토픽 모델링 기반 한중관계 이슈 분석 :
2019~2023년 온라인 뉴스 댓글을 중심으로

백지영[1]·성민지[2]· 염원[3]·이가흔[4]
[1]성균관대학교, 데이터사이언스융합전공, 2021312576
[2]성균관대학교, 스포츠과학전공, 2020311521
[3]성균관대학교, 데이터사이언스융합전공, 2021312050
[4]성균관대학교, 데이터사이언스융합전공, 2021313368

Analysis of Korea-China Relations Issues Based on Topic Modeling : Focusing on Online News Comment from 2019 to 2023

Jiyeong Baek[1] · Minzy Sung[2*] · Won Yeom[1] · Gaheun Lee[1]

[1]Undergraduate Student, Applied Data Science, Sungkyunkwan University, Seoul, 03063 Korea
[2]Undergraduate Student, Sports Science, Sungkyunkwan University, Seoul, 03063 Korea

요 약

한국과 중국은 전략적 협력동반자관계로서 경제, 정치적으로 서로 가장 중요한 파트너 국가이다. 하지만, 코로나19이후 깊어진 반한/반중의 정서가 깊어지면서 양측의 국민들은 온라인 상에서는 서로를 향한 혐오가 환산되고 있다. 2019년부터 2023년까지 온라인 뉴스 기사 댓글을 분석하여 국민이 바라보는 양국의 관계과 시선은 어떤지 살펴보고자 한다. 토픽모델링 기법 중 LDA를 사용해 댓글은 어떤 토픽으로 구성되어 있는지, 각 토픽별로 어떤 키워드로 이루어져있는지 파악하고자 한다. 이를 통한 키워드 조합으로 각국 국민이 예민하게 반응하는 이슈를 토픽으로 정형화하고자한다. 이는 앞으로 양국의 국민이 중요시 하는 한중관계 이슈를 파악하여 신뢰할 수 있는 연대 방법을 찾아 한중 관계 개선 회복을 위한 방향성을 제시할 것이다.

ABSTRACT

Korea and China are the most important partner countries economically and politically as strategic partnerships. However, as anti-Korean/anti-Chinese sentiment deepened after COVID-19, people on both sides are spreading their hatred for each other online. From 2019 to 2023, we will analyze the comments on online news articles to examine the relationship and perspective of the two countries that the people see. Among the topic modeling techniques, LDA is used to determine what topics the comments are composed of and what keywords are composed of for each topic. Using this technique, the issue that the people of each country react sensitively to is formalized as a topic through a keyword combination. In the future, this will help to identify the issues related to Korea-China relations that the people of the two countries value and find a reliable solidarity method to suggest a direction for improving and recovering Korea-China relations.

키워드 : 중국, 한국, 온라인 뉴스, 자연어, 감성분석, 토픽 모델링

Keywords : China, Korea, Online News, Natural Language, Sentiment Analysis, Topic Modeling

Ⅰ. INTRODUCTION

한중 양국은 상호 간에 경제적으로 가장 중요한 파트너 국가이며, 정치적으로는 '전략적 협력동반자 관계'라는 최상위 외교 관계에 있다[1]. 우리 국민 대다수도 '중국은 싫지만 한중관계는 중요하다'는 인식을 가지고 있었다.[2] 중국도 10명 중 4명을 한국을 비호감[3]이라 여겼지만, 현재 중국은 미중분쟁으로 서방의 군사안보와 글로벌 공급망 측면에서 반도체 수출통제와 같은 강력한 견제를 받고 있는 상황에서 한국과의 관계 개선의 필요성이 커졌다. 이와 같은 상황에서 중국의 경제, 국제 관계에 있어서 한국과 우호적인 관계가 유지되는데 있어 중요하다.[4] 그러나 국제정세에서 중요한 현안과 국민이 통감하는 문제는 다를 수 있다. 지금까지 한중관계에 관한 주요 연구는 기자회견 텍스트[5]나 뉴스 기사[6]를 중심으로 이루어져 왔다. 본 연구는 2019년부터 2023년까지 5년간의 온라인 뉴스 기사 댓글을 분석하여 국민이 집중했던 주요 이슈는 무엇이며, 어떤 특징이 있는지 살펴보고자 한다. 특히 온라인 상 부정적인 댓글에 나타난 혐오표현과, 이와 관련있는 한중 이슈에 대해 초점을 두고자 한다. 이를 통해 혐오표현이 자주 일어나는 정치, 경제, 문화 이슈를 알 수 있을 뿐만 아니라, 각 이슈를 이해하는 한중 양국 국민의 입장 차이에 대해 살펴보고자 한다. 이는 앞으로 한국/중국 국민 간의 의견 차이를 이해하여 서로를 존중하며 이해하며 연대하는 방법에 대해 논할 수 있을 것이다.

Ⅱ. DATA SET

본 연구의 분석대상은 한국과 중국이 상대 국가에 관해 작성한 온라인 뉴스 기사의 댓글이다.

2.1 한국 데이터 수집

기사 댓글을 수집하기 위한 포털은 네이버 뉴스로 결정하였다. 이것은 우리나라 포털 사이트 이용자의 92.1%가 뉴스를 네이버를 통해 얻고 있기[7] 때문이다. 본 연구에서는 댓글을 수집하기 위해 먼저 기사 URL을 수집했다. 온라인 뉴스 기사를 수집하기 위한 검색어로는 '중국'을 사용했으며, 기간은 2019년 1월 1일부터 2023년 12월 31일까지의 5년으로 제한했다. 이때, 네이버 뉴스 알고리즘 상 검색어 '관련도순'으로 설정하여도, 이용자에게 최신 기사를 주로 제공한다. 따라서 설정한 검색 기간의 후반 기사만 수집되는 문제가 발생하였다. 특정 시간대에 편향된 데이터를 얻는 것을 방지하기 위해, 기간을 한 달 단위로 나누어 수집을 진행하였다. 설정한 기간 내에서 검색어를 입력하였을 때, 7번 스크롤한 후의 화면에서 기사를 수집하였다. 어떤 기사 하단에는 관련 기사를 함께 제공하는데, 이런 경우 관련 기사의 URL까지 수집하였다. 이를 통해 달별 100~150개, 총 7786건의 데이터를 얻을 수 있었다. 이후 Python의 Selenium, BeautifulSoup 패키지를 이용해 수집된 기사 링크 안에 게재된 기사 제목, 작성일, 언론사, 댓글을 웹 크롤링(Web Crawling)하였다.

2.2 중국 데이터 수집

중국의 기사 댓글을 수집하기 위한 포털은 Sina Weibo(시나 웨이보, 이하 웨이보)로 결정하였다. 륙치금 외 1인(2021)에 따르면, 이는 중국 내에서 가장 많이 방문하는 웹 사이트/플랫폼이자, 중국의 버전의 X(구 트위터)라 불린다고 한다. 웨이보는 온라인 상에서 감정 표현과 소통을 하는 핵심 통로로서 이용자들이 정보를 기록하고 소통을 하여 오락과 사회적 교류를 할 뿐만 아니라 인기 뉴스의 탄생을 촉진하는 매체로서 작용한다. 따라서 중국 대중이 사회활동에 참여하고 의견을 표현할 수 있는 공간이기에 '웨이보'를 댓글 수집을 위한 포털로 선정하였다. 본 연구에서는 검색어 키워드로 '韩国(한국)'을 사용하였으며, 기간은 2019년 1월 1일부터 2024년 3월 25일로 최근 5년으로 제한했다. Python의 Pymongo, Requests, BeautifulSoup 패키지를 이용하여 한국 관련 게시물과 댓글을 웹 크롤링(Web Crawling)을 하였다. 이때 웨이보 계정의 쿠키 정보를 이용한 데이터 수집 시, 매일 50페이지이내의 공식 신문사계정에서 게시되는 뉴스 게시물만 모을 수 밖에 없는 제약이 존재하였다. 따라서 월별로 일정한 게시물을 추출할 수 없는 한계가 존재하였다. 웹 크롤링 결과, 937개의 게시물 데이터를 얻을 수 있었다.

2.3 데이터 전처리

네이버에서 수집한 데이터에서 이상값이나 불일치는 존재하지 않았다. 다만, 댓글이 없는 기사는 댓글 데이터가 빈 리스트 형태로 존재하였다. 본 연구의 분

석대상은 댓글이기 때문에 데이터가 없는 행은 모두 제거하였다. 해당 과정 이후 얻은 댓글 데이터는 15,177건이다. 이후 자연어로 작성된 댓글과 기사 제목에 전처리를 진행하였다. 줄 바꿈, 이모티콘 및 문장 기호, 양옆 공백 문자열을 제거하여 분석하기 적절한 형태로 정제하였다. 전처리 코드는 다음과 같다.[1)]

시나 웨이보에서 수집한 데이터에서는 이상값, 불일치, 결측치가 존재하지 않았다. 수집된 게시물에 작성된 댓글은 중복된 데이터를 제거하여 총 13818개이다. 이후 네이버 수집 댓글 전처리 과정과 동일하게 댓글에 포함된 이모티콘, 비표준 문자를 제거하여 분석하기 적절한 형태로 정제하였다. 전처리 코드는 다음과 같다. [2)]

Ⅲ. EDA

3.1 네이버 데이터 시각화

3.1.1 언론사별 기사 수 막대그래프

한국 언론사별 기사 수를 막대그래프로 표현하여, 어떤 언론사의 기사가 많이 수집되었는지 파악하고자 했다. 너무 많은 막대가 그려지는 것을 방지하기 위하여, 기사 수가 10개인 언론사는 제거하였다. 그림[3)]을 통해 연합뉴스가 총 1,009건으로 가장 많은 기사를 가지고 있었으며, 연합뉴스, 뉴시스, 뉴스1 등 뉴스통신사의 기사가 많이 수집되었다는 것을 확인할 수 있었다.

3.1.2 기사 제목 워드클라우드

네이버 기사 제목을 워드클라우드로 표현하기 위해, 토큰화한 제목에 품사를 태깅하고 명사만 추출하였다. 이 과정에서 '중국', '중국인'은 불용어 처리하였다. 검색어를 '중국'으로 하여 자료를 수집하였기 때문에, 데이터 내에 다수 존재하나 분석할 이유가 없기 때문이다. 그림[4)]을 통해 표1과 같은 결과를 확인할 수 있다.

<표 1> 워드클라우드에서 나타나는 단어

구분	단어
국가명	미국, 한국, 대만, 일본, 러시아, 홍콩
사건	코로나, 전쟁, 시위
경제 관련	경제, 반도체, 기업, 시장, 수출, 투자
정치 관련	시진핑, 트럼프

3.1.3 댓글 워드 클라우드

위와 같은 과정을 거쳐 네이버의 댓글 내용을 워드클라우드로 표현하였다. 불용어는 그림을 확인하며 필요하지 않은 조사, 접속부사, 감탄사 등을 직접 제거하였다. 그림[5)]을 통해 표2와 같은 결과를 얻었다. '중공'과 '조선족', '바퀴벌레'를 혐오 표현으로 분류하였는데, 실제 의미와 달리 멸칭으로 사용되고 있었기 때문이다. 표3에 있는 예시 댓글을 통해 확인할 수 있다.

<표 2> 워드클라우드에서 나타나는 단어

구분	단어
국가명	미국, 한국, 북한, 일본, 홍콩, 러시아
사건	미세먼지, 바이러스, 전쟁
정치 관련	시진핑, 문재인, 트럼프, 민주당
혐오 표현	중공, 짱깨, 조선족, 바퀴벌레

<표 3> 혐오적인 표현이 포함된 예시 댓글

단어	예시
중공	- 하여튼 중공놈들은 민패덩어리임 - 알면 알수록 싫어진다 중공놈들
조선족	- 조선족 추방해야 대한민국 광명찾는다 - 그 직원 조선족이냐 당장 짤라라
바퀴벌레	- 중국은 바퀴벌레의 탈을 쓴 민족 - 전세계 왕따 바퀴벌레국

3.1.4 감성분석

감성분석을 진행한 코드는 다음과 같다.[6)] 하나의 댓글을 형태소로 토큰화한 후, KNU 한국어 감성 사전

을 이용하여 극성값들의 합을 계산하였다.

3.1.4.1 기사 제목

그림[7]은 기사 제목을 대상으로 한 감성분석 결과를 파이 그래프로 시각화한 것이다. 중립이 48.8%로 가장 높은 비율을 차지하는 것으로 보아, 언론의 중립성이 지켜지고 있음을 알 수 있다. 그러나 중립을 제외한 긍/부정의 비율만을 살펴보았을 때는 부정이 36.1%로 15.2%인 긍정의 2배 이상을 차지하였다.

3.1.4.2 기사 댓글

그림[8]은 기사 댓글을 대상으로 위와 같은 과정을 진행한 결과다. 부정이 44.1%로 가장 높은 비율을 차지하였고, 긍정의 비율은 12.8%로 기사 제목을 분석하였을 때보다 더 줄어든 것을 확인할 수 있었다. KNU 한국어 감성 사전을 이용한 한계로써, 사전에 포함되지 않은 '짱깨'와 같은 혐오 표현의 극성값은 계산되지 않았다. 따라서 중립의 비율이 43.1%로 계산되었으나, 실제로는 부정적인 댓글의 비율이 훨씬 더 높을 것이라고 예상할 수 있다.

3.1.4.3 기사 제목 워드클라우드

감성분석을 통해 긍정과 부정으로 판단된 기사 제목만을 이용하여 각각 워드클라우드로 그렸다.[9] 긍정적인 제목은 초록색으로, 부정적인 제목은 빨간색으로 표현했다. 각 워드클라우드에서 나타난 의미 있는 단어는 표4와 같다. 국가명이나 정치인의 이름, 코로나와 같은 사건은 감성분석 결과에 상관없이 확인할 수 있었다.

<표 4> 긍정/부정 워드클라우드에서 나타나는 단어

감성	구분	단어
긍정	긍정어	인기, 최고, 기대
	경제 관련	이익, 투자, 지원
	정치 관련	나토, 협정
부정	부정어	비판, 갈등, 봉쇄, 논란, 위기
	경제 관련	화웨이, 반도체, 불법조업
	정치 관련	입국, 남중국해

3.1.4.4 기사 댓글 워드클라우드

위와 같은 과정을 이용하여 워드클라우드로 표현하였다.[10] 그러나 KNU 감성 사전의 한계로 인해, 제목을 이용한 시각화보다 명확한 구분이 이루어지지 않았다. 각 그림에서 나타난 단어는 표5와 같다.

<표 5> 긍정/부정 워드클라우드에서 나타나는 단어

감성	구분	단어
긍정	긍정어	최고, 인정, 발전, 도움
	경제 관련	이익, 반도체, 삼성, 기술
	정치 관련	공산당, 대통령
부정	부정어	재앙, 문제, 피해
	사건	코로나, 우한폐렴, 미세먼지
	정치 관련	조선족, 중공, 입국, 추방

3.2 웨이보 데이터 시각화

3.2.1 언론사별 게시물 수 막대그래프

웨이보 언론사 계정별 게시물 수를 막대그래프로 표현하여, 어떤 언론사의 기사가 많이 수집되었는지 파악하고자 했다.

그림[11]을 보면, 상위 5개의 신문사는 环球时报(환구시보), 每日经济新闻(매일경제신문), 央视财经(CCTV 재경, 중국국가방송의 금융채널), 看看新闻 KNEWS(칸칸신문, 이하 KNEWS), 澎湃新闻(영문 : The Paper, 상하이 미디어 그룹 산하의 인터넷 뉴스 미디어) 순으로 기사가 많이 수집되었다.

3.2.2 게시물 내용 워드 클라우드

웨이보의 게시물 내용을 워드클라우드로 표현하기 위해, 중국어 형태소 분석기인 jieba를 이용해 토큰화하였다. 그 다음, 중국어 특징에 따라 불용어를 식별하기 어려워 하얼빈 공업대학에서 제작한 불용어 단어장인 哈工大停用词表(HIT_stopwords)를 이용해 불용어를 제거하였다. 또, 검색어를 '한국(韩国)'으로 하여 자료를 수집했기에 데이터 안에 포함된 한국(韩国)이란 단어도 불용어로 취급하였다. 추가로, 그림[12] 의 워드클라우드 결과를 확인하며 필요하지 않은 단어를 불용어 단어장에 추가하여 직접 제거한 결과는 표6과 같다.

<표 6> 게시물 - 워드클라우드에서 나타나는 단어

구분	단어
국가 및 수도	中国(중국), 美国(미국), 日本(일본), 朝鲜(북한), 首尔(서울), 北京(베이징)
문화	泡菜(파오차이), 视频(동영상), 媒体(미디어)
경제	显示(디스플레이)
사건	新冠(신종코로나), 疫情(전염병), 肺炎(폐렴), 口罩(마스크), 感染(감염), 确诊(확진), 韩国队(한국대표팀), 核污染(핵/원자력 오염), 疫苗(백신), 比赛(시합)
정치 관련	尹锡悦(윤석열), 文在寅(문재인), 朴槿惠(박근혜), 韩国政府(한국정부), 外交部(외교부)

3.2.3 댓글 워드 클라우드

웨이보의 댓글 내용을 워드클라우드로 표현하기 위해, 앞선 게시물 내용 워드 클라우드 표현 과정과 동일한 과정을 거쳤다. 그 결과는 그림 13)과 같다. <표7>을 보면 댓글에 코로나19 관련 단어들과 '중국, 일본, 미국' 국가명이 공통적으로 두드러지게 나타난 것을 확인할 수 있다.

<표 7> 댓글 - 워드클라우드에서 나타나는 단어

구 분	단어
국가명	中国(중국), 日本(일본), 美国(미국), 朝鲜(북한),
비하명사	棒子(한국인 비하)
부정동사	没有/没(없다), 不要(하지마세요), 不会/不能(할 수 없다)
형용사	真的(정말로)
문화	泡菜(파오차이), 申遗(문화유산등록), 辣白菜(매운배추김치, 김치와 같은 단어), 文化(문화)
사건	疫情(전염병), 邪教(사이비), 病毒(바이러스)

3.2.4 감성분석 결과

3.2.4.1 게시물

게시물 내용에 대해 감성분석한 결과14), 긍정은 43%, 부정은 41.1%, 중립은 15.9%로 긍정, 부정, 중립 순으로 나타났다. 부정과 긍정을 띠었던 게시물 내용에 대해 워드 클라우드로 나타낸 결과는 그림15) 와 같다.

<표 8> 웨이보 게시물 - 부정 워드클라우드에서 나타나는 단어

구 분	단어
국가	韩国(한국), 美国(미국), 日本(일본), 首尔(서울)
동사	发生(발생하다), 要求(요구하다)
정치관련	尹锡悦(윤석열), 在野党(야당에서)
사건	新冠(코로나19), 肺炎(폐렴), 感染(감염), 确诊(확진), 核污染(핵오염), 疫情(전염병진행상황)
문화	韩媒(한국미디어), 视频(영상)
부정적 어휘	死亡(사망), 抗议(항의), 没有(없다)

<표 9> 웨이보 게시물 - 긍정 워드클라우드에서 나타나는 단어

구 분	단어
국가	中国(중국),日本(일본), 美国(미국), 朝鲜(북한), 北京(베이징), 首尔(서울)
주요동사	举行(개최하다), 可能(가능하다)
사건	韩国队(한국대표팀), 比赛(시합), 疫情(코로나진행상황)
문화	韩媒(한국미디어), 韩联社(한국연합뉴스)
정치	尹锡悦(윤석열)

위 <표 8>과 <표9>를 보면 공통적으로 일본, 미국과 같은 주요 외교국가들의 국가명과 정치인의 이름이 나타났다. 또, 부정 워드 클라우드에서 코로나19과 연관된 단어들이 많이 나타났다. 반면, 긍정 워드 클라우드에 스포츠(올림픽, 월드컵, 아시안컵)과 같은 경기와 연관된 단어들이 나타난 것을 발견하였다.

3.2.4.2 댓글 감성분석

댓글에 대해 감성분석한 결과인 그림16)을 보면, 긍정 21.8%, 부정은 34.3%, 중립은 43.8%로 부정적인 댓글의 비율이 긍정적인 댓글의 비율보다 높았다. 댓글에 대한 긍/부정 워드 클라우드17) 를 분석한 결

과인 <표10>과 <표11>를 보면, 앞선 게시물 워드클라우드와 공통적으로 감성분석 결과와 상관없이 코로나19와 관련된 단어들이 나타났다. 또, 파오차이/김치와 같은 문화 분쟁과 관련 단어들도 긍/부정 상관없이 두드러지게 나타난 것을 발견했다.

<표 10> 웨이보 댓글 - 부정 워드클라우드에서 나타나는 단어

구 분	단어
국가	中国(중국), 韩国(한국) , 日本(일본), 美国(미국)
비하단어	棒子(한국인 비하)
동사	死亡(사망), 不能/不会(할 수 없다), 没有/没(없다), 可能(가능하다)
형용사	可怕(두렵다)
문화	泡菜(파오차이, 절임채소음식), 申遗(세계문화유산을 신청하다)
기타	狗(개, dog)

<표 11> 웨이보 댓글 - 긍정 워드클라우드에서 나타나는 단어

구 분	단어
국가	中国(중국), 韩国(한국) , 日本(일본), 美国(미국)
표현	哈哈哈(웃음소리, “하하하”)
동사	确实(확실하다), 可以/可能(가능하다), 支持(지지하다), 希望(희망하다), 爱(사랑하다), 觉得(~라고 느끼다),
형용사	真的(정말로), 很多(많다), 好吃(맛있다)
문화	泡菜(파오차이, 절임채소음식)

Ⅳ. METHODOLOGY

댓글 데이터는 맞춤법이 제대로 지켜지지 않거나, 신조어나 비속어를 사용한 경우가 많다. 문장을 형태소로 토큰화한 데이터에 품사를 태깅하고, 특정 품사의 단어들을 눈으로 직접 확인할 것이다. 기존에 사용한 감성사전에 추가적인 긍정어와 부정어를 더함으로써, 분석에 필요한 최종 사전을 구축할 것이다.

이후, 긍정 댓글과 부정 댓글 각각 LDA 분석을 수행할 것이다. LDA 분석은 어떤 토픽이 어떤 비율로 구성되어있는지 분석하고, 해당 토픽이 어떤 키워드로 구성되었는지 정보를 제공한다. 따라서 키워드 조합을 통해 인사이트를 도출할 수 있다는 장점이 있다. 양국 국민이 예민하게 반응하는 이슈를 파악하기에 효과적인 분석 방법이라고 생각하여 선택하게 되었다. LDA 기법에서 α값, β값은 하이퍼파라미터(hyperparameter)로, α값은 토픽의 문서 밀집도를 결정하고, β값은 키워드의 토픽 밀집도를 결정한다. [8]

V. CONCLUSION

본 연구는 지금까지 이루어진 연구와 달리 양국 국민이 중요시하는 한중관계 이슈를 파악할 수 있을 것이라는 의의가 있다. 정부는 이를 통해 양국 국민의 상호신뢰관계 회복을 위해 노력할 수 있을 것이다. 그러나 다음과 같은 한계도 존재한다.

5.1 네이버 기사 데이터의 편향

네이버 데이터를 수집할 때 여러 편향이 발생하였다. 우선 네이버 뉴스 알고리즘은 사용자에게 최신 기사를 주로 제공하여, 검색 기간의 후반 데이터만 주로 수집되었다. 이는 그림[18]을 통해 월말 기사만 수집된 것을 명확히 볼 수 있다. 또한, 보수 성향의 언론사가 주로 수집되었다. 이와 같은 편향으로 인해 결과를 해석할 때 오류가 발생할 수 있다.

5.2 웨이보 게시물 데이터의 편향

웨이보 게시물 데이터를 수집 시 특정 월에 집중되는 편향 문제가 발생했다. 웹 크롤링 시, 하루에 가져올 수 있는 게시물의 양이 정해져있어 각 월별마다 동일한 게시물의 수를 수집할 수 없어 특정 월에 수집된 게시물이 많은 편향이 일어나기도 하였다.

5.3 사용자에 대한 인구통계학적인 정보 부재

댓글 작성자에 대한 성별, 연령대, 직업, 정치적 성향 등의 정보는 크롤링으로써 얻을 수 없었다. 연령대나 정치적 성향에 따라 관심 있는 주제가 매우 달라지며, 온라인 뉴스에 댓글을 남기는 비율도 다를 것이다. 따라서 부정적인 댓글이 많이 달렸던 토픽이라고 하여 국민 전체의 생각을 대표하기 어렵다는 한계가 있다.

5.4 무분별한 혐오 표현

기사 제목이나 내용의 감성이 기사 댓글에 영향을 준다는 가정을 기반으로 산점도를 그려보았고[19], 서로 연관이 없다는 것을 확인하였다. 이는 기사와 관련

없는 혐오 표현이 다수 존재하기 때문이다. 따라서 기사 댓글을 대상으로 토픽모델링을 진행하였을 때도, 명확한 토픽이 결과로 나오지 않을 수 있다. 이런 경우 부정적인 댓글이 일정 비율 이상 달린 기사 내용을 LDA 분석하여 어떤 주제에 국민이 반응하는지를 파악할 것이다.

5.5 중국어 구문

중국어의 경우, 단어가 단독으로 사용되어, 명사와 동사의 의미를 뜻할 뿐만 아니라, 두 개 이상의 단어가 결합되어 의미를 이루기 때문에 이를 명확히 파악하여 정확한 토큰화의 결과를 얻기에 한계가 있었다. 이를 방지하기 위해 불용어 단어장을 활용하여 정확한 토큰화를 진행하고자 하였지만, 너무 많은 불용어의 취급 의미상 중요한 단어도 불용어를 취급할 수 있다는 문제점이 발생하였다.

REFERENCES

[1] 황태연, 김태종. (2023). 토픽 모델링 기반 한중관계 이슈 분석:1990~2022년 뉴스 빅데이터를 중심으로. 평화학연구, 24(1), 91-144쪽.

[2] 동아시아연구원(EAI). 2023년 EAI 동아시아 인식조사 (2): 중국과 한중관계. 2023.09. https://www.eai.or.kr/new/ko/pub/view.asp?intSeq=22109&board=kor_issuebriefing&keyword_option=&keyword=&more=

[3] 박혜진. (2023.05.30) 중국인 10명 중 4명 "한국 비호감"… 만만찮은 반한 정서. KBS https://news.kbs.co.kr/news/pc/view/view.do?ncd=7687461 (접근 2024.05.13.)

[4] 김환용(2024.04.30.). Voa뉴스. 한국-중국 관계 물꼬 트일까? https://www.voakorea.com/a/7591608.html (접근 2024.05.13)

[5] 이상국. (2022). 2011년-2021년 기간 중국 안보·국방 토픽 분석 - 중국 국방부 정례기자회견 텍스트에 대한 동적 토픽 모델링 적용. 국방연구 65(3), 65-96쪽.

[6] 정원준. (2018). 사드(THAAD) 이슈를 둘러싼 한국과 중국 간 갈등 쟁점의 변화 추이 연구. 한국광고홍보학보, 20(3), 143-196, 10.16914/kjapr.2018.20.3.143

[7] 한국언론진흥재단. 2023 언론수용자 조사. 2023.12, 101쪽. https://kpf.or.kr/synap/skin/doc.html?fn=1706080724689.pdf&rs=/synap/result/research/

[8] David M. Blei, Andrew Y. Ng and Michael I. Jordan, (2003). "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, 993-1022.

APPENDIX

1) 자연어 전처리 코드

```
#텍스트 정제 함수
import re
import ast

def clean_reviews(data):
    #문자열을 리스트로 변환
    result = ast.literal_eval(str(data))

    #정제 작업
    cleaned_list = []
    for item in result:
        #줄바꿈 제거
        cleaned_item = re.sub(r'₩n+', ' ', item)

        #이모티콘 및 문장 기호 제거
        cleaned_item = re.sub(r'[^₩w₩s]', '', cleaned_item)

        cleaned_item = cleaned_item.strip()
        #공백 문자열이 제거
        if cleaned_item:
            cleaned_list.append(cleaned_item)

    return cleaned_list
print(clean_reviews(df_na.reviews[2]))
```

['중국 사람을 고용하지 말라고 했자나', '우리나라도 마찬가지 중국인이나 조선족 중국에 관련된 사랑은 절대로 중요한 문서나 역활을 맡기지 말라', '글로벌 산업 중국 놈들 악의 축이구나 몇년씩 걸려 진행한 프로젝트를 하루사이 꿀꺽 중국 첨단산업의 비약적 발전은 도둑질로 이룬 산업', '그냥 중국인은 고용안하는게 답이구만', '중국이 찜찍으면 어케든 델꼬 간다이미 넘어간 업체 수두룩 수년내에 수면위에 올라오면 손 못 쓰지 대기업은 정보유출조심해야 안망함국가도 마찬가지']

2) 중국어 웨이보 게시물 댓글 전처리 코드

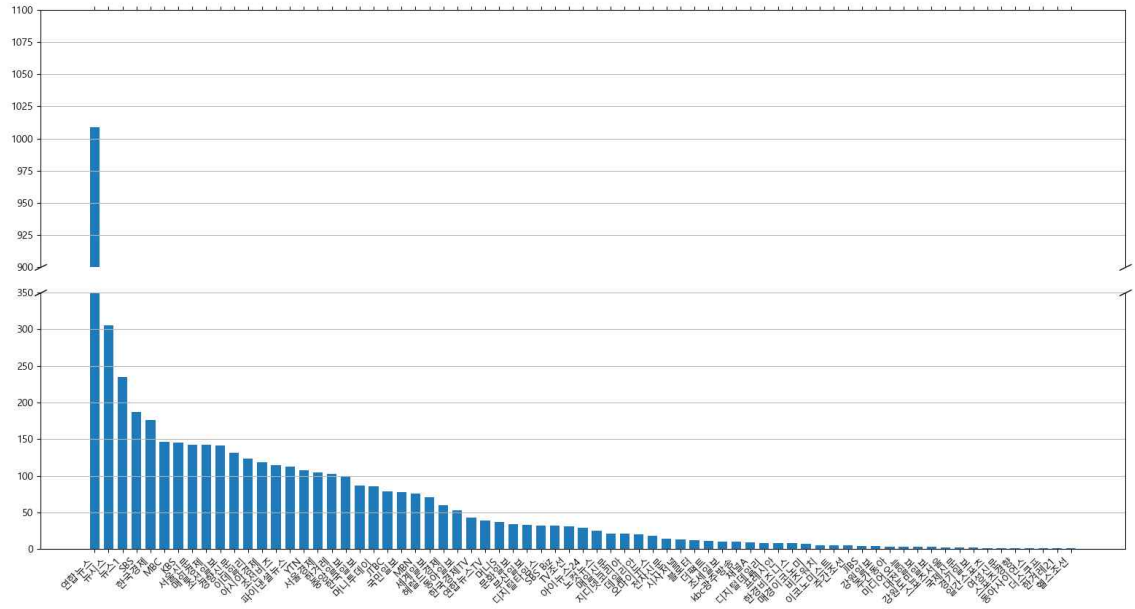
```
# 이모지와 비표준 문자를 제거
def remove_emojis_and_nonstandard(text):
    # Keep only letters, numbers, and standard punctuation
    cleaned_text = re.sub(r'[^\w\s,.!?;:]', '', text)
    return cleaned_text

# 'comment_text' 열을 정리하기 위해 함수를 적용
df['comment_text'] = df['comment_text'].apply(remove_emojis_and_nonstandard)

# 빈 문자열만 있는 행을
df = df[df['comment_text'].str.strip().astype(bool)]
df.to_excel('/content/cleaned_comments.xlsx', index=False)

# 데이터프레임을 동일한 Excel 파일에 저장하여 기존 데이터를 덮어씁니다.
excel_file_path = 'cleaned.xlsx'  # Replace with your file's path
df.to_excel(excel_file_path, index=False)
```

3) 네이버 언론사별 기사 수 막대그래프



4) 네이버 기사 제목 워드클라우드

5) 네이버 기사 댓글 워드클라우드

6) 네이버 감성분석 코드

감성분석

```
# 감성분석에 이용할 감성어 사전
import json
import pandas as pd

# 해당 파일은 https://github.com/park1200656/KnuSentiLex/tree/master/KnuSentiLex/data 에서 설치 필요
with open('/content/drive/MyDrive/Capstone_Project_YW/SentiWord_info.json', encoding='utf-8-sig', mode='r') as f:
  SentiWord_info = json.load(f)

sentiword_dic = pd.DataFrame(SentiWord_info)
```

기사제목 감성 분석

```
# 기사제목 칼럼
title = data['title']
```

```
# 형태소 토큰화
from konlpy.tag import Mecab

mecab = Mecab()
tokens = [mecab.morphs(word) for word in title]
tokens = list(map(lambda x : " ".join(x), tokens))
tokens[:10]
```

```
["글로벌 1 · 2 위 합친 ' 메가 조선사 ' 탄생 … 중국 맹추격 따돌린다",
 '자율 주행 차 기밀 훔쳐 중국 업체 에 이직 하 려고 한 애플 직원',
 '애플 자율 주행 차 기밀 훔쳐 중국 업체 로 이직 하 려 한 직원 기소 돼',
 "또 터진 애플 ' 스파이 ' 사건 … 중국인 , 자율 주행 차 기밀 절 도 혐의",
 '중국 에 자율 주행 차 기밀 빼돌리 려던 애플 직원 기소',
 '황치열 " 중국 공기 안 좋 다 " 한마디 에 中 누리꾼 들 댓글 폭탄',
 '대만 지지 드러냈 나 … 美 백악관 지도 , 중국 과 별도 로 대만 표시',
 '백악관 지도 중국 · 대만 별도 표기 파장 … 지원 책 일환',
 '중국 서 가장 흔한 성 ( 姓 ) 은 王 씨 … 왕 씨 만 한국 인구 2 배',
 '미국 · 중국 무역 담판 사작 … 전 세계 촉각']
```

```
import pandas as pd
from tqdm import tqdm

df_title = pd.DataFrame(columns=["title", "sentiment"])  # 리뷰별 극성을 저장하기 위한 데이터프레임 생성
idx = 0                                                  # 다음 리뷰로 넘기기 위한 초기값

for token in tqdm(tokens):                                    # 전체 리뷰에서 문장 하나씩 가져옴
  sentiment = 0                                          # 초기 감성값 0으로 설정
  for i in range(0, len(sentiword_dic)):                 # 감성사전의 모든 단어를 하나씩 선택
    if sentiword_dic.word[i] in token:                   # 리뷰 문장에 감성 단어가 있는지 확인
      sentiment += int(sentiword_dic.polarity[i])        # 감성단어가 있다면 극성값 합계를 구함
  df_title.loc[idx] = [token, sentiment]                      # 리뷰별 극성값을 데이터프레임으로 저장
  idx += 1                                               # 다음 리뷰 문장으로 넘어감
```

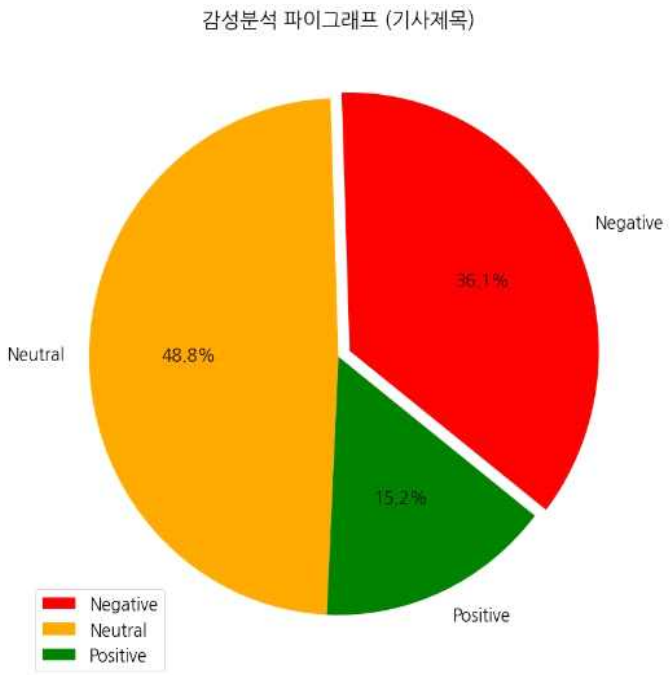
```
100%|██████████| 4806/4806 [10:27<00:00,  7.66it/s]
```

```
data['title_sentiment'] = ''
for i in range(len(data)):
  data['title_sentiment'][i] = df_title['sentiment'][i]
data.head()
```

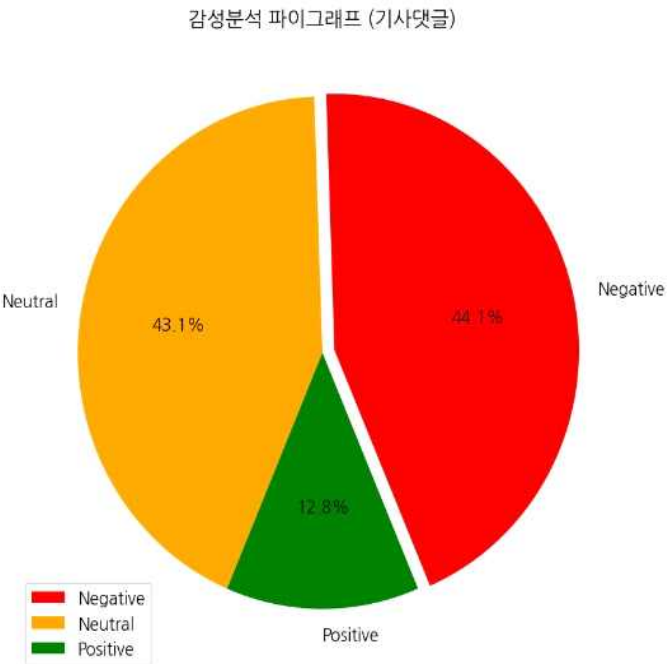
	title	date	company	cleaned_reviews	classification	title_sentiment
0	글로벌 1·2위 합친 '메가 조선사' 탄생…중국 맹추격 따돌린다	2019-01-31 00:00:00	한국경제	['노조 파업해 평생 자손까지 직장 보장 놀아도 회사 망해도 국가가 연봉 1억 보장...	[1, 1, 0]	-2
1	자율주행차 기밀 훔쳐 중국 업체에 이직하려고 한 애플 직원	2019-01-31 00:00:00	서울신문	['앞으로 미국에선 중국애들 직원으로 절대 안뽑을듯', '한국계가 이러지 않았으면 ...	[0, 0]	-1
2	애플 자율주행차 기밀 훔쳐 중국 업체로 이직하려 한 직원 기소돼	2019-01-31 00:00:00	YTN	['중국 사람을 고용하지 말라고 했자나', '우리나라도 마찬가지 중국인이나 조선족 ...	[0, 0, 1, 1, 0]	-1
3	또 터진 애플 '스파이' 사건…중국인, 자율주행차 기밀 절도 혐의	2019-01-31 00:00:00	아시아경제	['중국인 직원들은 위장 취업한 산업스파이로 보면 정확하다', '아마 이사건으로 중...	[0, 1, 0]	0
4	중국에 자율주행차 기밀 빼돌리려던 애플 직원 기소	2019-01-31 00:00:00	연합뉴스	['전세계의 해충이네', '이런걸 보면 중국인 유학생이나 직원을 채용을 안해야하는거다']	[1, 0]	0

```
data.to_csv('/content/drive/MyDrive/Capstone_Project_YW/data_title_sentiment.csv')
```

7) 네이버 기사 제목 감성분석 결과에 따른 기사 수 파이 그래프



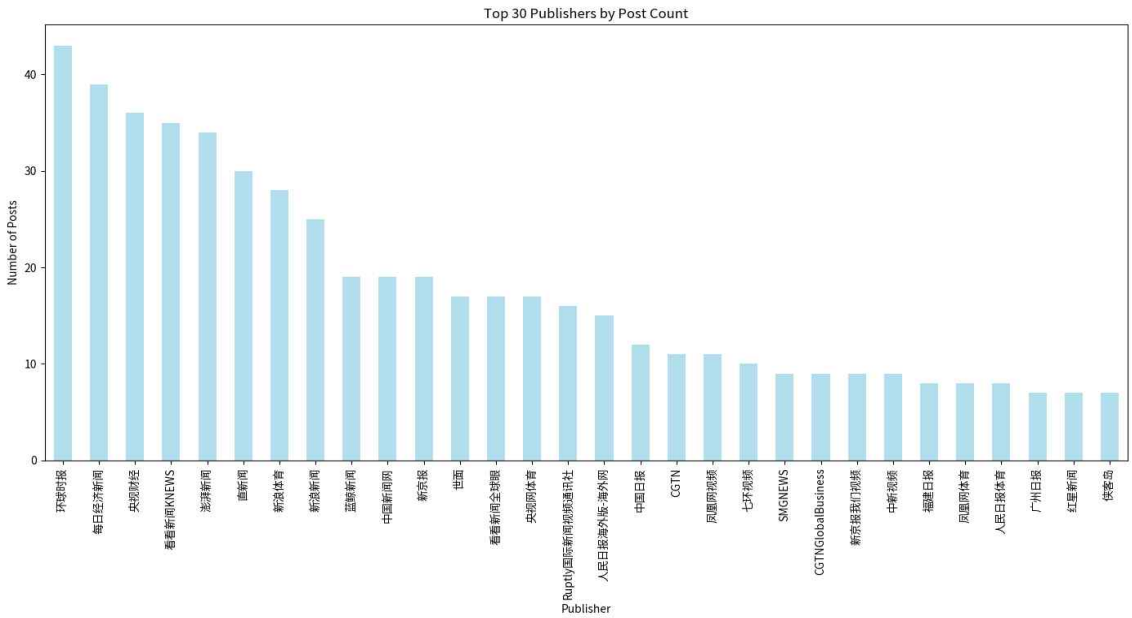
8) 네이버 기사 댓글 감성분석 결과에 따른 댓글 수 파이 그래프



9) 네이버 기사 제목 긍/부정 워드클라우드

10) 네이버 기사 댓글 긍/부정 워드클라우드

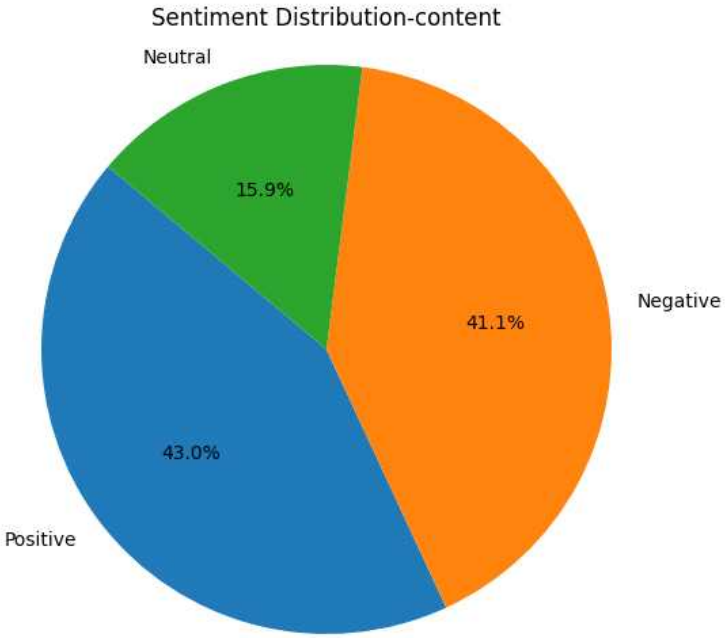
11) 웨이보 언론사별 기사 수 막대 그래프



12) 웨이보 게시물 워드 클라우드

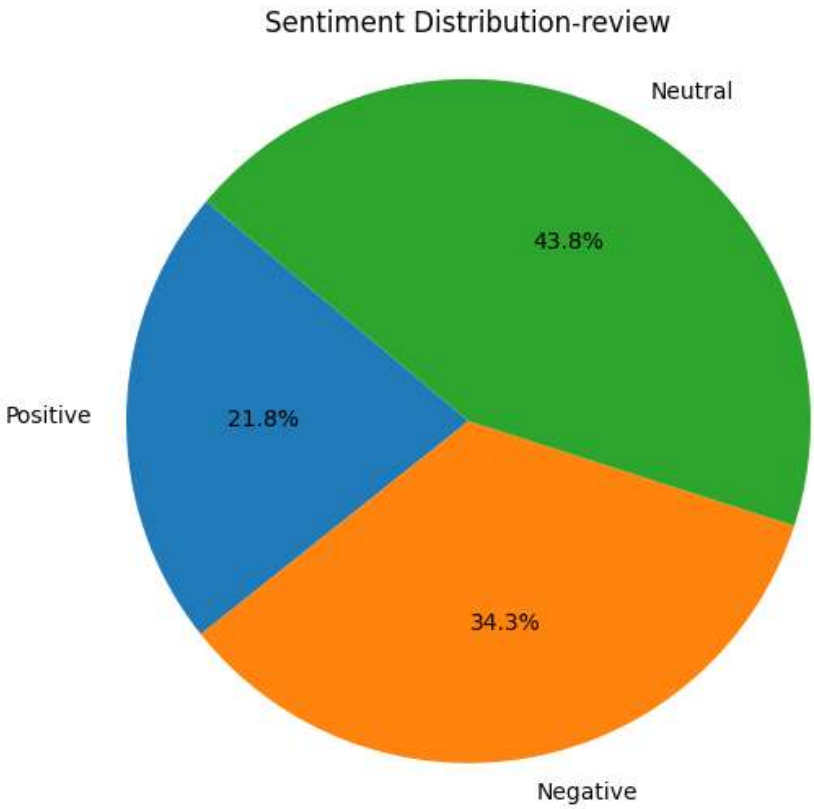
13) 웨이보 댓글 워드 클라우드

14) 웨이보 게시물 감성분석 결과에 따른 긍정, 부정, 중립 비율 수 파이 그래프



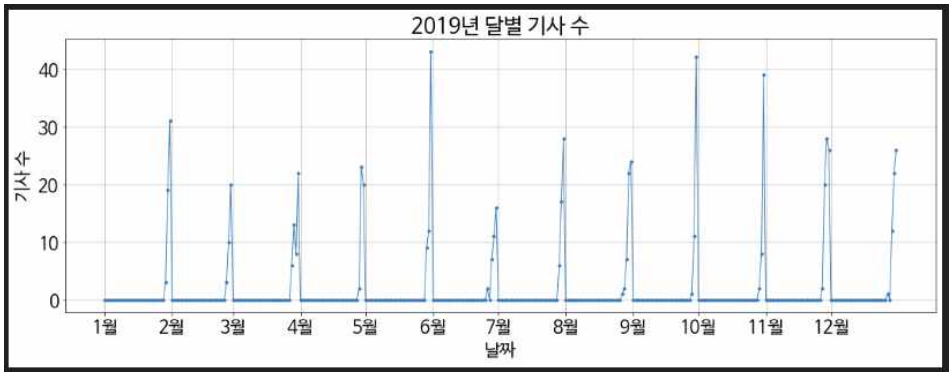
15) 웨이보 게시물 긍/부정 워드클라우드

16) 웨이보 댓글 감성분석 결과에 따른 긍정, 부정, 중립 비율 수 파이 그래프



17) 웨이보 댓글 긍/부정 워드 클라우드

18) 2019년 월별 네이버 뉴스 기사 수 그래프



19) 기사 제목/내용에 따른 기사 댓글의 감성 분석 결과 (한국/중국)