2023-1 학부생 연구 참여 프로그램(UROP)

코로나19 뉴스 기사의 감성 분석과 언론의 감성 동향 비교

성신여자대학교 통계학과 심 대 희 2023년 8월 18일

1. 서론

2019년 12월 중국에서 발생한 코로나 바이러스는 전 세계 유례 없는 팬데믹 위기를 만들었다. 2020년 1월 20일 해외 유입 국내 첫 감염자 발생을 시작으로 코로나 바이러스의 누적 확진자 수는 기하급수적으로 증가하여 사회 전반의 생활양식에 많은 변화를 만들었다. 세계보건기구(World Health Organization)는 코로나19에 대하여 20 20년 3월 11일 전염병 경보 최고 단계인 '세계적 대유행'이라는 의미의 '팬데믹(Pandemic)'을 선언하였다. 이는 단순 공중 보건의 위기가 아닌 사회, 경제 등 모든 분야에 영향을 미치는 위기이며 이를 극복하기 위하여 국가 차원의 대응이 필요했다. 3년 4개월이 지난 2023년 6월, 정부는 코로나19 엔데믹(Endemic) 선언을 했지만 코로나바이러스는 그 사이 동안 사회, 경제, 정치, 문화 등 다방면에서 우리의 삶에 많은 영향을 미쳤다.

사회 네트워크 서비스(SNS)의 발전과 스마트 기기의 대중화로 대중들은 다양한 매체들을 통하여 코로나19와 관련된 정보를 접할 수 있었다. 언론 매체, SNS, 포털 사이트 등을 통해서 발병 상황에 대하여 언급하고 소통하였으며, 이렇게 기록된 텍스트형태의 빅데이터들은 소셜미디어나 인터넷의 텍스트 데이터를 기반으로 코로나19 관련 텍스트를 분석하는 연구들이 많이 진행되었다. 특히 코로나19에 대한 주요 주제를 파악하기 위한 텍스트 마이닝 연구는 이전부터 수행되어왔다. 해외의 Amin 연구팀과 lyad AlAgha 연구팀은 페이스북, 트위터와 같은 소셜 미디어 자료를 기반으로 코로나19 관련 텍스트 마이닝 분석을 수행한 바 있다. 국내에서는 트위터 및 네이버 지식인 자료를 활용하여 코로나19와 대중의 인식, 불안 확산에 관해 수행한 연구들이 있다. 그러나 이전 국내 연구들은 1년 이하의 비교적 짧은 조사 기간으로 연구가 수행되었으며 전체적인 동향을 파악하는데 어려움이 있을 수 있다.

본 연구는 2020년 2월부터 2021년 1월까지의 1년간 코로나19 관련 뉴스기사(코로나19, 코로나, 코로나 바이러스, 신종 코로나바이러스, COVID-19, 코비드19) 제목에 대한 종단적 연구로써 텍스트마이닝과 lexicon-based 방법의 감성 분석을 진행한다. 이를 통하여 관련 키워드들과 단어 빈도를 파악하고, 코로나19에 관한 언론 기사들을 실제 사건들과 연관하여 시간의 흐름에 따른 정서적 동향을 관찰하고자 한다. 또한 각 언론 매체의 정서를 비교 분석하고, 매체 간의 정서 변화의 추이와 매체의 편향성에 대한 가설 검정을 해보고자 한다.

2. 선행연구정리

2.1 감성분석

감성분석이란 감성 어휘(sentiment terms)와 어휘의 긍정 및 부정의 정도를 나타내는 극성(polarity)으로 구성된 감성 사전(sentiment dictionary or sentiment lexico n)을 활용하여 감성을 정량화하는 것을 의미한다.[1, 2] 따라서 감성분석의 경우 텍스

트 데이터에 대한 분석 이전에 데이터에 내재되어 있는 '감정'이 무엇인지 결정하며 텍스트에서 특정한 감정(부정적, 중립적, 긍정적)이 도출되는지 혹은 감정 점수 등을 통해 얼마나 나타나는지를 분석하는 것이 목적이다(백영민, 2017).

감성 분석에는 다양한 방식이 있는데, 본 연구에서는 감성 사전을 사용한 lexicon-ba sed 기반 감성 분석을 진행할 것이다. 한국어 감성분석의 경우 서울대학교에서 개발한 한국어 감성분석 코퍼스 KOSAC(Korean Sentiment Anaysis Corpus)과 군산대학교에서 구축한 KNU 한국어 감성사전이 많이 사용되고 있다. KNU 한국어 감성 사전은 감성 어휘를 매우 긍정(2), 긍정(1), 중립(0), 부정(-1), 매우 부정(-2)으로 상세구분하고 있다고 판단해 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[3, 4].

3. 연구 방법

3.1 자료 수집

본 연구의 분석범위는 시간적으로 2020년 2월부터 2021년 1월까지 한정하여 해당 기간 내에 발행된 코로나19 관련 보도기사를 수집 및 활용하고자 한다. 뉴스 기사들은 언론 데이터 수집이 용이한 빅카인즈(bigkinds.or.kr)의 뉴스 검색 기능을 사용하여 수집하며 엑셀 형태로 저장하여 분석에 활용한다. 코로나19 관련 검색 키워드는 다음과 같다: 코로나19, 코로나, 코로나 바이러스, 신종 코로나바이러스, COVID-19, 코비드19. 해당 기간동안 작성된 1,399,961건의 뉴스 기사 데이터의 언론사, 제목, 발행일자, 기관 등을 분석에 활용한다.

3.2 전처리 과정

수집한 뉴스 기사의 제목들을 토큰화하기 위하여 R 프로그래밍에서 tidytext 패키지의 unnest_token 함수를 사용한다. 이는 텍스트 데이터를 토큰화하여 단어나 어절과 같은 작은 단위로 분해하는 데 사용되어 데이터를 구조화하고 분석이 가능한 형태로 변환시켜준다. 이후 숫자와 특수문자 등을 제거하였고, 한 글자 단어들도 필터링하여 제거하였다. 이는 '1차', '2차'와 같이 차수를 나타내는 단어와 '3명', '104명'처럼 인원 수를 나타내는 '차'와 '명'과 같은 단어들과 조사를 제거해주기 위하여 추가하였다. 또한 긍정(1)으로 구분된 '대상'의 경우, 표준국어대사전에 명시된 '여러 가지 상 가운데 가장 큰 상'의 의미가 아니라 '어떤 일의 상대 또는 목표나 목적이 되는 것'의 의미로 많이 사용되었다고 판단하여 불용어로 제거했다.

3.3 감성분석

뉴스 기사의 제목을 토큰화한 단어들과 KNU 감성사전에 정의된 긍정어/부정어를 연계하고, 감성 사전에 등재되지 않아 감성 점수가 부여되지 않은 단어들은 중립으로 간주하여 감성점수에 0을 할당하여 총 감성 점수를 합산하였다. 문서마다 합산된 감성 점수가 양수인 경우 긍정 기사로, 음수인 경우 부정 기사로, 0인 경우 중립 기사로 분류하였다. 이를 기반으로 각 극성의 상위 10개의 단어를 비교하였고, 월별 긍정/부정 기사 개수와 감성 수치 등을 계산하였다. 본 연구에서는 언론 매체들의 코로나 19 언론 보도의 편향성을 확인하기 위해 '언론사별 감성 점수는 같다'라는 귀무가설을 세우고 이를 Tukey 검정을 수행하여 가설 검정을 해보고자 한다. 이를 통해 코로나19에 대한 언론매체들의 감성 동향을 분석하였다.

모든 분석은 R 통계 소프트웨어를 이용하였으며 텍스트 마이닝 분석을 위해 'KoNL P', 'tidyr', 'rstatix' 등의 R packages를 이용하였다.

4. 연구결과

4.1 코로나19 텍스트 마이닝

총 1,399,961건의 뉴스 기사의 제목을 토큰화하여 11,589,762개의 단어에서 숫자와 특수문자, 불용어 등을 제거하여 총 9,374,257개의 단어에 감성 점수를 부여할 수 있었다. 이를 극성별로 분류하여 개수를 파악해본 결과 긍정의 단어가 총 133,368개(1.4%), 부정의 단어가 총 142,450개(1.5%), 중립의 단어가 9,098,439개(97.1%)로 나왔다. 2020년 2월부터 2021년 1월까지의 감성점수의 총합은 -11,598이 나왔는데, 이는 매우 부정(-2)의 감성 점수를 가지는 '공포', '최악', '불안' 등의 부정적 단어들이 많이 출현하여 나타난 결과라는 것을 <표 1>을 통해서 알 수 있다.

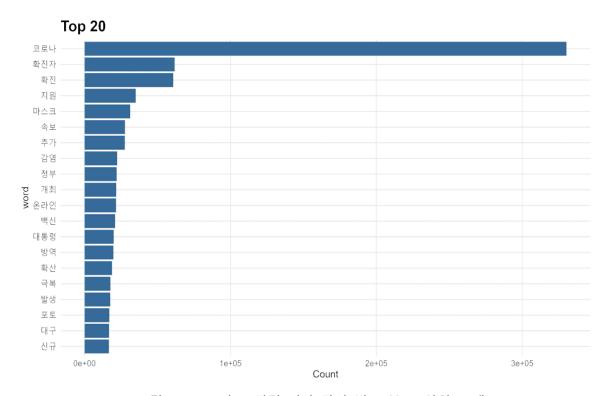
<표 1> 감성 어휘 개수(왼쪽)과 매우 부정(-2)의 단어 빈도(오른쪽)

점수	빈도	단어	빈도
-2	60888	공포	2344
-1	81562	최악	2238
0	9098439	불안	2121
1	74996	아니다	1835
2	58372	어려운	1770

중립으로 분류한 단어들을 포함하여 코로나19에 대한 뉴스 기사들 중에서 많이 등장

한 단어들은 <그림 1>과 같다. 분석 대상이 코로나19와 관련된 데이터로서, 기사 제목에서 가장 빈번하게 등장한 단어는 '코로나'(330359회)로 나타났다. 그리고 두 번째로 빈도가 높은 단어는 '확진자'(61709회)이지만 두 단어의 빈도는 5배 이상의 차이를 보여주었다.

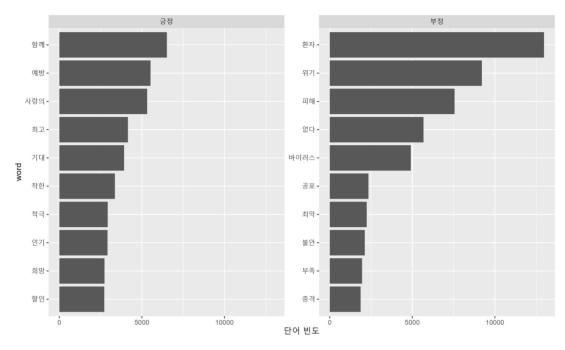
텍스트 분석을 통하여 <그림 1>에서 '확진자'와 '확진', '발생'의 단어처럼 코로나19에 대한 현황을 보도하는 기사와 '마스크', '방역', '백신', '극복' 등 의료 대응 체계 혹은 예방 및 해결과 관련된 기사들을 확인할 수 있었다.



<그림 1> 코로나19 관련 기사 단어 빈도 분포: 상위 20개

4.2 코로나19의 긍정적, 부정적 이미지

감성 사전을 기반으로 하여 코로나19의 긍정적 감성을 분석한 결과, '함께', '예방'처럼 공동체적 목표를 비롯한 감성과 '사랑의', '최고', '착한'을 비롯한 호감 감성 등이 도출되었다. 코로나19의 부정적 감성을 분석한 결과, '환자', '피해', '위기' 등 바이러스로 인한 위험 상황에 대한 감성과 '공포', '최악', '불안', '충격' 등의 부정적인 감성들이 도출되었다. 위의 감성들 외에도 다양한 감성들이 도출되었으나 모두 분석하기는 한계가 있으므로 각 극성 단어들의 상위 10개의 감성 단어들을 정리하였다. 상위 10개의 데이터는 <그림 2>와 같다.



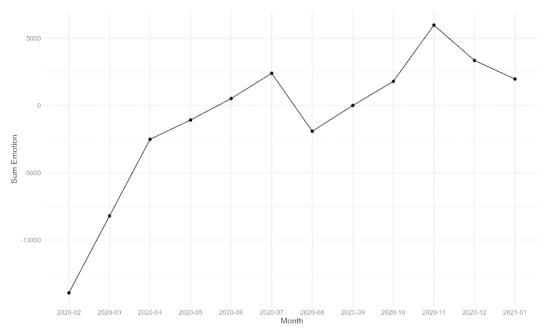
<그림 2> 긍정, 부정 단어 출현 빈도 상위 10개

4.3 시간별 감성 변화

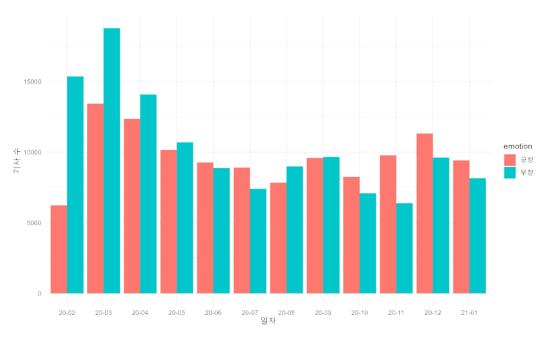
보도된 기사들의 감성 점수들을 통하여 월별 감성 변화 흐름을 관측하고자 했다. 이에 대한 결과는 <그림 3>에서 확인할 수 있으며, 평균 감성 변화도 이와 유사하게 나타난다.

<그림 3>에서 감성 점수의 총합의 가장 낮은 기간은 2020년 2월이고, 가장 높은 기간은 2020년 11월이다. 전체적으로 시간의 흐름에 따라 감성 점수의 합이 증가하는 경향이 있으며 이는 초기에 언론에서 코로나19에 대한 부정적인 감성을 보도하였으나, 시간이 지남에 따라 그 양상이 달라진 것이라는 것을 확인할 수 있다. <그림 3>에서 보면 4월의 감성 점수 합이 지난달에 비하여 급격히 증가하였으며, 이후 증가하는 경향을 보이다가 8월에 감소한 것을 확인할 수 있다. 이후 11월에 가장 높은 점수를 기록하였고, 이후 다시 감소하는 모습이 보이며 이에 대한 원인을 찾고자 4월과 7월, 8월, 11월에 보도된 기사들의 제목 텍스트 분석을 수행했다. 그 결과 4월의 경우, 총선 결과와 관련하여 '압승'(매우 긍정:2) 단어의 사용이 많았으며, 7월에는 15일에 모더나 임상 실험 성공과 국내 첫 임상 실험 진행 등의 사건이 나타났고, 한국 수(水) 자원공사에서 비전을 선포하는 등의 내용을 묘사하며 '최고'(매우 긍정:2) 단어의 사용이 많았다. 8월에는 사랑 제일 교회 확진자 급증과 관련된 부정적인 내용이 주를 이루었고, 11월에는 바이든의 당선으로 '승리'(매우 긍정:2)를 묘사하는 기사가 많아이러한 경향이 나타난 것으로 해석된다.

이후, 매월 발행된 기사들의 감성 분류를 확인하고 한 달 동안의 긍정 및 부정, 중립 문서의 변화를 분석해보고자 기사 제목의 감성 점수 총합을 계산하였다. 이때 감성 단어 점수 총합이 0일 경우 중립 문서로 간주하였고, 0을 초과한 값은 긍정적인 기사, 미만인 값은 부정적인 기사로 분류했다. 이러한 분류 기준으로 월별 극성 문서들의 개수를 파악하였고, 이를 토대로 월별 전체 문서의 수를 분모로, 각 극성 문서의 개수를 분자로 하여 비율을 계산하였다. 이때, 매월 보도된 전체 기사 중에서 중립 문서가 80% 이상의 높은 비중을 차지하여 중립 문서를 제외하고 긍정, 부정 기사들의 개수를 비교하였다.



<그림 3> 코로나19의 월별 감성 점수 합 추이



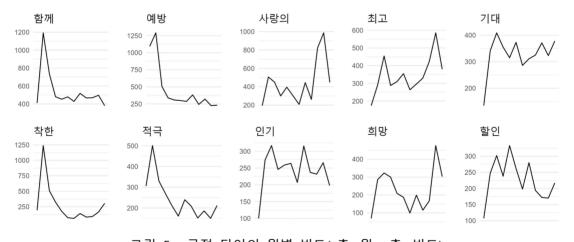
<그림 4> 코로나19의 월별 긍정 기사와 부정 기사 개수

<그림 4>를 보면 2월에는 부정적인 제목의 기사가 많았지만, 3월이 되면서 부정적인 제목의 기사보다 긍정적인 제목의 기사의 증가량이 더 많은 것을 확인할 수 있다. 이후 6월에는 긍정 감성의 기사가 더 많아졌으며, 이후 8, 9월을 제외하고 긍정 제목의 기사가 더 많이 보도되었다. 이를 통하여 전체적인 감성 총합의 변화와 유사한 추이를 갖는 것 확인할 수 있다.

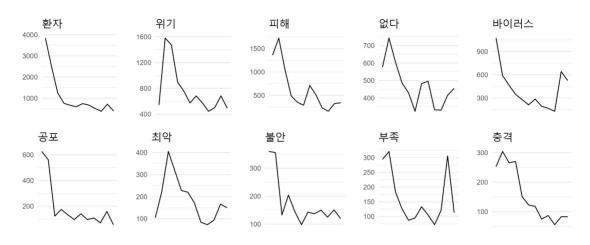
이를 <그림 5>와 <그림 6>의 월별 단어 출현 빈도와 함께 확인해보았다. <그림 5>의 긍정 단어의 경우, '함께', '예방', '착한', '적극'의 단어가 3월에 많이 등장하였으며 이후에는 빈도가 감소한 것을 확인할 수 있다. '최고', '희망'의 감성 단어는 후반에 많이 등장하였으며, '기대', '인기'는 2월을 제외하고 매월 200회 이상 나타났다.

<그림 6>의 부정 단어의 경우, 2020년 상반기에 많이 등장하였다. 특히 '환자', '바이러스', '공포', '불안' 등의 코로나로 인한 피해와 두려움의 감성 단어들은 2월과 3월에 많이 등장하였다가 이후 급격히 빈도가 감소하는 추세를 보인다. '위기'와 '피해', '없다'등 위험 상황에 대한 감성 단어들은 3월에 가장 높은 빈도를 기록했다. '부족'의 경우, 2, 3월에 많이 등장했다가 급격히 감소하였고 이후 12월에 다시 높은 빈도로 나타났다.

대체적으로 긍정 단어의 경우, 시간의 흐름에 따라 빈도가 증가하며, 부정 단어의 경우에는 시간의 흐름에 따라 감소하는 경향을 확인할 수 있었다. 이는 <그림 3>, <그림 4>와 함께 연관된 동향이라는 것을 알 수 있다.



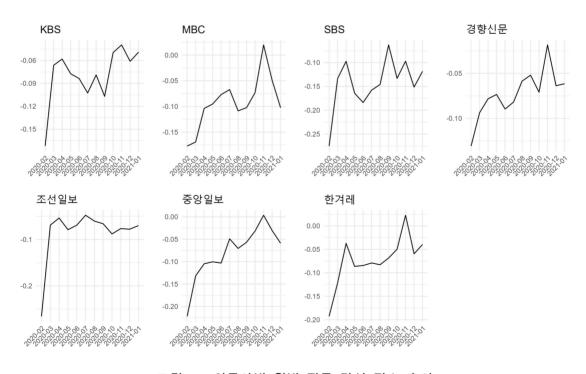
<그림 5> 긍정 단어의 월별 빈도(x축: 월, y축: 빈도)



<그림 6> 부정 단어의 월별 빈도(x축: 월, v축: 빈도)

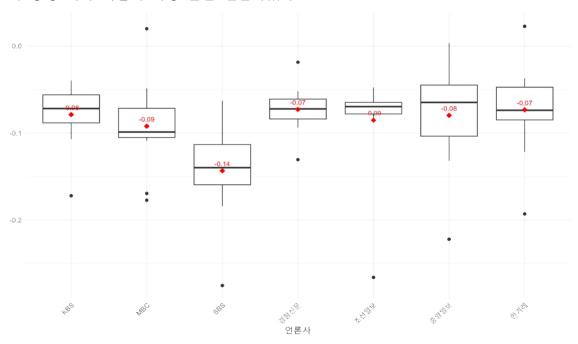
4.3 시간별 감성 변화

언론사들의 평균 감성 점수의 동향을 비교하기 위하여 언론사별로 데이터를 분류하고 월별 감성 점수의 합을 월간 보도한 기사 개수의 합으로 나눠서 평균 감성 점수를 구 하였다. 이후 언론사의 인지도와 성향 등을 고려하여 총 54개의 언론사 중에서 KBS, SBS, MBC, 경향신문, 조선일보, 중앙일보, 한겨레 총 7개의 언론사를 필터링하여 시 간의 흐름에 따른 변화를 <그림 7>처럼 나타내었다.



<그림 7> 언론사별 월별 평균 감성 점수 추이

7개의 언론사 모두 2020년 2월과 2021년 1월의 평균 감성 점수를 비교했을 때, 확연히 증가하였다. 지상파의 경우, KBS, SBS는 2월과 3월 사이에 급격히 증가했지만, MBC는 3월과 4월 사이에 증가하였다. 지상파 언론사 모두 하반기에도 큰 폭으로 증가하는 구간이 있는데, 여기서도 MBC가 10월부터 11월 사이에 증가하며 늦은 변화를 보인다. 월별 평균 감성 점수가 0을 넘겨본 언론사도 지상파 중에서는 MBC만 있으며 지상파 외에는 중앙일보와 한겨례가 11월에 평균 감성 점수가 0을 넘었다. 이후 '언론사별 평균 감성 점수가 같다.'라는 가설을 검정하였다. 7개의 언론사의 월별 평균 감성 점수들의 경향과 분포를 확인하기 위하여 Box-plot을 그려본 결과 <그림 8>과 같은 결과를 얻을 수 있었다. 6개 언론사의 평균 감성 점수는 -0.1에서 -0.06 사이에 위치했는데, SBS는 -0.14로 타 언론사들보다 코로나19에 대하여 더 부정적이었다. 실제로 <그림 9>를 보면 SBS의 부정 기사 비율이 15.2%로 가장 높은 것을 확인할 수 있다. 이외에 KBS가 중립 기사 비율이 가장 높은 언론사였고, 중앙일보가 긍정 기사 비율이 가장 높은 언론사였다.



<그림 8> 언론사별 평균 감성 점수 Box-plot



<그림 9> 언론사별 기사 제목의 감성 비율

'지상파의 평균 감성 점수는 같다.'라는 가설을 검정하기 위하여 Post hoc analysis 의 Tukey (HSD) test를 실시하였다. HSD는 Tukey에 의해 개발된 Honestly Significant Difference(HSD) test이다. 해당 검정의 주요 아이디어는 두 평균 사이의 HSD를 계산하는 것으로, 이를 위해 Student에 의해 정의된 통계 분포인 q 분포를 사용한다(Abdi, Hervé, and Lynne J. Williams. "Tukey's honestly significant difference (HSD) test." Encyclopedia of research design 3.1 (2010)). 검정 결과 <표, 그림>에서 KBS, MBC, SBS 3개의 언론사 각 각의 평균 감성 점수가 같다는 귀무가설을 유의수준 0.05에서 기각하고 있음을 확인할 수 있다.

< 상자수염>에서 나타난 7개의 언론사 중 지상파를 제외한 4개의 언론사의 평균 감성점수의 분포가 다소 유사하다고 판단하여 '경향신문, 조선일보, 중앙일보, 한겨례 언론사들의 평균 감성 점수는 같다.'라는 귀무가설을 추가적으로 검정하였다. 그 결과는 <표, 그림>에서 나타난 것처럼 유의수준 0.05에서 '평균 감성 점수는 다르다.'라는 대립가설을 기각할 수 있어, 귀무가설을 받아드릴 수 있었다.

group1	group2	null.value	estimate	conf.low, conf.high	p.adj
KBS	MBC	0	-0.0298	-0.0462, -0.0135	5.80e-05
KBS	SBS	0	-0.0732	-0.0935, -0.0529	2.96e-14
MBC	SBS	0	-0.0434	-0.0661, -0.0207	2.18e-0.5

<표 2> HSD 검정 결과 - 지상파(KBS, MBC, SBS)

∠ ∓	2 <	HCD	거저	결과 _	겨햐시므	조선일보.	주아인ㅂ	하거레
$\leq \pm$. > >	עוכח	7177		687 T.	그는 얼마	중하글모.	YI / H LII

group1	group2	null.value	estimate	conf.low, conf.high	p.adj
경향신문	조선일보	0	-0.0062	-0.0210, 0.0086	0.704
경향신문	중앙일보	0	-0.0112	-0.0263, 0.0038	0.220
경향신문	한겨레	0	-0.0007	-0.0198, 0.0183	1.000
조선일보	중앙일보	0	-0.0050	-0.0179, 0.0078	0.745
조선일보	한겨레	0	0.0055	-0.0119, 0.0229	0.849
중앙일보	한겨레	0	0.0105	-0.0071, 0.0281	0.415

5. 결론 및 요약

본 연구는 대규모의 비정형 텍스트 데이터를 활용하여 텍스트 마이닝과 감성 분석을 수행하였다. 이를 통해 코로나19와 관련된 뉴스 기사의 특징과 감성 변화를 탐색하여 주요 결과를 도출하였다.

첫째, 기사 제목에서 빈도가 높은 단어 20개를 통해 코로나19와 관련된 핵심 주제 및

이슈를 확인할 수 있었다. 두 번째로, 긍정과 부정 기사의 개수와 감성의 단어 빈도를 분석하여 각각 10개의 주요 키워드를 도출하였다. 세 번째, 시간의 변화에 따른 감성 변화를 관찰하였다. 코로나19의 발생 초기에는 부정적인 감성이 우세하였으나시간이 지남에 따라 변화하는 양상을 확인하였다. 네 번째로, 시간별로 긍정과 부정단어의 빈도 변화를 분석하여 시기별 감성 단어 빈도를 파악하였다. 다섯 번째, 언론사별로 감성 동향을 파악하고 보도한 기사들의 감성 비율을 확인하였다. 또한 감성점수를 비교하고 Tukey 검정을 통해 가설 검정을 진행하였다. 이를 통해 언론매체간의 감성 점수 차이를 확인하고, 각 언론사의 특징을 파악하는 데 도움을 얻을 수있었다.

본 연구의 결과는 코로나19와 관련된 뉴스 기사의 감성 변화 및 특징을 다각도로 분석하여 이해하는 데 기여할 것으로 기대된다. 그러나 감성사전으로 단어의 긍정과 부정을 구분하였지만, 실제 문맥 속에서 사용되는 의미를 반영하여 구분하는 것이 필요하다. 사용된 단어의 문맥과 사용된 의미의 중요성을 강조하며, 더욱 정밀한 분석을위해 더 다양한 매체와 시기를 고려하는 연구를 추후에 기대한다.

참고문헌

- 1. Dubey AD. Twitter sentiment analysis during COVID-19 Outbreak [Internet]. Amsterda m: Social Science Research Network; 2020 [cited 2022 Jan 5]. Available from: https://papers.ssrn.com/so13/papers.cfm?abstract_id=3572023.
- 2. Li N, Wu DD. Using text mining and sentiment analysis for online forums hotspot de tection and forecast. Decis Support Syst 2010;48:354-368.
- 3. Suh H, So J. A study on the topic and sentiment of national petition data using te xt analysis. Korean Data Anal Soc 2020;22:999-1011.
- 4. Park SM, Na CW, Choi MS, Lee DH, On BW. KNU Korean sentiment lexicon: Bi-LSTM-base d method for building a Korean sentiment lexicon. J Intell Inf Syst 2018;24:219-240.
- 5. 김희정, 권오병 (2023), 뉴스 기사 감성분석을 활용한 고령층 연령차별 양상 및 COVID-1 9 조절효과 연구.한국전자거래학회지,28(1),55-76
- 6. 허성민,양지연,Heo Seong-Min,and Yang Ji-Yeon. "텍스트 분석을 이용한 코로나19 관련 국내 논문의 주제 및 감성에 관한 융합 연구." 한국융합학회논문지 12.4 (2021): 31-42.
- 7. 권나현,오종민,and 하은희. "텍스트 마이닝을 활용한 COVID-19에 대한 대중의 관심 주제와 정서 분석." EMJ (Ewha medical journal) 45.2 (2022): 46-54.
- 8. 문태헌(Moon Tae Heon), and 김혜림(Kim Hye Lim). "신문기사 감성분석과 아파트매매실거 래가의 관계 분석: 부산지역 사례." 주거환경(한국주거환경학회논문집) 18.3 (2020): 135-149.

- 9. 김진순. "감성분석을 중심으로 한 도시브랜드와 관광이미지 연구." 국내석사학위논문 동의대학교 대학원, 2022. 부산
- 10. Heo, S.-M., & Yang, J.-Y. (2021). A Convergence Study on the Topic and Sentiment of COVID19 Research in Korea Using Text Analysis. Journal of the Korea Convergence So ciety, 12(4), 31–42. https://doi.org/10.15207/JKCS.2021.12.4.031