지긋지긋하다		짜증나다
고통	고통스럽다	혐오	꺼림칙하다
	골치아프다		꼴불견
	아프다		부담스럽다
	지끈거리다		징그럽다
관심	궁금하다	걱정	걱정하다
	부러워하다		근심스럽다
	신비롭다		냉가슴
	흥미롭다		노파심

제 3 장 결 과

제 1 절 주요 키워드의 빈도

2020년 1월 1일부터 2021년 3월 14일까지 네이버 지식인에서 전문가의 답변이 있는 질문을 대상으로 ‘코로나’ 검색을 통해 수집한 글은 23,463건으로 관찰되었다. 이 중 단어 검색 결과의 상위 25개의 키워드와 빈도를 내림차순으로 나열하였다. <표 2>

‘코로나’ 단어가 22,281회로 가장 많이 나타났으며, ‘병원’은 6,307회로 두 번째로 많이 등장하였다. 이 외에도 ‘증상’ 6,237회, ‘급여’ 6,136회, ‘근무’ 5,560회, ‘회사’ 5,477회, ‘검사’ 4,147회 등으로 나타났다.

전문가의 직업으로는 ‘의사’가 11,676건에 대해 가장 많은 답변을 하였고 ‘노무사’ 5,242건, ‘변호사’ 3,888건, ‘운동전문가’ 1,143건 순으로 나타났다. <표 3>

텍스트 데이터 분석 결과에서 두드러지게 강조되는 키워드를 분석하기 위해 상위 빈도 100개의 단어를 워드 클라우드로 시각화하였다.

[그림 3]

<표 2> 지식 포털 사이트 내 ‘코로나’ 관련 상위 25개 단어 검색 결과

키워드	빈도
코로나	22,281
병원	6,307
증상	6,237
급여	6,136
근무	5,560
회사	5,477
검사	4,147
실업	3,615
아프다	3,347
기침	3,344
생각	3,159
걱정	3,141
계약	3,103
퇴사	2,910
보험	2,757
통증	2,657
퇴직금	2,536
월급	2,523
운동	2,296
머리	2,215
연락	2,070
수술	2,000
치료	1,966
지급	1,914
수당	1,824

<표 3> 지식 포털 사이트 내 ‘코로나’ 관련 문서의 전문가 답변 수

전문가	답변
의사	11,676
노무사	5,242
변호사	3,888
운동전문가	1,143
세무사	276
약사	274
회계사	197
법무사	183
관세사	96
손해사정사	49
변리사	42
영양사	41

[그림 4] 지식 포털 사이트 내 ‘코로나’ 관련 상위 빈도 100개 워드 클라우드

제 2 절 토픽 모델링 분석 결과

‘코로나’ 관련 글에 대한 토픽 모델링 분석 결과 3개의 테마와 8개의 주제로 분류 되었다. <표 4>, [그림 5]

첫 번째 테마는 사회 분야로 분류하였고 Topic 1, Topic 2, Topic 5, Topic 6가 포함되었다. 두 번째 테마는 보건 의료이며 Topic 3, Topic 4, Topic 7이 이에 해당한다. 세 번째 테마는 법률로 명명하였고 Topic 8로 구성되었다.

첫 번째 테마인 ‘사회’ 의 토픽별 주요 키워드를 살펴 보면 Topic 1은 전체 문서의 24.7%로 가장 높은 비중을 차지했고 ‘급여’, ‘회사’, ‘근무’, ‘실업’, ‘퇴사’, ‘보험’, ‘월급’, ‘지급’, ‘고용’, ‘직원’ 의 단어가 키워드로 나타났다. 이를 바탕으로 Topic 1은 ‘고용 불안정’ 에 대한 주제라고 판단하였다.

Topic 2는 ‘연락’, ‘전화’, ‘계약’, ‘소송’, ‘대출’, ‘월세’, ‘보증금’, ‘합의’, ‘본인’, ‘주인’ 이 주요 단어로 나타났다. 전체에서 18.7%를 차지해 두 번째로 높은 비중을 차지했고 ‘주요 계약 건의 비대면 접촉’ 이라는 주제를 선정하였다.

Topic 5는 전체 문서의 9.1%를 차지하였으며 ‘한국’, ‘미국’, ‘비자’, ‘격리’, ‘입국’, ‘해외’, ‘발급’, ‘활동’ 이 주요 단어로 나타났다. 이러한 결과를 바탕으로 Topic 5는 ‘해외 출입국 절차 강화’ 라는 주제를 생각해 볼 수 있다.

Topic 6은 6.8%의 비중을 차지하며 ‘수당’, ‘연차’, ‘해고’, ‘출근’, ‘사장’, ‘근로계약서’, ‘주휴’, ‘퇴근’, ‘시급’, ‘알바’ 가 포함되었다. 위 단어들의 조합으로 미루어 볼 때, ‘근무 조건 변화’ 라는 주제로 분류하였다.

‘보건 의료’ 로 정한 두 번째 테마는 Topic 3, Topic 4 그리고 Topic 7으로 구성되었다.

Topic 3은 전체 문서의 15.3%의 비중을 차지하였으며 ‘증상’, ‘기침’, ‘머리’, ‘마스크’, ‘가래’, ‘감기’, ‘콧물’, ‘근육통’, ‘미

열’, ‘몸살’ 등이 포함되었다. Topic 3의 주제는 ‘코로나 의심 증상’으로 명명하였다.

Topic 4는 전체 비중의 13.9%를 차지하였고 ‘병원’, ‘수술’, ‘의사’, ‘신경’, ‘선생’, ‘스트레스’, ‘진료’, ‘처방’, ‘치료’가 주요 키워드로 나타났다. ‘병원 방문 우려’라는 주제로 선정하였다.

Topic 7은 6.5%의 비중으로 ‘운동’, ‘허리’, ‘다리’, ‘어깨’, ‘근육’, ‘손가락’, ‘무릎’, ‘자세’, ‘디스크’, ‘발목’ 키워드를 포함하였다. 결과를 종합했을 때 ‘운동’이라는 주제로 분류하였다.

세 번째 테마인 ‘법률’의 토픽은 전체 문서의 5.2%를 차지하며 Topic 8에 해당한다.

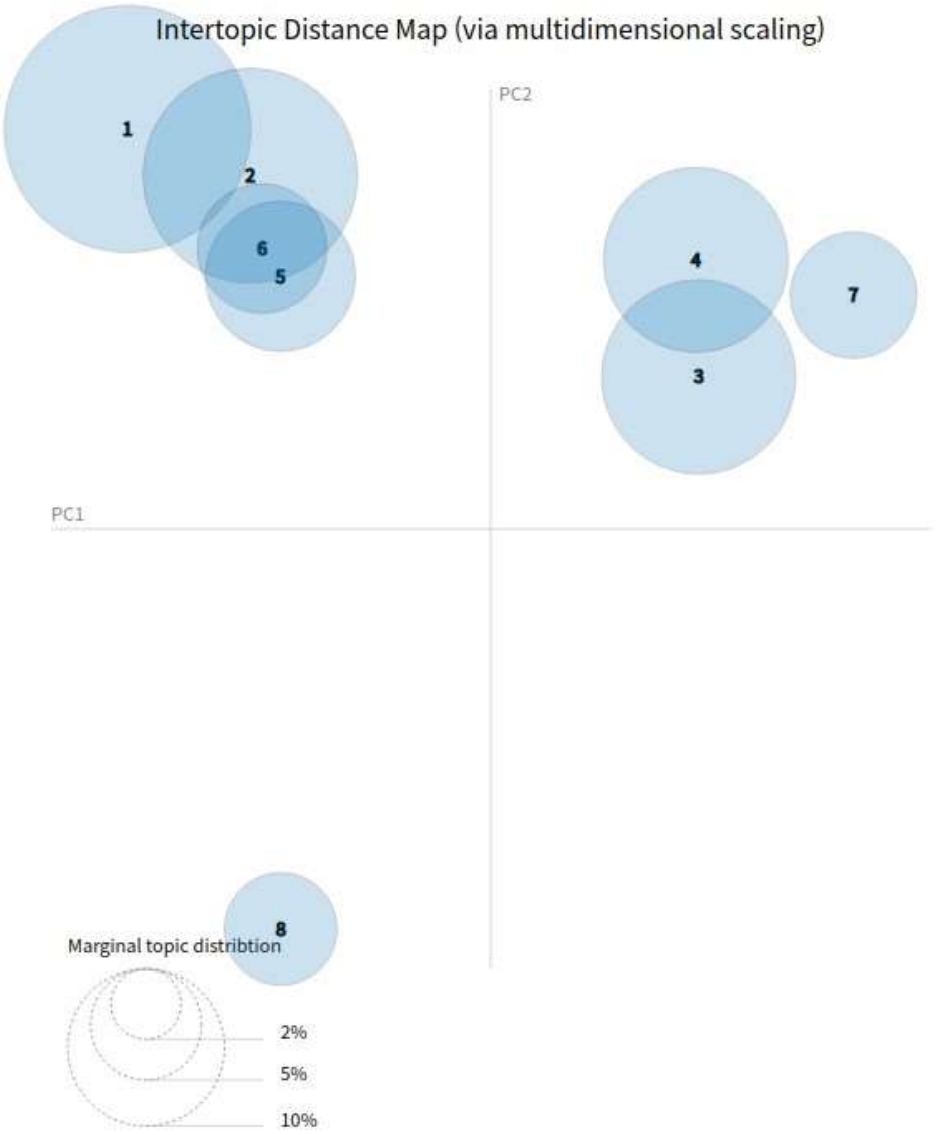
Topic 8은 ‘재판’, ‘조사’, ‘감염’, ‘사건’, ‘벌금’, ‘피해자’, ‘헬스’, ‘확률’, ‘무증상’, ‘구속’ 등으로 구성되어 있으며 ‘행정 처분’이라는 주제로 선정하였다.

토픽별 전문가 답변의 비율을 조사한 결과, 의사가 가장 많은 비중을 차지한 직업으로 나타났다.

‘사회’ 테마로 분류된 Topic 1은 노무사 답변이 가장 많았고 Topic 1의 78.2%를 차지하였다. Topic 2는 변호사 답변이 가장 많았으며 Topic 2의 55.8%를 차지하였다. Topic 5의 전문가 답변은 의사가 가장 많았으며 Topic 5의 38.1%를 차지하였다. Topic 6의 전문가 답변 또한 의사가 30.5%로 가장 많은 답변을 한 것으로 나타났다.

‘보건 의료’로 분류된 테마는 모두 의사 답변이 가장 많은 것으로 나타났고 Topic 3은 의사 답변이 86.1%를 차지하였다. Topic 4, Topic 7은 의사 답변으로 각각 67.3%, 73.2%를 차지하였다.

‘법률’로 분류된 테마인 Topic 8 역시 의사 답변 수가 가장 많이 등장하였고 62.4%의 비율을 나타냈다. <표 5>



[그림 5] 지식 포털 내 ‘코로나’ 관련 토픽 모델링 결과

<표 4> ‘코로나’ 관련 토픽 모델링 주제 및 키워드 결과

구분	토픽(%)	주제	키워드
사회	Topic 1(24.7%)	고용 불안정	급여, 회사, 근무, 실업, 퇴사, 보험, 월급, 지급, 고용, 직원
	Topic 2(18.7%)	비대면 계약	연락, 전화, 계약, 소송, 대출, 월세, 보증금, 합의, 본인, 주인
	Topic 5(9.1%)	해외 출입국 절차 강화	한국, 미국, 비자, 격리, 입국, 해외, 발급, 활동, 접종, 도착
	Topic 6(6.8%)	근무 조건 변화	수당, 연차, 해고, 출근, 사장, 근로계약서, 주휴, 퇴근, 시급, 알바
보건의료	Topic 3(15.3%)	코로나 의심 증상	증상, 기침, 머리, 마스크, 가래, 감기, 콧물, 근육통, 미열, 몸살
	Topic 4(13.9%)	병원 방문 우려	병원, 수술, 의사, 신경, 선생, 스트레스, 진료, 처방, 치료, 입원
	Topic 7(6.5%)	운동	운동, 허리, 다리, 어깨, 근육, 손가락, 무릎, 자세, 디스크, 발목
법률	Topic 8(5.2%)	행정 처분	재판, 조사, 감염, 사건, 벌금, 피해자, 헬스, 확률, 무증상, 구속

<표 5> ‘코로나’ 관련 토픽별 전문가 답변 비율 (%)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
의사	11.8	16.0	86.1	67.3	38.1	30.5	73.2	62.4
노무사	78.2	12.2	5.4	6.9	6.9	19.7	6.0	7.1
변호사	5.6	55.8	3.7	9.9	35.4	20.8	6.8	12.3
운동전문가	1.3	1.8	3.0	4.8	10.0	2.9	10.9	6.5
세무사	1.4	4.7	0.2	0.6	1.0	1.0	0.3	0.7
약사	0.3	0.5	2.0	1.4	1.2	1.0	1.3	1.7
회계사	0.6	1.9	0.2	0.5	0.5	0.7	0.2	0.5
법무사	0.5	2.1	0.3	0.8	0.9	1.1	0.4	0.9
관세사	0.1	1.2	0.1	0.2	1.1	0.4	0.2	0.5
손해사정사	0.1	0.4	0.1	0.5	0.2	0.3	0.1	0.3
변리사	0.1	0.4	0.0	0.1	0.2	0.2	0.3	0.1
영양사	0.1	0.1	0.1	0.2	0.2	0.1	0.5	0.1

Topic 1 : 고용 불안정, Topic 2 : 비대면 계약, Topic 3 : 코로나 의심 증상, Topic 4 : 병원 방문 우려, Topic 5 : 해외 출입국 절차 강화, Topic 6 : 근무 조건 변화, Topic 7 : 운동, Topic 8 : 행정 처분.

제 3 절 동시 출현 단어

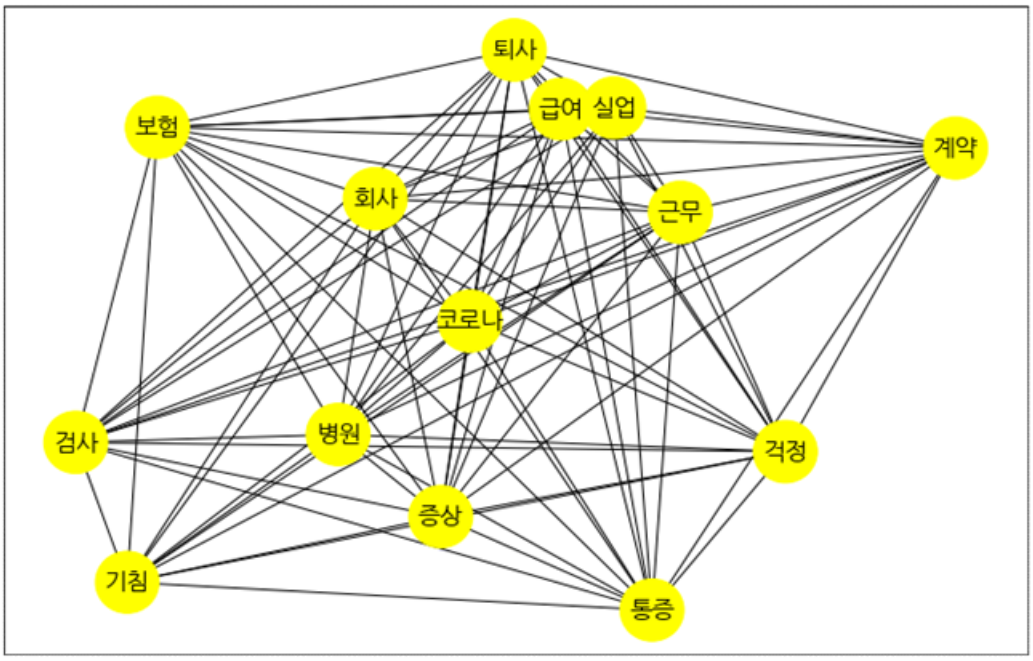
‘코로나’ 관련 문서에서 가장 빈도가 높았던 키워드 중 상위 15개 단어에서 동사인 ‘아프다’를 제외한 후, 상위 14개의 키워드를 대상으로 동시 출현 단어 분석을 실시하였다. <표 6>

‘코로나’ 키워드와 동시 출현 빈도가 가장 높은 단어는 ‘증상’이고 8,216회 등장하였다. ‘급여’ 7,156회, ‘병원’ 6,855회, ‘근무’ 6,601회, ‘검사’ 5,461회 등으로 나타났다.

상위 14개의 키워드 중 유일하게 감정을 표현하는 단어인 ‘걱정’과 동시 출현 빈도가 가장 높은 단어는 ‘코로나’가 3,356회 등장했고 ‘병원’ 1,394회, ‘증상’ 1,310회, ‘검사’ 862회, ‘기침’ 731회 순이었다. [그림 5]은 동시 출현하는 단어를 연결 그래프로 나타낸 것이다.

<표 6> 지식 포털 사이트 내 ‘코로나’ 관련 상위 14개 단어-단어 분석 결과 (단위 : 건)

	코로나	실업	병원	증상	급여	기침	퇴사	근무	회사	검사	걱정	계약	통증	보험
코로나	32,779	4,164	6,855	8,216	7,156	3,266	3,366	6,601	6,280	5,461	3,356	3,529	2,578	3,077
실업	-	10,757	212	22	11,880	0	3,777	4,263	4,213	157	124	1,745	20	3,055
병원	-	-	17,867	3,504	429	1,334	261	740	815	3,731	1,394	83	3,096	648
증상	-	-	-	13,551	34	2,821	39	101	275	2,702	1,310	10	2,225	56
급여	-	-	-	-	23,366	1	5,933	8,273	7,336	177	207	2,384	54	4,745
기침	-	-	-	-	-	6,090	7	54	43	947	731	7	580	2
퇴사	-	-	-	-	-	-	8,248	3,969	4,696	60	118	1,206	20	1,566
근무	-	-	-	-	-	-	-	20,250	5,390	212	206	2,374	123	3,235
회사	-	-	-	-	-	-	-	-	20,495	642	484	2,119	127	2,607
검사	-	-	-	-	-	-	-	-	-	10,677	862	81	911	199
걱정	-	-	-	-	-	-	-	-	-	-	4,165	223	671	186
계약	-	-	-	-	-	-	-	-	-	-	-	14,199	8	1,077
통증	-	-	-	-	-	-	-	-	-	-	-	-	11,235	100
보험	-	-	-	-	-	-	-	-	-	-	-	-	-	8,537



[그림 6] ‘코로나’ 관련 키워드의 단어 네트워크 분석 결과

토픽 모델링 분석에서 ‘보건의료’로 분류된 테마 내 많은 비중을 차지했던 신체 부위에 해당하는 키워드 중에서 연결된 의미를 추가적으로 파악하기 위해 네트워크 분석을 통해 동시 출현 단어를 분석하였다.

<표 7>, [그림 7]

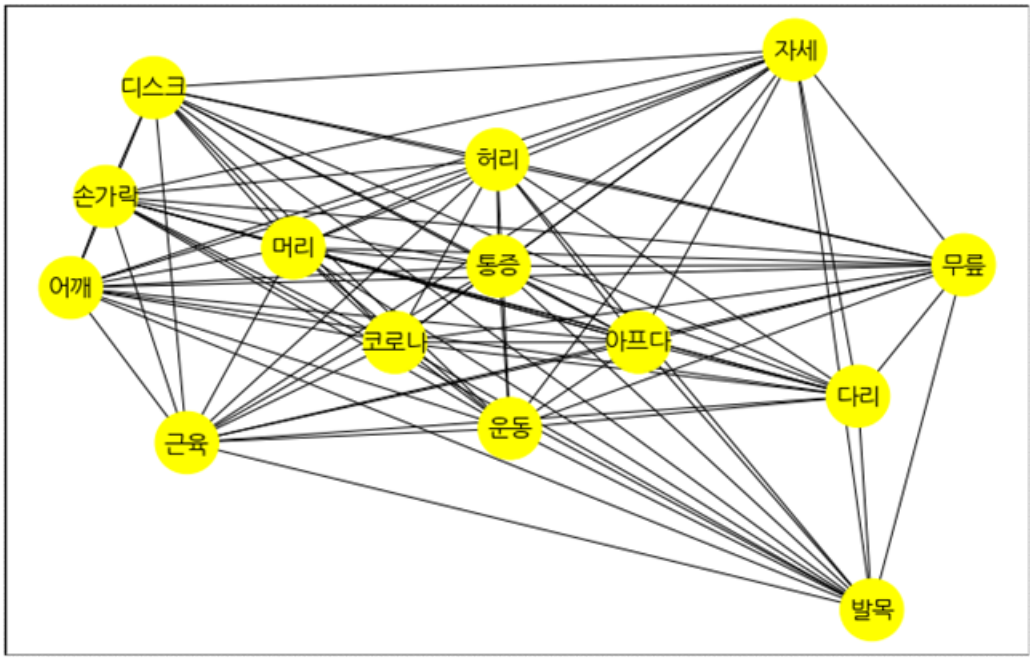
단어 빈도 분석에서 추출했던 주요 키워드에서 신체 부위와 짝을 이룰 수 있는 키워드인 ‘코로나’, ’아프다’, ’통증’ 단어를 추가하여 동시 출현하는 단어의 빈도를 측정하였다. 또한, Topic 7에서 신체 부위와 함께 포함되었던 ‘운동’, ‘자세’ 역시 신체 부위와의 연관성을 알아보기 위해 키워드로 추가하였다. ‘코로나’와 가장 많이 동시 출현하는 단어 중 신체 부위에 해당하는 단어는 ‘머리’가 2,154회로 가장 많았으며 두 번째로 많이 출현한 ‘허리’는 699회 등장하였다. ‘손가락’ 591회, ‘다리’ 541회, ‘어깨’ 525회 등으로 나타났다.

‘아프다’와 동시 출현하는 단어는 ‘머리’가 1,633회로 가장 많이 등장하였고 ‘허리’ 512회, ‘다리’ 321회, ‘무릎’ 316회, ‘어깨’ 298회 순이었다.

‘통증’과 동시 출현하는 단어는 ‘허리’가 1,621회로 가장 높은 빈도를 보였으며, ‘손가락’ 1,434회, ‘머리’ 1,131회, ‘어깨’ 914회, ‘다리’ 902회 순으로 나타났다.

<표 7> 지식 포털 사이트 내 ‘운동’ 관련 단어-단어 분석 결과 (단위 : 건)

	코로나	머리	운동	허리	다리	어깨	근육	손가락	무릎	자세	디스크	발목	통증	아프다
코로나	32,779	2,154	1,527	699	541	525	345	591	352	186	264	258	2,578	3,088
머리	-	7,487	265	397	112	1,745	364	4,433	24	76	894	13	1,131	1,633
운동	-	-	6,882	562	532	662	807	127	597	230	212	233	1,488	601
허리	-	-	-	2,630	458	363	376	539	153	232	594	52	1,621	512
다리	-	-	-	-	1,594	57	246	22	328	143	98	114	902	321
어깨	-	-	-	-	-	2,568	356	2,857	32	132	603	19	914	298
근육	-	-	-	-	-	-	1,444	556	98	74	179	33	756	234
손가락	-	-	-	-	-	-	-	8,642	32	98	1,504	7	1,434	125
무릎	-	-	-	-	-	-	-	-	1,702	96	19	107	869	316
자세	-	-	-	-	-	-	-	-	-	563	67	42	464	155
디스크	-	-	-	-	-	-	-	-	-	-	826	14	674	98
발목	-	-	-	-	-	-	-	-	-	-	-	975	492	208
통증	-	-	-	-	-	-	-	-	-	-	-	-	11,235	1,643
아프다	-	-	-	-	-	-	-	-	-	-	-	-	-	5,679



[그림 7] ‘운동’ 관련 주요 키워드의 단어 네트워크 분석 결과

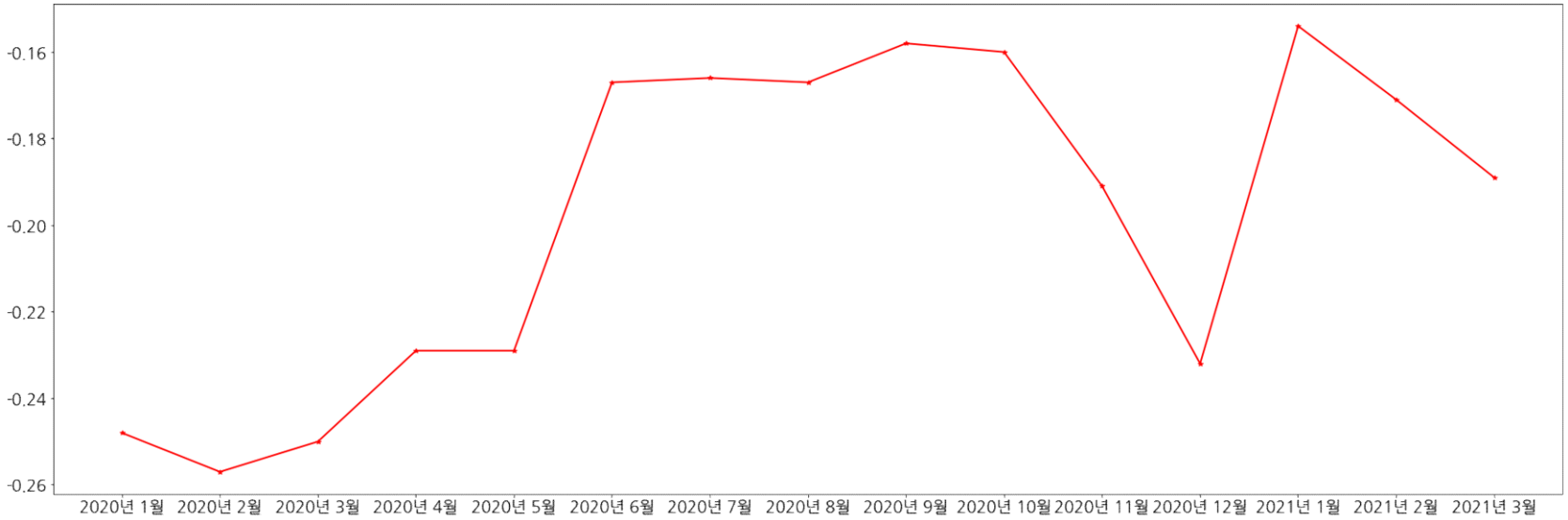
제 4 절 감성 분석 결과

‘코로나’ 관련 문서에서 전체 기간인 2020년 1월 1일부터 2021년 3월 14일까지 감성 분석을 실시하였다. KOSAC의 polarity 사전을 이용한 감정 지수 측정 결과, 전체 문서에서 -0.201로 나타나 전반적으로 부정적인 감정 및 태도의 문장들로 구성되었다. 2020년 2월에 -0.257로 가장 부정적인 감정 지수가 나타났고 점차 상승하다가 2020년 12월에 -0.232로 감정 지수가 하락하였다. <표 8>,[그림 8]

기본 감정 범주에 대한 분석에서 ‘고통’이 4,707회 등장하여 66.7%로 가장 많은 비중을 차지하였다. [그림 9] ‘불안’ 감정은 1,128회(16.0%), ‘우울’ 407회(5.8%), ‘당혹’ 293회(4.1%), ‘만족’ 248회(3.5%), ‘화’ 176회(2.5%), ‘지루함’ 64회(0.9%), ‘혐오’ 39회(0.6%) 순으로 나타났다.<표 8> 그 외 감정 범주에 해당했던 ‘걱정’, ‘관심’, ‘기타’는 한 번도 등장하지 않았다. 감정 단어로는 ‘아프다’가 3,347회로 가장 높은 빈도를 보였고, ‘아야’ 1,303회, ‘무섭다’ 961회, ‘앗’ 183회, ‘울다’ 172회 순으로 출현했다.<표 10>

<표 8> 지식 포털 내 ‘코로나’ 관련 문서에서의 감정 지수 월별 추이

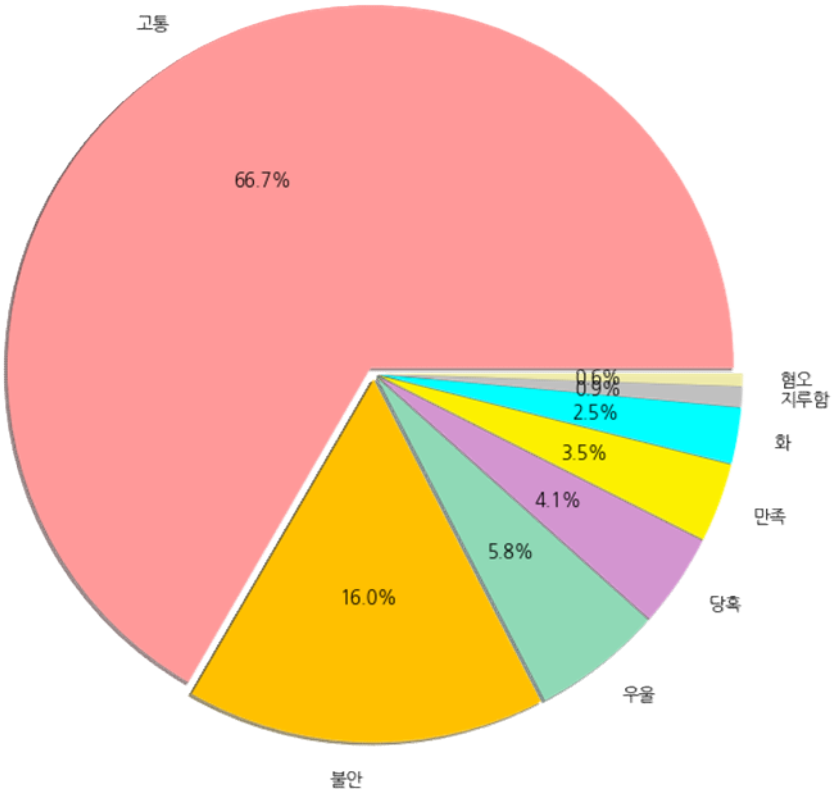
연월	감정 지수
2020년 1월	−0.248
2020년 2월	−0.257
2020년 3월	−0.25
2020년 4월	−0.229
2020년 5월	−0.229
2020년 6월	−0.167
2020년 7월	−0.166
2020년 8월	−0.167
2020년 9월	−0.158
2020년 10월	−0.16
2020년 11월	−0.191
2020년 12월	−0.232
2021년 1월	−0.154
2021년 2월	−0.171
2021년 3월	−0.189
전체	−0.201



[그림 8] 지식 포털 ‘코로나’ 관련 문서에서의 월별 감정 지수 추이

<표 9> 지식 포털 내 ‘코로나’ 관련 문서에서의 감정 범주 빈도 분석 결과 (2020.01 ~ 2021.03)

감정 범주	빈도	%
고통	4,707	66.7
불안	1,128	16.0
우울	407	5.8
당혹	293	4.1
만족	248	3.5
화	176	2.5
지루함	64	0.9
혐오	39	0.6



[그림 9] 지식 포털 ‘코로나’ 관련 문서에서의 감정 범주 분포

<표 10> 지식 포털 내 ‘코로나’ 관련 문서에서의 감정 단어 빈도 분석 결과 (2020.01 ~ 2021.03)

감정 범주	감정 단어	빈도
고통	아프다	3,347
고통	아야	1,303
불안	무섭다	961
당혹	앗	183
우울	울다	172
만족	웃다	87
불안	겁나다	84
불안	두렵다	81
화	나쁘다	79
당혹	갑작스럽다	75
지루함	귀찮다	58
고통	지끈거리다	57
만족	고맙다	49
우울	괴롭다	38
화	어이없다	36
우울	아깝다	32
우울	아쉽다	24
만족	기쁘다	23
우울	속상하다	22
당혹	깜짝	21

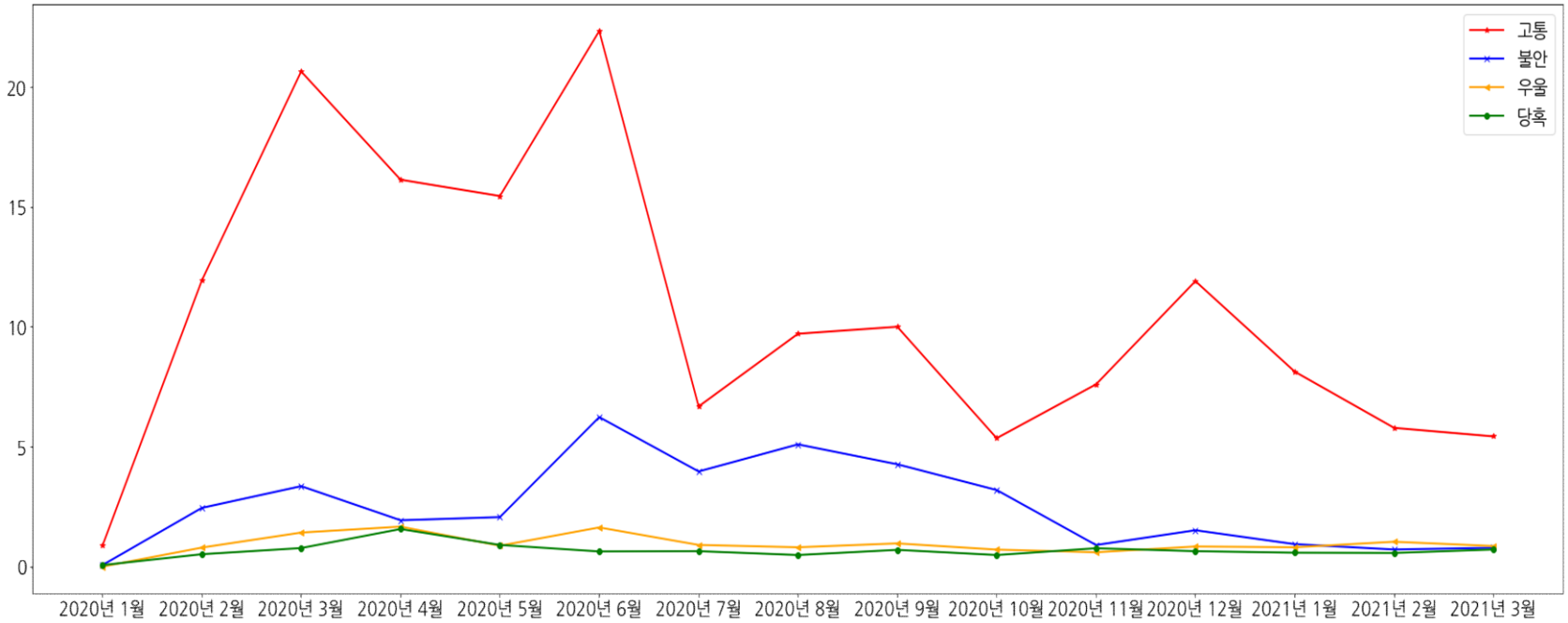
시간의 흐름에 따른 감성 분석을 위해 월별 추이를 조사하였다. 단위는 해당 월에 존재하는 날짜 수 만큼 나누어 일 평균 회수로 계산하였다. 예를 들면, 1월에는 31일을 나누고 6월에는 30일을 나누어 계산하였다. 마지막 월에 해당하는 2021년 3월은 14일까지 데이터만 존재하므로 14를 나누어 계산하였다. <표 11>

8가지 감정 범주는 구분을 위해 상위 4가지(고통, 불안, 우울, 당혹)와 [그림 10] 하위 4가지(만족, 화, 지루함, 혐오)로 나누어 그래프로 나타냈다.[그림 11]

가장 출현 빈도가 높은 감정 범주인 '고통' 을 살펴 보면, 2020년 6월에 일 평균 지수 22.33으로 가장 높았고 2020년 3월에는 20.65, 2020년 4월 16.13, 2020년 5월 15.45, 2020년 2월 11.93 순으로 나타났다. 두 번째로 출현 빈도가 높았던 '불안' 의 경우, 2020년 6월에 일 평균 지수가 6.23으로 가장 높았고 2020년 8월 5.10, 2020년 9월 4.27, 2020년 7월 3.97, 2020년 3월은 3.35 였다. 다음으로 '우울' 은 2020년 4월에 일 평균 지수가 1.67로 가장 높았고 2020년 6월 1.63, 2020년 3월 1.42, 2021년 2월 1.04, 2020년 9월 0.97 순이었다. 마지막으로 '당혹' 은 2020년 4월에 일 평균 지수가 1.57로 가장 높았고 2020년 5월 0.90, 2020년 3월과 11월 모두 0.77, 2021년 3월에 0.71 순으로 나타났다. 두 번째 그래프에서 가장 출현 빈도가 높은 감정 범주인 '만족' 은 2020년 3월에 일 평균 지수가 1.71로 가장 높았고, 2020년 6월 0.87, 2021년 1월 0.77, 2020년 9월 0.73, 2020년 5월 0.61 순이었다. 두 번째로 출현 빈도가 높은 '화' 의 경우 2020년 6월에 일 평균 지수가 0.83으로 가장 높았다. 이어 2021년 1월 0.61, 2020년 3월 0.58, 2021년 3월 0.57, 2020년 7월 0.52의 순서로 나타났다. '지루함' 은 2020년 5월에 일 평균 지수가 0.32로 가장 높았고 2020년 6월과 11월에 0.27, 2020년 3월과 2021년 1월에 0.16 순이었다. '혐오' 는 2020년 3월에 일 평균 지수가 0.26으로 가장 높았고 2020년 6월 0.23, 2020년 10월 0.16, 2021년 3월 0.14, 2020년 9월 0.13 순으로 나타났다.

<표 11> 지식 포털 내 ‘코로나’ 관련 월별 감정 범주 변화 결과 (일 평균)

시간	고통	불안	우울	당혹	만족	화	지루함	혐오
2020년 1월	0.87	0.06	0	0.06	0.03	0.00	0	0.00
2020년 2월	11.93	2.45	0.79	0.52	0.21	0.10	0.1	0.07
2020년 3월	20.65	3.35	1.42	0.77	1.71	0.58	0.16	0.26
2020년 4월	16.13	1.93	1.67	1.57	0.50	0.47	0.2	0.10
2020년 5월	15.45	2.06	0.87	0.90	0.61	0.45	0.32	0.06
2020년 6월	22.33	6.23	1.63	0.63	0.87	0.83	0.27	0.23
2020년 7월	6.68	3.97	0.9	0.65	0.26	0.52	0.06	0.00
2020년 8월	9.71	5.10	0.81	0.48	0.32	0.32	0.1	0.03
2020년 9월	10.00	4.27	0.97	0.70	0.73	0.40	0.13	0.13
2020년 10월	5.35	3.19	0.71	0.48	0.45	0.29	0.13	0.16
2020년 11월	7.60	0.90	0.6	0.77	0.50	0.27	0.27	0.07
2020년 12월	11.90	1.52	0.84	0.65	0.45	0.32	0.1	0.03
2021년 1월	8.13	0.94	0.81	0.58	0.77	0.61	0.16	0.00
2021년 2월	5.79	0.71	1.04	0.57	0.46	0.36	0.07	0.07
2021년 3월	5.43	0.79	0.86	0.71	0.57	0.57	0.07	0.14



[그림 10] 상위 4개 감정 범주(고통, 불안, 우울, 당혹)의 월별 추이 (일 평균)

[그림 11] 하위 4개 감정 범주(만족, 화, 지루함, 혐오)의 월별 추이 (일 평균)

제 4 장 고 찰

본 연구에서는 ‘코로나’ 키워드로 검색된 지식 포털의 질문 내용을 수집하고 텍스트 마이닝 방법을 사용하여 분석하였다. 토픽 모델링 분석에서 ‘코로나’ 관련 문서는 총 8개의 토픽을 3가지 주제(사회, 보건 의료, 법률)로 구분하였다. 이 중 ‘사회’ 분야는 59.3%를 차지하는 주된 이슈였고 ‘보건 의료’ 35.7%, ‘법률’ 5.2% 순으로 분류되었다.

키워드 빈도 분석 결과에서 전문가 답변 중 노무사의 답변이 5,242건으로 11,676건의 답변을 한 의사 다음으로 높은 순위를 차지하였다. 노무사의 답변이 의사보다 적었음에도 불구하고 토픽 모델링 분석에서 가장 높은 비율[Topic 1 : 고용 불안정(24.7%), Topic 6 : 근무 조건 변화(6.8%)]을 차지하였다는 것은 노무사와 관련된 키워드가 여러 차례 반복해서 언급되었다는 것을 유추할 수 있다. 이는 국내에서 고용이나 임금에 관련한 이슈가 포스트 코로나 시대에 중요한 과제임을 암시한다. 토픽 모델링 분석 결과 중 Topic 1에서 ‘급여’, ‘실업’, ‘퇴사’, ‘월급’, ‘지급’의 단어들이 나타났고 Topic 6에서 ‘수당’, ‘연차’, ‘해고’, ‘출근’, ‘근로계약서’, ‘주휴’, ‘퇴근’, ‘시급’, ‘알바’가 포함되었다. 2개의 토픽에서 고용과 임금에 대한 언급을 주로 함으로써 직장과 관련한 이슈임을 확인할 수 있다. 동시 출현 단어 분석에서도 ‘코로나’와 함께 등장한 단어로 ‘급여’, ‘근무’, ‘회사’ 등이 나타났고 최근 감염병으로 인한 불안정한 사회 속에서 직장 활동의 변화에 적응하기 위한 준비를 하려는 대중의 관심으로 해석된다. ‘급여’의 문제는 사회적으로 다양한 현상이 나타날 수 있는데 고용주가 직원에게 정기적으로 지급하는 노동의 대가 보상 이외에도 정부가 사회 보장 제도의 일환으로 국민의 기초생활보장을 위해 사회보장급여의 대상자를 상대로 지급하는 금액을 의미하기도 한다. 실제 국내 건설업 고용시장은 코로나19

로 인해 시장 충격이 발생하였다. 이와 함께 구직급여와 관련된 통계자료를 살펴보면, 코로나19가 발생한 이후 시점인 2020년도의 구직급여 신청자 및 구직급여 지급액이 2019년도에 비해 상당수 증가한 것으로 나타났다.[28] 구직급여 지급액은 2019년 6월 6,816억원에서 2020년 6월 11,103억으로 크게 증가하였다. 코로나19는 고용과 임금 문제에 있어서도 밀접한 연관이 있다. beland 연구팀[29]은 미국의 고용과 임금에 대한 코로나19의 단기 결과를 조사하였다. 그들의 연구 결과는 코로나19가 실업률을 증가시키고, 노동 시간과 노동력 참여를 감소시켰으나, 임금에는 큰 영향을 미치지 않았음을 시사했다. Lena 연구팀은 스웨덴 최대 온라인 취업 게시판에서 채용 공고에 대한 게시물의 실시간 데이터를 활용하여 코로나19에 대한 구직 반응을 측정하였다.[30] 코로나 발생 이후 3개월 전후를 비교할 때, 채용 공고가 40% 정도 줄었고 구직자들은 안정적인 직종으로 구직 활동을 하는 경향이 있다고 보고하였다. ‘근무’ 키워드와 관련하여 사회적으로 나타난 가장 큰 변화는 재택근무이다. 사회적 거리 두기가 본격적으로 시행된 이후 재택근무에 대한 수요도 높아졌으며 많은 공공기관과 기업에서 재택근무를 실시하였다.[31] 코로나19의 장기화는 근무 시스템의 대대적인 변화를 야기하였다.

토픽 모델링 결과에서 두 번째로 빈도가 높은 주제인 Topic 2는 ‘연락’, ‘전화’, ‘계약’, ‘소송’, ‘대출’, ‘월세’, ‘보증금’, ‘합의’를 포함하였다. ‘연락’, ‘전화’는 주제를 대표하는 주제 내에서 상위 빈도의 키워드에 해당한다. 이는 계약이나 합의 등이 전화 연락 등의 비대면 접촉을 통해서 이루어질 수 있다는 의미로 해석하여 ‘주요 계약에 대한 비대면 접촉’이라는 주제로 선정하였다. 전체 주제에서 18.7%에 해당하는 큰 비중을 차지하는 것을 통해서 계약과 거래와 관련하여 비대면 접촉에 대한 대중의 관심을 파악할 수 있다. Onder 등의 연구[32]에서 가장 두드러진 변화 중 하나로 사람과 사람 사이에 대면 접촉을 피하는 인공 지능 기반 비접촉식 거래 소프트웨어 프로그램이 활성화 되고 발전이 가속화 되었음을 확인하였다. 물리적 접촉에 기반한 전통적인 판

매는 적합하지 않으며 사람들은 감염으로부터 안전하지 않은 곳에서 쇼핑하지 않음을 강조하였다. 소매 부문의 기술 혁신은 정보와 디지털 기반 상거래를 위한 길을 닦았고 코로나19를 퇴치하는데 도움을 주고 있음을 확인하였다. 현재 화폐에 기반한 지불 시스템은 코로나19의 전염을 촉진하고 있으며 지폐 살균 또는 디지털 화폐와 같은 비접촉식 지불 시스템의 장려가 필요하다는 연구는 비대면 시스템에 대한 필요성을 시사한다.[33] 코로나19는 거래 시스템 뿐만 아니라 기존의 업무 체계, 소비 습관, 사회 생활에 깊은 영향을 끼쳤고 감염으로부터 보호하기 위한 방향으로 변화가 진행되고 있다. 학교에서 교사와 학생들이 교실에서 면대면 수업을 하는 것은 감염의 우려가 있기 때문에 교육부에서는 중앙재난안전본부와 협의를 하였으며, 4월 9일부터 전국 초·중·고등학교를 비롯하여 특수학교를 대상으로 하는 온라인 개학을 발표하였다.[34] 비대면 진료에 대한 사람들의 만족도가 높고 재택 근무의 생산성 역시 일정 수준 일상이 되었다. 또한, 비접촉이 일상화 되면서 사회를 변화시키며 새로운 일상으로 불리고 있다.

토픽 모델링 결과에서 사회 분야의 세 번째 비중을 나타내는 Topic 5의 주제는 출입국 절차 강화이며 '한국', '미국', '비자', '격리', '입국', '해외', '발급' 등의 출입국 관련 키워드를 포함하였다. 최근 출입국을 하는 상황에서 절차가 복잡해지는 것을 반영하는 것으로 유추할 수 있다. 2020년 4월 1일 이후 한국에 입국하는 사람들은 2주간 격리 조치가 의무화 되었다.[35] 외교 및 공무 목적의 비자 소지자 등과 사전에 '자가 격리 면제서'를 발급받은 경우는 자가 격리 의무에서 면제를 받을 수 있다. 2020년 3월부터 2020년 7월까지 실시된 미국의 비자 발급 중단[36] 이슈는 총 문서 수집 기간인 1년 3개월 중 4개월에만 해당 되었음에도 상위 키워드로 나타나 '비자' 키워드가 출입국 절차와 관련되어 대중의 중요한 이슈였음을 확인할 수 있다. 중국 우한에서 시작된 코로나19가 전 세계로 빠르게 전파되는 데 있어서 가장 중요한 역할을 한 것 중 하나는 출입국으로 인한 전염병의 확산이었다. 이에 많은 나라들이 추가 전염병을 막기 위한 조치로 국경 폐쇄, 부분적인 여

행 제한, 출입 심사, 여행객 격리 등 국제 여행과 관련된 통제 조치를 시행하였다. WHO 글로벌 데이터베이스를 기반으로 한 연구에서 여행 관련 통제 조치에 대한 실제 증거는 부족하지만 대체로 국경을 넘는 여행을 제한하는 것은 코로나19의 확산을 제한할 수 있고 PCR 검사, 검역 및 관찰을 병행하면 효과가 개선될 수 있다는 결과가 보고된 바 있다.[37] 국내에서 수행된 코로나19와 출입국 제한에 관련된 연구에서는 세계 각국이 자국으로의 바이러스 유입을 차단하기 위해, 전면적인 입국 금지 조치를 취하거나, 비자 발급 중단 등의 까다로운 특별입국절차를 마련해 국제적인 확산을 막기 위해 노력하고 있음을 확인하였다.[35] 전문가 답변 빈도 결과를 통해 사회적인 이슈는 노무사와 변호사가 중요한 역할을 할 수 있음을 유추할 수 있다.

‘사회’ 분야에 이어 중요한 쟁점은 ‘보건 의료’ 의 이슈이다. 주요 키워드 분석에서 ‘코로나’ 다음으로 ‘병원’ 6,307회, ‘증상’ 6,237회가 나타난 바 있고 전문가 답변 중 의사의 답변이 가장 높은 비중을 차지하였다. 또한, 동시 출현 단어 분석 결과에서 ‘걱정’ 과 함께 나타난 단어로 ‘병원’ 과 ‘증상’ 과 같은 보건 의료 분야의 키워드가 ‘코로나’ 다음으로 가장 많은 빈도를 차지하였다. 대중이 코로나19에 대한 감염 여부 및 감염시 나타나는 증상에 대한 걱정을 하고 있다는 것으로 풀이하였다.

증상과 유사한 개념으로 통증 또한 중요한 키워드로 나타났다. 토픽 모델링 분석에서 보건 의료 주제에 해당하는 신체 부위와 관련된 Topic 5, Topic 7에 나온 키워드는 ‘머리’, ‘허리’, ‘다리’, ‘어깨’, ‘근육’ 등과 ‘아프다’, ‘통증’ 키워드로 동시 출현 단어 분석을 통해 단어 간 연관성을 그래프로 표현하였다[그림 7]. 코로나19로 입원한 환자를 대상으로 임상적 특징을 추적 관찰한 영국의 코호트 연구에서 근골격계 통증(근육통, 관절통)을 동반한 확진자 그룹을 확인하였다고 보고하였다.[38] 이전 연구[38]의 결과와 마찬가지로 본 연구에서 언급된 ‘어깨’, ‘근육’ 이 코로나19에 감염되었을 때 발현되는 통증 중 관절과

근육에 통증이 나타나는 증상에 해당할 수 있는 가능성을 유추할 수 있다.

코로나19의 증상은 무증상 감염에서 중증 호흡기 장애에 이르기까지 개인마다 다르다.[39] 중국에서 진행된 코로나19의 증상에 대한 7개 연구를 포함한 메타 분석에 따르면 1,576명의 감염환자 중에서 발열, 기침, 피로, 약간의 호흡곤란 순으로 증상이 나타났고 기저 질환이 있는 중증 환자가 합병증의 위험이 있다는 결론을 도출하였다.[40] 감염자의 증상과 확진 결과에 대해 평가한 연구에서 50% 이상의 민감도를 나타낸 증상은 기침, 인후통, 발열, 근육통, 피로, 두통이 이에 해당했고 주로 노년층에서 발생하였다.[41] 따라서, 코로나19의 증상은 일반적이면서 다양하기 때문에 증상으로만 확진을 판단하기 힘들고 다른 호흡기 질환에서 나타나는 증상과도 구별하기 어렵다.[42-44]

본 연구의 토픽 모델링 결과의 Topic 3는 '머리', '기침', '가래', '콧물', '미열' 등의 감기와 유사한 증상의 키워드를 포함하였고 동시출현 단어 분석에서 '아프다' 키워드는 '머리' 와 가장 많이 출현함으로써 선행 연구에서 언급한 기침, 발열과 두통과 같이 일반적이면서 다양한 증상과 유사한 패턴의 키워드가 추출되었다.

토픽 모델링 결과에서 Topic 4의 주제는 병원 방문에 대한 우려이며 '병원', '수술', '의사', '신경', '선생', '스트레스' 등의 키워드를 포함하여 병원에서 발생할 수 있는 감염에 대해 민감할 수 있음을 추측하였다. 코로나19에 상응하는 감염병 관리가 가능한 병원 시스템 도입과 관리의 필요성을 시사한다. 코로나19는 전 세계적으로 빠르게 확산되어 의료 시스템을 한계까지 확장시키고 있으며 병원 방문에 대한 공포와 두려움을 가중시키고 있다. 터키의 네브셰히르주립병원과 우악수련연구병원의 피부과 방문 환자에 대한 연구에서는 코로나19가 발생한 후 외래 피부과 진료 건수가 크게 줄었다는 것을 발표하여 터키 정부의 행정 조치와 병원에 대한 사람들의 우려를 짐작할 수 있다.[45] 이 같은 변화에 대한 대응으로 독일의 우에르츠부르크 대학병원 외과 병동에서는 응급 수술과 종양학 환자의 관리와 병원 직원의 복지를 유지하고 기하급수적으로 증

가하는 코로나19 감염 환자를 안전하게 수용하기 위해 병원 관리 시스템을 개발하였다. 코로나19 감염자와 수술환자 집단의 우선 순위는 높은 수술 기준을 유지하면서 생명을 보존하는 데 핵심이며 응급실, 진료실 및 수술 구역의 환자 그룹을 엄격히 분리하면 바이러스 확산을 방지하고 병원 직원을 적절하게 교육하고 신중하게 선택하여 업무를 성공적으로 수행할 수 있다는 결론을 내린 바 있다.[46] 토픽 모델링 분석을 한 선행 연구들[12,13]에서도 병원과 관련된 토픽에서 워드 클라우드에 상위 빈도로 병원이 등장하였고 '감염', '마스크', '침대' 등의 키워드와 함께 출현해 감염에 대한 우려를 짐작할 수 있다. 전문가 답변의 결과에서 보건 의료와 관련된 이슈는 의사의 역할이 중요할 수 있음을 암시한다.

토픽 모델링 분석에서 Topic 8은 전체에서 가장 낮은 비중을 차지한 '법률'에 관한 이슈였다. '행정 조치'라는 주제로 분류 되었고 '재판', '조사', '감염', '사건', '벌금', '피해자', '확률', '무증상', '구속' 키워드가 의미하는 것은 국내에서 역학 조사를 거부하거나 방역 지침을 위반했을 때, 구속 수사를 하는 강도 높은 처벌에 대한 연관성이 반영될 수 있음을 나타냈다. 코로나19은 기존에 바이러스로 유입되었던 사스나 메르스와 다르게 사회 경제 전반에 미치는 영향이 종전과는 비교할 수 없을 정도로 충격적인 결과를 낳고 있다. 재조합과 변이가 심한 코로나19의 특성으로 경제적 피해와 필수적으로 마스크를 착용해야 하는 등의 평범한 일상생활이 불가능한 실정이다.[47] 이를 위해 국가 단위의 감염병 통제가 필요하게 되었고 두 번의 감염병 사례로 인해 감염병 예방 및 관리체계가 정비되었지만 코로나19에 대한 예방은 최근에서야 법률로 개정되어 반영되었다. [48] 감염병의 예방 및 관리에 관한 법률, 검역법, 의료법을 포함하며 국민들의 불안과 불편을 해소하고자 정부는 법률 개정안을 의결하였다. 이 중 감염병의 예방 및 관리에 관한 법률 (이하 '감염병예방법')은 코로나19의 대응을 위해 감염병 환자의 진료 정보, 이동 수단, 이동 경로 등의 의무적인 공개와 이의

신청 등을 내용으로 개정되었다. 감염병 환자 등의 격리와 치료 거부, 고의적인 감염병 전파에 대한 벌칙의 내용도 추가되었다. 이는 Topic 8에서 등장한 '조사', '감염', '벌금' 등의 키워드가 개정된 감염병예방법 내 벌칙에 관한 내용을 반영하고 있다는 가능성을 보여주었다. 대중은 행정 처분에 대해서 관심을 가지고 새롭게 개정된 법에 대한 대응을 필요로 한다는 것을 유추할 수 있다. 토픽 모델링의 답변 비율 결과를 바탕으로 법률과 관련된 문제에서도 의사가 전문가로 관여할 수 있음을 알 수 있다.

감성 분석에서 긍정, 부정 감정 지수를 분석한 결과 다소 부정적인 것으로 나타났다. 또한, 2020년 1월 이후부터 5월까지 5개월간 감정지수는 -0.2를 밑돌며 코로나19 발병 초기가 이후 시기에 비해 상대적으로 부정적인 감정이 많이 드러났다. 온라인 교육에 대한 소비자 인식을 변화 분석한 연구[18]에서는 온라인 개강과 팬데믹 선언 직후인 2020년 3월부터 긍정 또는 부정에 대한 사람들의 인식이 나타나기 시작하였고 부정의 빈도가 긍정보다 높았다. Lwin 연구팀[49]은 2020년 1월 28일부터 2020년 4월 9일 까지 코로나19과 관련된 2,000만건 이상의 트위터 글을 분석하여 전염병의 진행 과정에 따라 대중의 감정이 불안에서 화로 옮겨 가는 것을 확인하였다. 이 같은 선행 연구의 결과들은 바이러스 발생 초기에 불안을 포함한 부정적인 감정이 나타났음을 보여주었다.

감정 범주를 사용한 감성 분석을 통해 나타난 결과로 '고통'과 '불안'이 주된 감정 범주로 등장하였다. '고통'에 해당하는 감정 빈도는 2020년 3월에 일 평균 20.65회 등장하고 감소 추세를 보이다가 같은 해 6월에 22.33회 등장하여 가장 높은 빈도로 나타났다. '불안' 감정은 2020년 6월에 일 평균 6.23회 등장하여 가장 높은 빈도를 나타냈으며 같은 해 8월과 9월에 각각 5.1회, 4.27회 등장하였다. 감성 분석에서 '고통'으로 분류된 감정은 신체와 연결되어 감각에 해당하는 '아프다'와 '아야'를 제외하면 매우 낮은 빈도 (57회)로 대중의 감정을 대변하기에 다소 무리가 있다고 볼 수 있다. 반면, '불안'을 나타내는 키

워드인 ‘무섭다’, ‘겁나다’, ‘두렵다’는 신체 감각과 분리되어 정서적인 감정을 기반으로 하고 키워드 빈도가 고르게 분포 되어 있어 ‘불안’이 라는 감정을 충분히 대변할 수 있다.

문서 내 ‘불안’ 감정은 2020년 8월에 다시 증가하는 모습을 보였는데 집단 감염에 의해 국내 일일 확진자 수가 100명이 넘어감에 따라 8월 16일부터 시행된 서울과 경기 지역의 사회적 거리두기 2단계 조치[50]와 관련이 있을 수 있다. 코로나19와 관련하여 약 10만건의 소셜 미디어의 문서를 대상으로 진행된 스페인의 감성 분석 연구[51]에서 우울, 혐오, 화 등의 감정들이 연구 기간 동안 일상 속에서 매우 혼재하고 있음을 보여주었다. 스페인 정부의 경보 시스템 도입 및 통행 제한 조치등의 발표가 사람들의 정서에 큰 영향을 끼쳤고 새로운 질병의 발생과 감염이 일어날 수 있는 가능성에 대한 불안은 압도적일 수 있다고 보고하였다. 확진자 수의 증가와 정부의 감염병 대응 강화 조치가 대중의 불안감을 조성할 수 있다는 결론을 뒷받침한다.

[그림 8]의 월별 감정 지수는 2020년 12월에 급격히 감소하였는데 이는 2020년 12월 12일에 국내 신규 확진자가 950명으로 역대 최다 기록을 경신하고 뒤이어 13일에 1,030명의 확진자가 나오며 정부의 사회적 거리두기 3단계를 검토한 것과 연관이 있을 수 있다.[1] 3차 확산기로 분류되며 한국에서 인구 밀도가 높은 서울 및 수도권을 중심으로 집단 감염이 발생했다는 점에서 대중은 부정적인 감정이 증가하여 온라인 텍스트 문서에 반영 되었을 것이다.

AlAgha 연구팀[13]에서도 토픽 모델링과 감성 분석을 동시에 수행하였고 경제, 정치, 격리 조치, 행정, 증상 등 다양한 분야의 토픽이 분류된다는 것과 일부 주제가 일치하다는 점이 공통점으로 나타났다. 캐나다, 미국, 영국을 대상으로 한 감성 분석에서는 미국과 영국은 대중이 부정적인 감정이 있는 것으로 나타났지만 캐나다는 긍정적인 감정을 가지고 있는 것으로 연구 결과가 나타났다. 이는 캐나다가 코로나19의 영향을 적게 받는 나라임에 의미가 있다고 보고하였다. 한국에서 확진자가 많이 발생했던 2020년 12월에 감정 지수가 크게 하락한 것과 비교하였

을 때, 미국과 영국 등의 감염자 발생이 많은 나라와 대중의 부정적인 감정은 연관이 있을 것으로 생각된다. Amara 연구팀[12]은 세계 다양한 국가와 여러 언어로 코로나19와 관련된 토픽 모델링 분석을 실시하였다. 추출된 주제는 보건 의료와 검역 중 하나인 보호를 위한 절차와 사회, 경제, 정치 등의 다양한 주제가 추출되는 것으로 드러나 선행 연구들과 공통점을 나타냈다.

본 연구에서 사용된 텍스트마이닝 방법은 코로나19와 관련된 대중의 감정과 관련해 사회적으로 형성된 주요 이슈가 무엇인지 실증적으로 분석한 연구이다. 장점으로는 토픽 모델링 방법으로 주요 이슈를 세분화하여 잘 알려지지 않은 주제에 대한 정보를 제공할 수 있고 코로나19에 관련된 주제와 주요 키워드 및 시각화한 그래프를 제시함으로써 의료 정책 담당자 및 연구원들이 현상을 파악할 때 다양한 관점을 제공해 줄 수 있다. 또한, 비정형 데이터를 분석하여 의견이나 감정, 태도 등의 단순한 감정 지수 측정을 넘어서 풍부한 주관적 정보를 반영하고 시간의 추이에 따른 감염병에 대한 대중의 인식을 파악할 수 있다. 토픽 모델링 및 감성 분석을 활용한 국내 코로나19 관련 선행 연구는 대부분 1년 미만의 기간동안 진행되었다. 1년 3개월간 소셜 미디어의 한국어 텍스트 데이터를 대상으로 분석한 연구라는 것에 의의가 있다. 전문가는 '코로나'와 관련 없는 광고성 글이나 논점에서 벗어난 질문에 답변하지 않기 때문에 전문가의 답변이 있는 질문 글을 선별하여 수집 함으로써 연구 결과의 질을 높일 수 있다. 텍스트 마이닝 분석 연구를 활용한 예로 캐나다 AI기반 건강 모니터링 플랫폼 블루닷(BlueDot)은 2019년 12월 31일 중국 우한에서 해산물, 동물 거래 시장과 관련 있는 27건의 폐렴 발생 기사의 텍스트 데이터를 분석하고 전세계 항공 발권 데이터 등을 함께 활용하여 감염자의 동선을 추정하였다. 바이러스의 확산 방향 및 추세와 강도를 예측하였고 홍콩, 마카오, 서울, 도쿄 등으로 확산될 가능성이 있다고 경고했는데 WHO의 2020년 1월 9일 발표보다 일주일 이상 앞서 발표하였다. 2018년 메르스 해외 유입에 관한 텍스트 마이닝 분석

연구 역시 주요 온라인 이슈를 파악하고 국가 위기적 질병에 대해 실효성 높은 대응책을 마련하는데 공헌 할 수 있다는 함의점을 갖는다고 보고하였다. 본 연구도 코로나19의 이슈 분석을 통해 얻은 정보를 고용이나 병원 또는 법률 기관의 데이터와 결합하여 코로나19에 대한 예측과 대응 방안 시스템을 구축할 필요가 있다.

본 연구의 제한점으로는 토픽 모델링 분석에서 연구자가 자의적으로 해석할 수 있기 때문에 같은 데이터가 있다고 하더라도 연구자마다 결과에 차이가 발생할 수 있다. 코로나19와 관련된 주제 내 분포된 사회, 보건 및 법률적 이슈가 상대적으로 중요성을 나타내고 있다는 것을 시사하였지만 명확한 인과관계를 밝히는 구체적인 연구 모형이 필요하다. 사전을 기반으로 한 감성 분석으로써 등록된 사전이 실생활에서 쓰이는 어휘와 다를 수 있다는 점과 문장의 어순을 고려하지 못한다는 점에서 감정 분포에 대한 결과 반영이 적절하게 이루어지지 않을 수 있다. 뿐만 아니라, 네이버 지식 포털의 질문 게시글을 근거로 얻어진 결과이기 때문에 코로나19의 전반적인 문제들을 대표한다고 보기는 어렵다. 이를 보완할 수 있도록 SNS와 함께, 뉴스, 논문초록, 정부 보도자료 등을 추가하여 대규모의 데이터를 수집해 신뢰성을 높인 분석을 수행하여 주제를 구체화하는 것을 제언한다.

제 5 장 결 론

본 연구는 코로나19와 관련해 제기된 우리 사회 전반의 이슈 및 보건 의료를 파악하기 위해 ‘코로나’ 키워드로 검색된 지식 공유 포털 내 문서를 텍스트마이닝 기법으로 분석한 탐색적 연구이다. 크롤링 방법을 통해 수집한 23,463건의 문서를 대상으로 텍스트 마이닝 기법을 통해 토픽 모델링 분석, 동시 출현 단어 분석 그리고 감성 분석을 수행하였다. 토픽 모델링 분석을 통해 8개의 토픽을 도출하고 각 토픽에 토픽명을 부여하였다. 동시에 출현하는 단어가 얼마나 등장하였는지에 따라 중요 키워드 간 연관성을 분석하였고, 감성 분석을 통해 대중이 코로나19에 대한 감정이나 태도를 조사하였다.

사회적 측면에서 고용과 임금을 둘러싼 회사 문제가 주요한 쟁점으로 나타났다. 출입국 절차가 강화되어 격리 조치가 시행되고 비대면 접촉을 통해 학교에서는 온라인 수업을 진행하고 회사에서는 화상 회의를 진행하는 등의 사회 구조적 변화를 암시하는 키워드를 통해 코로나19가 사회에 미치는 영향이 큰 것을 알 수 있다.

보건 의료 측면에서는 코로나19의 의심 증상과 통증에 대한 대중의 우려, 관심과 관련된 키워드가 다수 등장하였다. 의심 증상이라는 주제로 분류된 토픽 모델링 결과에 나타난 주요 증상들은 코로나19가 일반 호흡기 질환과 구분하기 어렵다는 연구 결과와 비슷하게 일반적인 호흡기 질환 및 감기 증상을 나타내는 키워드가 발견되었다.

법률 측면에서 행정 처분 관련 키워드를 통해 대중은 사회적인 이슈 외에도 감염병 관련 법률에 대한 관심을 나타낼 수 있다는 것을 알 수 있다.

전반적으로 코로나19에 대한 대중의 감정은 다소 부정적인 것으로 나타났으며 바이러스에 감염될 수 있다는 불안감을 가지고 있는 것을 짐작해 볼 수 있다. 대중은 감염병의 유행 초기 단계에 불안을 느끼고 부정적인 태도를 취한다는 감성 분석의 결과를 도출하였고 이후 부정적인 감

정과 태도는 완화되었다.

텍스트 마이닝 기법을 적용한 본 연구의 분석 결과가 기존 연구자의 직접 분석을 통한 결과와 유사하게 도출된 것에 의의가 있다. 또한, 새로운 방법론을 제시하여 보건 의학 분야의 다양한 주제를 분석하는 데 활용될 수 있을 것이다. ‘코로나19’에 대한 토픽은 지식 포털 뿐만 아니라 다양한 텍스트 데이터를 통해서 파악할 수 있다. SNS와 함께, 뉴스, 논문초록, 정부 보도자료 등을 추가하여 분석을 수행할 것을 제언한다.

참 고 문 헌

1. 코로나바이러스감염증-19 발생동향[웹사이트]. (2021.03.14.). URL:http://ncov.mohw.go.kr
2. Perlman, S. (2020). Another decade, another coronavirus. *New England Journal of Medicine 382* (8), 760-762.
3. Tyrrell, D. A. J., & Bynoe, M. L. (1966). Cultivation of viruses from a high proportion of patients with colds. *Lancet*, 76-7.
4. Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical medicine & international health, 25*(3), 278.
5. GISAID Global Initiative on Sharing All Influenza Data . Phylogeny of SARS-like betacoronaviruses including novel coronavirus (nCoV). (Available from: https://nextstrain.org/groups/blab/sars-like-cov).
6. Al-Salihi, K. A., & Khalaf, J. M. (2021). The emerging SARS-CoV, MERS-CoV, and SARS-CoV-2: An insight into the viruses zoonotic aspects. *Veterinary World, 14*(1), 190.
7. Fauci, A. S., Lane, H. C., & Redfield, R. R. (2020). Covid-19—navigating the uncharted. *New England Journal of Medicine 382* (13), 1268-1269.
8. 박선영, & 이재림. (2021). 소셜 빅데이터로 알아본 코로나 19와 가족생활: 토픽모델 접근. *한국콘텐츠학회논문지, 21*(3), 282-300.
9. 박은철. (2020). 코로나 19 이후 시대. *보건행정학회지, 30*(2), 139-141.
10. Ahmed, W. (2018). *Using Twitter data to provide*

qualitative insights into pandemics and epidemics (Doctoral dissertation, University of Sheffield).

11. Jordan, S. E., Hovet, S. E., Fung, I. C. H., Liang, H., Fu, K. W., & Tse, Z. T. H. (2019). Using Twitter for public health surveillance from monitoring and prediction to public response. *Data, 4*(1), 6.
12. Amara, A., Taieb, M. A. H., & Aouicha, M. B. (2021). Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. *Applied Intelligence, 51*(5), 3052-3073.
13. AlAgha, I. (2021). Topic Modeling and Sentiment Analysis of Twitter Discussions on COVID-19 from Spatial and Temporal Perspectives. *JOURNAL OF INFORMATION SCIENCE THEORY AND PRACTICE*, *9*(1), 35-53.
14. Yin, H., Yang, S., & Li, J. (2020, November). Detecting topic and sentiment dynamics due to Covid-19 pandemic using social media. In *International Conference on Advanced Data Mining and Applications* (pp. 610-623). Springer, Cham.
15. 김태종. (2020). 뉴스 빅데이터를 활용한 코로나 19 언론보도 분석: 토픽모델링 분석을 중심으로. *한국콘텐츠학회논문지, 20*(5), 457-466.
16. 김상미. (2020). 코로나 19 관련 온라인 교육에 관한 국내 언론 보도기사 분석. *한국디지털콘텐츠학회 논문지, 21*(6), 1091-1100.
17. 유소연, & 임규건. (2021). 텍스트 마이닝과 의미 네트워크 분석을 활용한 뉴스 의제 분석: 코로나 19 관련 감정을 중심으로. *지능정보연구, 27*(1), 47-64.
18. 손민성, 임미자, & 박경환. (2021). 텍스트 마이닝을 활용한 온

라인 교육에 대한 소비자 인식 변화 분석: COVID-19 전후를 중심으로. *디지털융복합연구, 19*(1), 29-43.

19. 박은정, & 조성준. (2014). KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지. *제26회 한글 및 한국어 정보처리 학술대회 논문집*, 1-4.

20. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

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

22. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

23. Roder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

24. Yin, H., Yang, S., & Li, J. (2020, November). Detecting topic and sentiment dynamics due to Covid-19 pandemic using social media. In *International Conference on Advanced Data Mining and Applications* (pp. 610-623). Springer, Cham.

25. Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX* (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

26. Jang, H., Kim, M., & Shin, H. (2013, November). KOSAC: A full-fledged Korean sentiment analysis corpus. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)* (pp. 366-373).
27. 손선주, 박미숙, 박지은, & 손진훈. (2012). 한국어 감정표현단어의 추출과 범주화. *감성과학*, *15*(1), 105-120.
28. Bae, S. H., Sin, W. S., & Gwak, H. S. (2020). 건설기술인의 COVID-19 팬데믹에 의한 영향분석과 정책적 과제. *Construction Engineering and Management, 21*(6), 33-40.
29. Beland, L. P., Brodeur, A., & Wright, T. (2020). The short-term economic consequences of Covid-19: exposure to disease, remote work and government response.
30. Hensvik, L., Le Barbanchon, T., & Rathelot, R. (2021). Job search during the COVID-19 crisis. *Journal of Public Economics, 194*, 104349.
31. 최성웅. (2020). 재택근무가 가능한 일자리의 특성과 분포: 물리적 근로환경을 중심으로. *한국경제지리학회지*, *23*(3), 276-291.
32. Sahinaslan, O., Sahinaslan, E., & Gunes, E. (2021, March). Review of the contributions of contactless payment technologies in the COVID-19 pandemic process. In *AIP Conference Proceedings* (Vol. 2334, No. 1, p. 070002). AIP Publishing LLC.
33. Auer, R., Cornelli, G., & Frost, J. (2020). Covid-19, cash, and the future of payments (No. 3). *Bank for International Settlements.*
34. 박혜진, 김한성, & 박한우. (2020). COVID-19 팬데믹과 온라

인 개학: 트위터 담화 분석. *Journal of the Korean Data Analysis Society (December 2020), 22*(6), 2535-2549.

35. 이재유. (2020). 코로나 19 대응 관련 한국의 출입국 제한조치 및 국제법적 평가. *서울국제법연구, 27*(1), 65-101.

36. 박인영. (2020). 미국의 이주민 정책 현황과 바이든 행정부의 과제. *국제사회보장리뷰, 2020*(겨울), 119-123.

37. Burns, J., Movsisyan, A., Stratil, J. M., Coenen, M., Emmert-Fees, K. M., Geffert, K., ... & Rehfuess, E. (2020). Travel-related control measures to contain the COVID-19 pandemic: a rapid review. *Cochrane Database of Systematic Reviews*, (9).

38. Docherty AB, Harrison EM, Green CA, et al. Features of 16,749 hospitalised UK patients with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol. *medRxiv* 2020; published online April 28. DOI:10.1101/2020.04.23.20076042 (preprint).

39. He F, Deng Y, Li W. (2020). Coronavirus disease 2019 (COVID-19):What we know? *J Med Virol. 92*, 719-725.

40. Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., ... & Zhou, Y. (2020). Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: a systematic review and meta-analysis. *Int J Infect Dis, 10.*

41. Struyf, T., Deeks, J. J., Dinnes, J., Takwoingi, Y., Davenport, C., Leeflang, M. M., ... & COVID, C. (2020). Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19 disease. *Cochrane Database of Systematic Reviews,* (7).

42. Li, L. Q., Huang, T., Wang, Y. Q., Wang, Z. P., Liang, Y.,

Huang, T. B., ... & Wang, Y. (2020). COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. *Journal of medical virology, 92*(6), 577-583.

43. Singhal, T. (2020). A review of coronavirus disease-2019 (COVID-19). *The indian journal of pediatrics, 87*(4), 281-286.

44. Sun, P., Lu, X., Xu, C., Sun, W., & Pan, B. (2020). Understanding of COVID-19 based on current evidence. *Journal of medical virology, 92*(6), 548-551.

45. Kutlu, O., Guneş, R., Coerdt, K., Metin, A., & Khachemoune, A. (2020). The effect of the "stay-at-home" policy on requests for dermatology outpatient clinic visits after the COVID-19 outbreak. *Dermatologic therapy, 33*(4), e13581.

46. Flemming, S., Hankir, M., Ernestus, R. I., Seyfried, F., Germer, C. T., Meybohm, P., ... & Wiegering, A. (2020). Surgery in times of COVID-19—recommendations for hospital and patient management. *Langenbeck's archives of surgery, 405*(3), 359-364.

47. 김은중, & 이동섭. (2020). 코로나바이러스: 사스, 메르스 그리고 코비드-19. *대한임상검사과학회지 (Korean Journal of Clinical Laboratory Science)*, *52*(4), 297-309.

48. 정태종. (2020). [감염병 예방 및 관리에 관한 법률] 의 내용과 개선방안. *법학연구, 20*(3), 289-316.

49. Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R., & Yang, Y. (2020). Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. *JMIR public health and surveillance,*

6(2), e19447.

50. MHOW-CDMCD. 코로나바이러스감염증-19 중앙재난안전대책본부 정례브리핑(8 월17 일). Retrieved May 25, 2021, from http://ncov.mohw.go.kr/tcmBoardView.do?brdId=&brdGubun=&dataGubun=&ncvContSeq=359039&contSeq=359039&board_id=&gubun=ALL

51. de Las Heras-Pedrosa, C., Sanchez-Nunez, P., & Pelaez, J. I. (2020). Sentiment analysis and emotion understanding during the covid-19 pandemic in spain and its impact on digital ecosystems. *International Journal of Environmental Research and Public Health, 17*(15), 5542.

52. 한국인터넷진흥원. (2021). 코로나19 대응을 위한 데이터 활용 현황 및 사례. *월간동향 보고서.*

53. 조재희, & 조인호. (2019). 2018 메르스 해외 재유입에 대한 주요 온라인 이슈 탐색: 토픽모델링 분석과 감성 분석을 중심으로. *한국디지털콘텐츠학회 논문지, 20*(5), 1051-1060.

Abstract

Issue analysis applying text mining on COVID-19

Song SungHee
Department of Preventive Medicine
The Graduate School
Seoul National University College of Medicine

This study was conducted to identify problems associated with COVID-19 from the point of view of public health and society in South Korea. A total of 23,463 postings was collected on the knowledge-sharing platform such as Naver Knowledge IN under the keyword 'Corona' from January 2020 to March 2021.

The collected texts were broken down into morphemes. After the pre-processing, we used the text-mining method to analyze the data. Frequency was measured by analyzing word frequency, and topics were classified using Latent Dirichlet Allocation (LDA). The co-occurrence network of main keywords in the text about COVID-19 has been analyzed. In this paper, we analyzed the public opinions and emotions through sentimental analysis.

We present three categories to analyze online postings related to COVID-19: the social, the public health, and the legal sectors.

The public' s concern was geared towards employment, making wage a main topic in the social sector. The suspected symptoms of COVID-19 were one of the main interests in the public health sector, while in the legal sector, the governmental administrative measures gained attention. We discovered that most people had negative feelings towards COVID-19, and in the center of the emotions lies pain and fear.

keywords : COVID-19, public health, society, text-mining, topic modeling, sentiment analysis
Student Number : 2019-23275