日本語學硏究 第65輯(2020. 9. 20. pp.25~43) http : //dx.doi.org/10.14817/jlak.2020.65.25

텍스트마이닝 기법을 활용한 일본 미디어의 한국 뉴스의 감정 추이에 대한 분석*

-파이썬을 활용한 단어감정극성대응표 분석기법의 수행을 통해-

金曙泳**

< Abstract >

Analysis of the Japanese media's emotions about Korea-related news using text mining techniques; by implementation the Semantic Orientations of Words analysis method using Python

In the era of big data, it is no longer possible to effectively collect, refine, and perform meaningful interpretations of necessary information only by methods such as conventional manual and intuitive insight. Therefore, in this paper, we built a large-scale text data independently and performed the text mining analysis based on the recognition that the text mining technique needs to be used in the field of Japanese studies as well, and that the verification of the practical technique itself is also necessary.

As a result, it was confirmed that an analysis of semantic orientations of the text using 'the Semantic Orientations of Words' is effective for the analysis of the 'emotion' of the text. Besides, it has been confirmed that Japanese media's 'semantic orientations' toward Korea has been deteriorating over the past decade. Above all, it was also confirmed that semantic orientations of Japanese media news articles about Korea were actively reflected in the issue of Korea-Japan relations.

Field : Lexicology

Keywords : Text mining, Semantic Orientations of Words, Machine learning, Japanese media news about Korea, Python

1. 들어가며

IT기술의 발전으로 인류가 생산해 내는 정보의 양은 기하급수적으로 증가하고 있다. KT경제경영연구소에 따르면, 매체를 불문하고 인류가 기록을 남기기 시작한 이래 2,000년대 초반까지 생산된 정보의 총량은 약 20엑사바이트(Exabyte, EB), 즉 200억 기가바이트(Gigabyte, GB)라고 한다. 그런데 그로부터 20년이 채 지나지 않은 2017년 기준, 전 세계에서 하루 생성되는 데이터양은 약 2.5엑사바이트로(이병욱, 2020:1) 이를 기가바이트로 환산하면 약 25억 기가바이트에 달한다. 다시 말하면 인류는 2017년 기준 단 8일 만에 지난 2,000년간 인류가 쌓아온 모든 정보량을 능가하는 정보를 생산했다는 것을 의미한다(연간 정보 생산량으로 환산하

* 본 연구는 2020년도 동덕여자대학교 연구년 제도 지원에 의하여 수행된 것임. This study was supported by the sabbatical of Dongduk Women's University in 2020.

** 동덕여자대학교 일본어과 부교수, 어휘론・코퍼스언어학・사회언어학.

면 약 9,125억 기가바이트). 그러나 이는 시작에 불과한데, 다가오는 2025년이 되면 전 세계에서 생산되는 연간 디지털 정보의 양은 약 163조 기가바이트, 하루 약 4,466억 기가바이트에 이를 것이라고 한다(ICT 시장조사기관 IDC, 2020). 이를 다시금 비교해 보면, 2025년에는 2017년도 연간 총 생산 정보량에 해당하는 분량을 단 2일 남짓 만에 생산해 낼 것이라는 놀라운 계산이 된다.

물론 이와 같은 방대한 정보 속에는 구조화된 데이터베이스 및 텍스트와 같은 정형 데이터뿐만 아니라 비구조화 텍스트, 그래픽, 음성, 영상 등 다양한 형태의 비정형 데이터 또한 섞여 있는 것이 사실이다. 하지만 이와 같은 모든 정보의 폭발적인 증가 상황 속에서 인류가 기존의 수작업과 직관적 통찰과 같은 방법만으로 필요한 정보를 수집・정제하고 이를 바탕으로 의미 있는 해석을 수행하는 것이 불가능에 가까워지는 시대, 즉 본격적인 빅데이터 시대를 맞이하게 되었다는 것을 의미한다.

이에 본 연구에서는 빅데이터 속에서 원하는 텍스트 데이터를 수집하여 처리하는 양적 분석 기법뿐만 아니라 기존의 직관적 내용 분석 기법의 약점을 극복하기 위해 텍스트마이닝(Text mining) 기법을 일본어학 분야에 적용해 보고자 한다. 구체적으로, 웹크롤링과 같은 자동화 처리 툴을 통해 일본 미디어의 한국에 대한 뉴스 기사를 대규모로 수집 및 정제하여 코퍼스를 작성했다. 또한 여기에 머신러닝에 기반을 둔 데이터 분석 기법을 도입하여 각각의 기사의 '감정'을 도출하고 이를 통시적으로 고찰하여 변화의 추이를 알아보는 것을 통해 일본의 한국에 대한 '감정' 및 '관심사'의 흐름을 명확히 하고자 했다. 이는 뉴스 기사에 대한 분석 결과뿐만 아니라 연구 과정인 빅데이터에 대한 텍스트마이닝 기법 자체 또한 일본어학 연구에 시사하는 바가 적지 않다고 판단된다.

2. 선행연구

지금까지 일본어학 분야에서 텍스트의 '감정'에 관한 연구는 '놀람', '안도' 등 개별적 감정표현 및 '의성어', '의태어' 등 품사별 감정어휘에 대한 어휘론적 분석 그리고 '문말 표현', '술어문' 등 감정표현의 특정 구문 및 문법적 출현 양상, 마지막으로 감정표현의 '발화유형', '지각' 등 음성적 분석 등등이 주로 이루어져 왔다. 그러나 최근 高村大也他(2006), 狩野達哉他(2012), 金原直也他(2018) 등 대규모 텍스트 데이터에 포함된 필자의 의견이나 태도를 포함한 '감정'을 자동으로 발견 및 특정하고 이를 통합적으로 분석하고자 하는 새로운 시도가 이루어지고 있다. 추가해서 中岡伊織他(2013), 森田晋也他(2017・2018), 武内達哉他(2019) 등, 트위터 및 페이스북 등 소셜미디어 및 다양한 텍스트 빅데이터에 대한 감정분석 및 이를 위한 기본 감정 사전 구축 등의 연구도 찾아볼 수 있다. 그 중에서 특히 高村大也他(2006:627)는 텍스트의 감정분석을 위한 중요한 재료로서 이차변수(긍정적 인상을 주는 단어와 부정적 인상을 주는 단어로 이분화)인 '감정극성(感情極性; Semantic orientations)'의 개념을 상정하고 이 값을 자동적으로 추출하는 방법으로 '단어감정극성대응표(単語感情極性対応表; Semantic Orientations of Words)'를 제안했다. 본고에서는 이러한 高村大也他(2006)의 '단어감정극성대응표'를 기반으로 프로그램 언어인 파이썬(Python)을 활용하여 독자적 분석 프로그램을 개발하고 일본 미디어의 한국에 관한 뉴스에 대한 감정 및 그 추이를 분석했다. 이를 통해 일본에서 소비되고 있는 한국 관련 뉴스 기사가 주로 어떠한 감정을 내포하고 있는지, 그리고 이와 같은 감정은 지난 10년간 어떠한 양상을 보이고 있는지 객관적으로 파악할 수 있었다.

3. 연구방법

텍스트마이닝은 크게 두 과정으로 나눌 수 있는데 첫 번째가 자료 처리과정(data processing) 그리고 두 번째가 자료 분석(data analysis)단계라고 할 수 있다. 우선, 자료처리과정은 비구조화 데이터를 분석 가능한 형태로 가공 및 정제하는 단계이고, 자료 분석은 데이터마이닝이나 기계학습(machine learning) 그리고 통계학(statistics) 등을 활용하여 텍스트로부터 유의미한 정보를 추출하는 단계이다.

본고에서는 2010년 7월부터 2020년 6월까지 10년간의 일본 미디어의 한국에 대한 뉴스 기사를 웹크롤링 방식으로 수집하여 뉴스 코퍼스를 독자적으로 구축했다. 그리고 이를 텍스트마이닝 기법을 사용하여 분석하는 것을 통해 한국에 대한 일본 미디어 뉴스의 감정극성 및 토픽의 추이를 통시적으로 고찰했다. 우선 자료처리과정 단계의 구체적인 조사대상 및 자료 수집 방법은 다음과 같다.

3.1 조사대상

본고에서는 다양한 일본의 미디어의 한국 뉴스 코퍼스를 구축하기 위해, 특정 미디어에 한정하지 않고 여러 미디어의 뉴스를 한데 모아 제공하는 일본의 뉴스 포털사이트의 뉴스 기사를 조사 대상으로 설정했다. 단, 뉴스 포털 사이트는 웹크롤링에 유리한 정형화된 포맷을 갖고 있는 일본 내 뉴스 부문 선호도 2위 뉴스 포털 「라이브도어 뉴스」(ライブドアニュース, livedoor ニュース)[1]를 조사 대상으로 선정했다. 참고로 뉴스 부문 1위의 「야후재팬 뉴스」[2]의 경우 한국 미디어의 일본어판 기사의 비중이 높아 일본 미디어에 의한 대량의 뉴스 기사 수집을 위해 배제했다.

또한 일본 미디어의 한국 관련 뉴스에 대한 '감정극성'의 특성을 분석하기 위해 「라이브도어 뉴스」에 게재된 한국 관련 뉴스 기사를 실험군 조사 대상으로 설정하는 한편, 한국 뉴스를 포함한 모든 뉴스 기사 코퍼스를 대조군 조사 대상으로 설정하여 일본 미디어의 한국 관련 뉴스에 대한 '감정극성'의 특성을 비교 대조 분석했다.

상기 조건 하에 구축한 실험군, 「일본 미디어의 한국 뉴스 코퍼스」(이하, AJ 일본 뉴스 코퍼스)는 각 연도별 4분기 4개 파일, 전체 10년 40개 파일로 구성되어 있다. 그리고 각 파일은 편차가 있으나 약 200건 내외의 기사를 담고 있는데, 「AJ 일본 뉴스 코퍼스」의 상세정보는 아래의 (1)과 같으며, 본 연구의 검증과 후속 연구 그리고 타 연구자를 위해 (1)g와 같이 전체 코퍼스 데이터는 웹을 통해 공개했다. 참고로 대조군에 해당하는 「AJ 일본 뉴스 코퍼스-대조군」의 경우, 실험군이 뉴스 기사를 분기별로 검색 후 크롤링을 실시한 것과 달리 전체 기간을 한 번에 검색하여 크롤링을 실시한 까닭에 아래의 (2)와 같이 파일 개수 및 정보량에 있어서 차이를 보이지만, 그 이외의 정보는 (1)의 실험군과 동일하며 마찬가지로 (1)g를 통해 코퍼스 데이터 전체를 웹상에 공개해 두었다.

(1) 실험군 : **「일본 미디어의 한국 뉴스 코퍼스; AJ 일본 뉴스 코퍼스」 Ver.1.202007**
   a. 수집기간 : 2010년 7월 1일 ~ 2020년 6월 30일 / 10년
   b. 수집사이트 : Livedoor ニュース / https://news.livedoor.com
   c. 정보필드 : Link, Title, SubTitle, Topic01, Topic02, Date, Media, MediaLink, Text
   (순서대로 뉴스기사 링크, 뉴스제목, 뉴스부제목, 뉴스상위분류, 뉴스하위분류, 게재일시, 미디어, 미

1) Livedoor ニュース : LINE의 종합 뉴스 포털 사이트. 2020년 기준 일본 웹사이트 인기 순위 28위, 뉴스 포털 사이트 한정 2위(출처 : SimilarWeb). https://news.livedoor.com
2) 2019년 기준 일본 웹사이트 인기 순위 2위, 뉴스 포털 사이트 한정 1위(출처 : SimilarWeb).

디어 링크, 본문)
d. 정보량 : 기사 총 7099건 / 9,216,077자(공백포함), 1,111,025단어 / 40개 파일, 23.5MB
e. 문자코드(인코딩) : UTF-8(유니코드)
f. 파일명 예시 : 2017-a.csv(2017년 일사분기), 2017-b.csv(이사분기)
g. 코퍼스 배포 : http://www.japanese.or.kr/JapaneseStudy_corpus.aspx

(2) 대조군 : **「일본 미디어 뉴스 코퍼스; AJ 일본 뉴스 코퍼스-대조군」**
a. 정보량 : 기사 총 279건 / 516,889자(공백포함), 60,710단어 / 1개 파일, 1.34MB
b. 파일명 예시 : 2010-a-2020-b_CG.csv

3.2 자료수집 및 정제 방법

본고에서는 앞서 언급한 바와 같이 약 10년간의 「라이브도어 뉴스」의 한국 관련 뉴스 기사 코퍼스를 작성하기 위해 수작업이 아닌 웹크롤링 기법을 도입했으며, 이를 위해 프로그래밍 언어 파이썬(Python) 기반의 전용 크롤러를 개발했다. 단, 「라이브도어 뉴스」 및 「야후재팬 뉴스」, 「goo 뉴스」 등 일본의 뉴스 포털 사이트는 기간 검색을 제공하지 않거나 제공한다고 하더라도 기간이 지난 1년간에 한정되는 등 제한이 많다. 이에 지난 10년간의 기사 수집을 위해 구글(Google)을 통해 우회적으로 「라이브도어 뉴스」 사이트 내 뉴스 기사를 검색하고 그 결과를 웹크롤링하는 방식으로 뉴스 코퍼스를 구축했다. 단, 웹크롤링 후 데이터를 정제하는 단계에서 한국 미디어의 일본어판 기사는 모두 삭제했으나, 일본 미디어에 의한 한국 미디어의 인용 및 언급은 유지했는데(e.g. 聯合ニュース, スポーツソウル日本版 등), 구체적인 기준은 (4)와 같다. 마지막으로 본 크롤러는 본 연구자가 개발한 트위터 크롤러 「AJ_Twitter_Crawling 1.0」[3]를 기반으로 새로 개발되었으며, 구체적인 사양과 예시 코드의 일부를 정리하자면 아래의 (3), 표 1과 같다.

(3) 「라이브도어 뉴스, 구글 경유 웹크롤링 프로그램」
a. 개발환경 : Python 3.8.3, Jupyter Notebook 6.0.3
b. 주 사용 패키지 : BeautifulSoup4, selenium, chromedriver 등

(4) 내용 중심 뉴스 기사 정제 기준
a. 한국 미디어에 의한 일본어판 뉴스 기사 제외. 그러나 일본 미디어에 의한 한국 미디어 및 중국 미디어에 대한 인용 뉴스 기사는 내용에 따라 포함
e.g. 포함:「韓国から強制追放された“お騒がせ”女性タレント、体重90キロに激変!! 「やはり只者ではない」」-「【ニュース提供＝スポーツ·ソウル】“お騒がせタレント”として知られるエイミが注目を集めている。…… 이하 생략」S-KOREA, 2018年8月29日
https://news.livedoor.com/article/detail/15226510
e.g. 제외:「韓国のクラブ集団感染 １１９人に＝ソウルで６９人」聯合ニュース, 2020年5月13日, https://news.livedoor.com/article/detail/18253072
b. 통계, 랭킹 등 한국 관련 단순 언급 혹은 일본에 초점이 맞춰진 뉴스 기사 제외
e.g. 제외: 단순 랭킹 언급, 일본 중심

3)「AJ_Twitter_Crawling 1.0」: 파이썬 기반 트위터 전용 크롤링 프로그램. 관련 연구 및 해당 프로그램 다운로드는 아래의 웹 페이지 참조.
http://www.japanese.or.kr/japaneseutill/AJ_Twitter_Crawling/AJ_Twitter_Crawling.aspx

「勤務先の会社への信頼度　日本は韓国と並んで28か国中最下位-【悲報】不信感が蔓延する日本　勤務先への信頼度は28か国中最下位、メディアへの信頼度も下から3番目」キャリコネニュース, 2018年2月16日
https://news.livedoor.com/article/detail/14311203

e.g. 제외: 일본 중심
「IZONE　宮脇咲良＆本田仁美＆矢吹奈子、韓国デビューの心境を語る「日本ファンが…」Kstyle, 2018年10月29日
https://news.livedoor.com/article/detail/15516548

c. 한국의 식문화, 미용, 의료 등 한국 문화에 중점을 둔 뉴스 기사의 경우 포함
d. 북한 및 재일 한국인에 초점을 맞춘 뉴스 기사는 한국 관련 기사로 인정하여 포함
e.g. 포함:「ヤクザの世界は人種差別がないと信じて、この世界に入った」NEWSポストセブン, 2017年7月15日, https://news.livedoor.com/article/detail/13342209
e. 기타 : 「朝日新聞デジタル」의 뉴스 기사 제외
f. 상기 외의 경우, 저자의 독자적 판단에 따라 분류함
e.g. 제외: 한국어 학습 기사는 예문에 의한 감정극성치 통계 왜곡 우려로 제외
「【韓国語会話】楽しみにしています」C CHANNEL, 2019年3月11日
https://news.livedoor.com/article/detail/16141343

〈표1〉 웹크롤러 코드 예시

```
# 구글 검색 결과 웹크롤링 예시 소스코드

from urllib.parse import quote_plus
from bs4 import BeautifulSoup
from selenium import webdriver

# baseUrl = ''
# plusUrl = input('검색 키워드를 입력해 주세요 :') # 본 연구의 경우, '韓国'
# url = baseUrl + quote_plus(plusUrl)

url = 'https://www.google.com/search?q=' # 구글의 기본 검색 주소창 쿼리를 입력

driver = webdriver.Chrome()
driver.get(url)

html = driver.page_source
soup = BeautifulSoup(html)

r = soup.select('클래스') # 검색하고자 하는 항목의 클래스 입력

# 본 연구의 경우 아래 결과를 링크의 내용을 포함하여 크롤링 후 csv 파일로 저장했음
for i in r:
    print(i.select_one('클래스').text) # 뉴스 제목 출력
    print(i.a.attrs['href']) # 링크 출력

driver.close()
```

3.3 텍스트마이닝

본고에서는 위와 같이 작성된 「AJ 일본 뉴스 코퍼스」에 대한 감정극성 및 추이 분석을 위해 각각의 파이썬 기반 전용 프로그램을 작성, 자동화 했다. 다만, 지면의 제약으로 본고에서는 텍스트마이닝 프로그램의 소스코드에 관한 정보는 생략하나, 구체적 프로그램 사양은 본고 말미의 (15)에, 구동 화면 및 코드 일부는 (1)g의 코퍼스 배포 웹페이지를 통해 함께 공개해 두었다.

4. 연구 결과

4.1. 감정극성 분석

Hiroya Takamura외(2005)에 따르면, 「단어감정극성대응표(単語感情極性対応表)」는 일본어 및 영단어의 '감정극성'의 대응표로서, '감정극성'은 특정 단어가 일반적으로 좋은 인상을 주는가(positive) 아니면 나쁜 인상을 주는가(negative)를 나타낸 이치속성(二値属性)을 가리킨다. 예를 들어 「良い」 혹은 「美しい」와 같은 단어는 긍정적(positive) 극성을, 「悪い」혹은 「汚い」 등의 단어는 부정적(negative) 극정을 갖게 된다. 그런데 이와 같은 이치속성의 감정극성에 단어 네트워크를 활용하여 -1부터 +1의 실수치를 부여한 것이 '감정극성치'라고 할 수 있다. -1에 가까우면 가까울수록 부정적, +1에 가까우면 가까울수록 긍정적으로 볼 수 있는데, 대표적인 예시는 아래 표2와 같다. 참고로 일본어의 기본 사전에는 『岩波国語辞書』가 사용되었으며, 영어는 『WordNet-1.7.1』이 사용되었다.

〈표2〉 단어감정극성대응표(単語感情極性対応表) 예시

기 준	일본어 상위 각 5개 단어	일본어 하위 각 5개 단어
단어 예시	優れる:すぐれる:動詞:1 良い:よい:形容詞:0.999995 喜ぶ:よろこぶ:動詞:0.999979 褒める:ほめる:動詞:0.999979 めでたい:めでたい:形容詞:0.999645	ない:ない:助動詞:-0.999997 酷い:ひどい:形容詞:-0.999997 病気:びょうき:名詞:-0.999998 死ぬ:しぬ:動詞:-0.999999 悪い:わるい:形容詞:-1

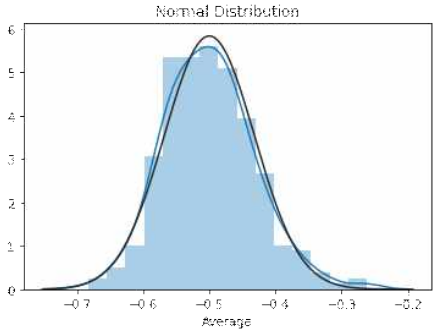
4.2. 「AJ 일본 뉴스 코퍼스 - 대조군」의 선행 분석

실험군 코퍼스에 대한 본격적인 '감정극성'에 대한 분석을 실시하기 전에 먼저 대조군 코퍼스와 실험군 코퍼스 중 2020년 이사분기 코퍼스에 대한 시험적 분석 결과를 확인하는 것을 통해 '단어감정극성대응표'와 웹 크롤링을 통한 텍스트 데이터의 신뢰도를 확인해 보았다. 우선, 3.1절의 (2)와 같이 대조군 코퍼스, 「AJ 일본 뉴스 코퍼스-대조군」는 1개 파일 전체 279건의 기사로 구성되어 있으며, '감정극성' 분석 결과는 다음의 표3과 같다.

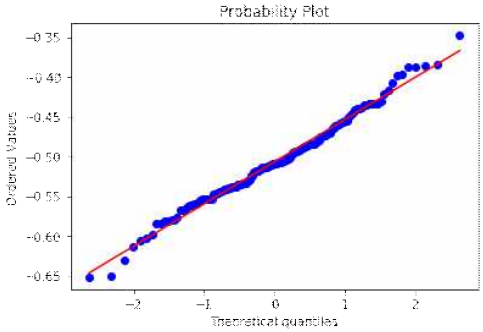
〈표3〉 「AJ 일본 뉴스 코퍼스 - 대조군」 감정극성치 요약

파일명	기사 개수	감정극성치			
		평균	최대값	최소값	표준 편차
semantic_2010-c-2020-b_CG.csv	279	-0.5005555556	-0.262	-0.684	0.068444

그리고 대조군 코퍼스의 감정극성 분석 결과에 대한 정규성 검정을 수행하기 위해 그림2와 같이 정규분포 히스토그램과 그림3과 같이 Q-Q plot을 통해 시각적으로 표현해 보았다. 그림1와 그림2 모두 정규분포와 매우 유사한 형태를 보여주고 있음을 알 수 있다.



〈그림1〉 대조군 정규분포 히스토그램



〈그림2〉 대조군 Normal Q-Q plot

이에 명확한 검정을 위해 Sapiro-Wilks Test(이하, 샤피로 검정)와 Anderson-Darling Test(이하, 앤더슨 검정)를 통해 정규분포 검정을 수행했다. 그 결과 이하 (5)a의 샤피로 검정의 경우, p-value가 본고에서 설정한 유의수준 0.01보다 크기 때문에(p-value 〉 α) 귀무가설에 대한 기각 실패, 즉 데이터는 정규성을 만족한다고 가정할 수 있다. 마찬가지로 (5)b의 앤더슨 검정의 경우, 통계 검정의 임계값과 비교하여 귀무가설H0을 기각 할 수 없다(귀무가설 H0 : 감정극성치 데이터가 정규분포를 따른다).

(5) 정규분포 검정 - 대조군

a. 샤피로 검정

검정통계량 : 0.989079475402832 / **p-value : 0.03400643914937973**

b. 앤더슨 검정

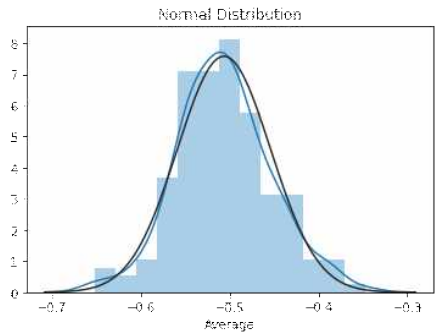
통계량 - Statistic: 0.600

임계값 - Critical_values : 0.568, **0.647, 0.776, 0.905, 1.077**

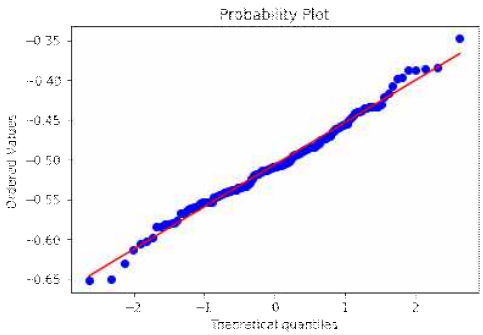
한편 3.1절의 (1)과 같이 실험군 코퍼스, 「AJ 일본 뉴스 코퍼스」는 전체 40개의 파일로 구성되어 있는데, 그 중에서 가장 최근의 2020년 이사분기(2020-b.csv)의 한국 관련 뉴스 기사 파일에 관한 '감정극정' 분석을 실시하였으며 간략한 개요는 표4와 같다.

〈표4〉 「AJ 일본 뉴스 코퍼스 - 실험군」 - '2020년 이사분기' 감정극성치 요약

파일명	기사 개수	감정극성치			
		평균	최대값	최소값	표준 편차
2020-b.csv	163	-0.5068957055	-0.348	-0.658	0.052998



〈그림5〉 실험군(2020-b) 히스토그램



〈그림6〉 실험군(2020-b) Normal Q-Q plot

(6) 정규분포 검정 - 실험군 2020년 이사분기(2020-b.csv)

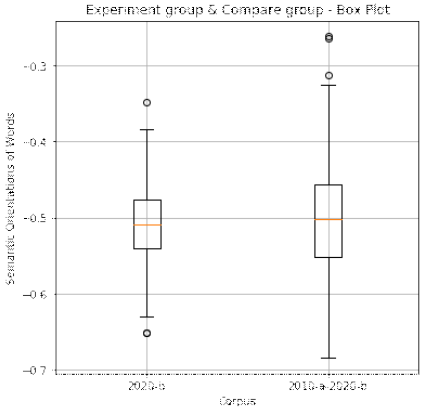
a. 샤피로 검정

검정통계량 : 0.9898890852928162 / **p-value : 0.29715994000434875**

b. 앤더슨 검정

통계량 - Statistic: 0.536

임계값 - Critical_values : **0.563, 0.641, 0.769, 0.897, 1.067**



〈그림7〉 실험군 대조군 박스 플롯

「실험군 코퍼스-2020-b.csv」의 대조군과 경우 마찬가지로 정규성을 만족한다고 가정할 수 있다((p-value 〉 α). 그리고 「실험군 코퍼스-2020-b.csv」의 요약 결과를 대조군 코퍼스와 비교해 개괄해 보자면, 표3과 표4와 같이 실험군 코퍼스의 감정극성치의 평균이 상대적으로 낮으며(-0.5078033473〈-0.5005555556), 최대값(-0.348〈-0.262) 역시 실험군 코퍼스의 값이 낮은 것을 확인할 수 있다. 또한 그림5와 같이 실험군 코퍼스의 감정극성치가 중간값을 비롯, 상대적으로 낮게 분포되어 있다는 것도 시각적으로 확인할 수 있다.

그러나 통계적으로 아래의 (7)과 같이 독립표본 t-Test를 실시해 본 결과, (7)c와 같이 실험군과 대조군의 감정극성치에는 큰 차이가 없다(p-value(0.2776967147921315 〉 0.1).

(7) 「실험군 코퍼스-2020-b.csv」와 「대조군 코퍼스」의 독립표본 t-Test

a. 실험군 및 대조군 모두 정규분포(p-value 0.01), (5)·(6) 참조

b. 등분산 검정

1) Levene test를 통해 등분산 검정 결과 기각. 등분산성을 만족하지 않음
LeveneResult(statistic=11.513498918273097, p-value=0.0007532952990579888)

2) Bartlett test를 통해 등분산 검정 결과 기각. 등분산성을 만족하지 않음
BartlettResult(statistic=12.847222578524075, p-value=0.0003379805271434057)

c. t-Test 결과
Ttest_indResult(statistic=-1.086173104914973, p-value=0.2780451377452789)

본격적인 10년간의 한국 관련 뉴스 기사에 대한 분석에 들어가기에 앞서, 선행분석의 실험군과 대조군 코퍼스의 가장 긍정적인 기사와 가장 부정적인 기사를 통해 감정극성치의 실효성을 확인해 보기 위해 감정극성치 최상・하위 뉴스 기사 목록을 정리했으며, 각각 아래의 표5・6・7・8과 같다.

〈표5〉 「AJ 일본 뉴스 코퍼스 - 대조군」 - 감정극성치 상위 5개 기사

번호	감정극성치	뉴스 기사 제목
258	-0.262	70歳・テリー伊藤、30代の自分は“手越祐也” 「遊んでいた」時代を回顧
13	-0.264	玉木宏、森田剛、柄本佑……2018年にゴールインした有名芸能人の結婚秘話
82	-0.312	常温よりもパワーアップ！「ホットヨーグルト」の効果がすごい
214	-0.326	コストコの46貫寿司『特選ファミリー盛』は穴子の一本握りがだいぶゴージャス
7	-0.334	LINEで通話できない! そんなときの対処法

〈표6〉 「AJ 일본 뉴스 코퍼스 - 대조군」 - 감정극성치 하위 5개 기사

번호	감정극성치	뉴스 기사 제목
211	-0.684	下半身太りの原因は❝大転子❞にあり！正しい位置に戻すトレーニング方法
10	-0.657	すぐ泣く、涙が出る人必見！ひょっとするとうつ病の初期症状かも・・・
269	-0.651	かんぽ不正販売で営業担当2448人処分
225	-0.643	「六波羅探題」オープニングスタッフ募集
90	-0.636	「撮り鉄」が踏切に三脚を置いて逮捕ーー「往来危険罪」は刑罰がとても重い犯罪

〈표7〉 「AJ 일본 뉴스 코퍼스」'2020년 이사분기' 감정극성치 상위 5개 기사

번호	감정극성치	뉴스 기사 제목
159	-0.348	【韓国コスメ】美白効果抜群☆洗い流さない炭酸パック
149	-0.384	2020年流行間違いなし！あか抜け顔になれる「韓国コスメブランド」7選
96	-0.386	日韓W杯の歴史的瞬間にFIFA再注目 韓国“初4強”も…海外ファン辛辣「史上最大の盗難」
121	-0.387	韓国唯一のノーベル賞にキズ「金大中・元大統領」子息、賞金めぐり骨肉裁判に
160	-0.388	女帝・金与正の日本通な素顔 アニメ好きで自宅にファミコン

〈표8〉 「AJ 일본 뉴스 코퍼스」'2020년 이사분기' 감정극성치 하위 5개 기사

번호	감정극성치	뉴스 기사 제목
141	-0.652	ゴルフ中に倒れた女性の頭から「弾丸」見つかる、韓国で衝撃の事件
20	-0.651	対韓国ビラ、1200万枚準備 北朝鮮「報復の時刻迫る」
92	-0.63	日本では常識でも韓国人に絶対にしてはいけない行為３つ
74	-0.628	韓国で少年がやった！「自分がボールになる始球式」がかわいい
65	-0.606	韓国政府が否定する「韓国軍のベトナム民間人虐殺」、ソウルの教育資料に掲載され物議

우선 감정극성치 상위 기사의 표5와 같이 일본 미디어의 일반 뉴스 기사의 코퍼스의 경우, 필자의 직관에 의한 정성분석과 유사한 감정극성치를 부여 받았다는 것을 확인할 수 있었다. 또한 표7의 일본 미디어의 한국 관련 뉴스 기사, 「실험군 코퍼스-2020-b.csv」도 역시 표7-121번 뉴스 기사를 제외하면 적절한 감정극성치가 부여되었다고 판단된다. 단, 표7-121번의 경우, 노벨상을 수상하게 된 경위, 상금, 의미 등을 설명하는 과정에서(e.g.賞:しょう:名詞:0.998943) 재판 관련 어휘(e.g.争う:あらそう:動詞:-0.880231)의 부정적 감정극성치가 완화되어 기사의 전체적인 감정극성치가 높아지는 결과를 낳았다고 볼 수 있는데, 앞으로 뉴스 기사의 제목과 본문의 통합적 분석은 보완해야 할 부분이라 하겠다.

또한 뉴스 기사의 언어량이 극단적으로 적은 표6-225번의 스태프 모집 기사[4]를 제외하면, 부정적 감정을 담고 있는 뉴스 기사가 정성분석과 유사하게 감정극성치 하위에 위치하고 있음을 확인할 수 있었다. 또한 한국 관련 뉴스 기사인 「실험군 코퍼스-2020-b.csv」의 경우에도 정성분석 결과와 동일하게 부정적 감정의 뉴스 기사에 대한 적절한 감정극성치가 부여 되어 있었다.

참고로 뉴스의 토픽에 있어서는 유사점과 차이점이 공존하는데, 일반 뉴스 기사에 다이어트(표6-211번) 및 질병(표6-10번) 등 건강과 관련된 기사 그리고 사회 일반 사건 사고(표6-269, 표6-90번) 등의 토픽을 가진 뉴스 기사가 부정적 감정극성치를 보이고 있으며, 한국과 관련된 뉴스 기사(실험군 코퍼스)의 경우에도 역시 건강(표8-74번[5]) 및 사건 사고 기사(표8-141번)가 부정적 감정의 감정극성치를 나타내는 점에서 서로 일치했다. 그러나 한국과 관련된 뉴스 기사(실험군 코퍼스)의 경우, 긍정적 감정극성치의 경우 표7-159・149과 같이 화장품을 필두로 한 미용, 패션 등의 토픽이, 부정적 감정극성치의 경우 북한(표8-20번) 관련 혹은 과거사(표8-20・65번) 등 역사 관련 토픽의 뉴스 기사가 상위와 하위를 차지하고 있다는 점에서 차이를 보였다.

4.3. 「AJ 일본 뉴스 코퍼스」 분석

4.3.1 감정극성치 통계 검정

앞선 절의 선행 분석을 통해 텍스트 마이닝 기법을 활용한 감정극성치에 대한 부여가 검증 된 바, 본 절에서는 본격적으로 전체 「AJ 일본 뉴스 코퍼스」에 대한 '감정극성' 분석을 수행했는데, 그 결과의 개요는 이하 표9와 같다. 또한 정규성 검정 결과 역시 아래의 (8)과 같다.

4) 2014년 8월 28일 기사로, 언어량(공백포함 283문자)이 작아 감정극성치(-0.643)를 그대로 수용하는 데에는 무리가 있음. 단, 본고에서는 기사의 분량을 통한 통제는 실시하지 않았음. https://news.livedoor.com/article/detail/9192139

5) 2020년 5월 8일 기사로 제목과는 달리 기사의 전반적인 내용은 코로나 관련 스포츠 이벤트의 중단을 소개하고 있음. https://news.livedoor.com/article/detail/18230434

〈표9〉「AJ 일본 뉴스 코퍼스」 및 선행 분석 감정극성치 결과 요약

파일명	기사 개수	감정극성치			
		평균	최대값	최소값	표준 편차
semantic_2010-c-2020-b_CG.csv	279	-0.5005555556	-0.262	-0.684	0.068444
2020-b.csv	163	-0.5068957055	-0.348	-0.658	0.052998
AJ 일본 뉴스 코퍼스	7,099	-0.475905	0.032	-0.781	0.074277

(8) 정규분포 검정 - AJ 일본 뉴스 코퍼스

a. 샤피로 검정(n(표본 개수)이 5,000을 상회하여 불가)

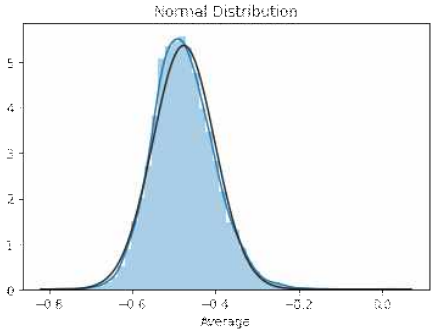
b. **앤더슨 검정**

통계량 - Statistic: **10.967**

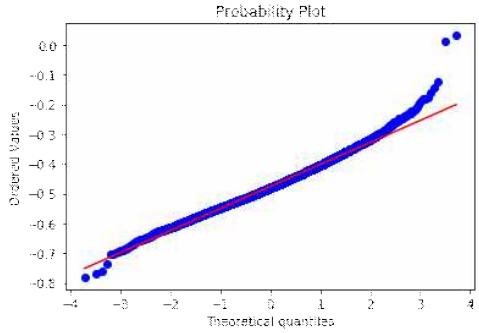
임계값 - Critical_values : 0.576, 0.656, 0.787, 0.917, 1.091

표9의 「AJ 일본 뉴스 코퍼스」의 전체 감정극성치 요약을 통해, 일본 미디어의 한국 관련 뉴스 기사는 일반 뉴스 기사와 비교할 때, 최대값과 최소값의 편차가 크며, 부정적인 감정을 가진 뉴스 기사의 정도가 상대적으로 강하다는 것을 알 수 있다. 또한 위 (8)과 같이 앤더슨 검정에 따르면 「AJ 일본 뉴스 코퍼스」의 감정극성치는 정규분포를 따르지 않는데, 이 또한 일반 뉴스 기사와 달리 감정극성치가 극단적이라는 것을 의미한다(단, 본고에서는 중심극한정리를 적용, 정규분포를 따른다고 가정). 이와 같은 「AJ 일본 뉴스 코퍼스」의 전체 감정극성치를 시각화 하면 아래 그림6・7・8과 같다.

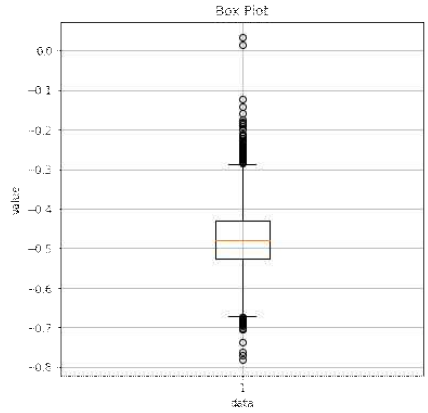
그림6의 정규분포 히스토그램을 통해 「AJ 일본 뉴스 코퍼스」의 감정극성치가 전체적으로 왼쪽, 즉 부정적 감정에 치우쳐 있으며, 그림8과 같이 중간값과 동떨어진 극단적인 감정의 뉴스 기사가 다수 발견된다는 점을 재확인 할 수 있다.



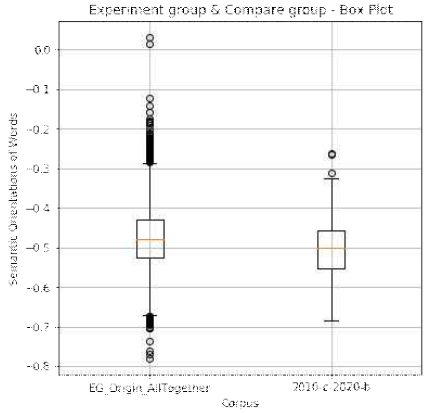
〈그림8〉 AJ 일본 뉴스 코퍼스 히스토그램



〈그림9〉 AJ 일본 뉴스 코퍼스 Q-Q plot



〈그림10〉 AJ 일본 뉴스 코퍼스 박스 플롯



〈그림11〉 AJ 일본 뉴스 코퍼스와 대조군 비교 박스 플롯

이어서 「AJ 일본 뉴스 코퍼스」와 「AJ 일본 뉴스 코퍼스-대조군」에 대한 비교를 수행하기 전에 양 코퍼스의 감정극성치를 시각화하여 비교하면 위 그림9와 같다. 시각적으로 볼 때, 양 코퍼스는 서로 다른 평균값을 가질 것으로 추정되며, 앞선 분석과 마찬가지로 일본 미디어의 한국 관련 뉴스 기사의 경우 일반 뉴스 기사보다 극단적인 감정극성치의 분포를 보인다는 것을 확인할 수 있다.

우선, t-Test의 경우 1)두 집단은 각각 정규분포를 따른다 그리고 2)두 집단은 분산이 동일하다는 가정을 만족해야 한다(단, (8)의 앤더슨 검정에 따르면 「AJ 일본 뉴스 코퍼스」는 정규분포를 따르지 않음). 그러나 본고에서는 「AJ 일본 뉴스 코퍼스」의 경우 n값이 5,000이상으로 충분히 크다고 판단, 중심극한정리를 적용하여 t-Test가 가능하다는 것을 지지하며, 이를 기반으로 아래 (9)와 같이 독립표본 t-Test를 수행했다. 그 결과 「AJ 일본 뉴스 코퍼스」와 「AJ 일본 뉴스 코퍼스-대조군」의 평균은 통계적으로 유의한 차이가 있음(p-value〈0.05)을 확인할 수 있었다.

이 결과를 다시 말하면, 일본 미디어의 일반 뉴스 기사에 대한 감정극성치와 한국 관련 뉴스 기사의 감정극성치의 평균이 서로 상이하기 때문에 한국 관련 뉴스 기사의 감정이 일반 기사의 감정과 상이한 모습을 보이고 있으며, 특히 부정적인 감정극성치에 치우쳐 있음을 의미한다.

(9) 「AJ 일본 뉴스 코퍼스」와 「대조군 코퍼스」의 독립표본 t-Test
  a. 정규분포 검정 : 실험군 정규분포 가정, 대조군 (5)와 같이 정규분포(p-value 0.01)
  b. 등분산 검정 : 등분산
    1) Levene test를 통해 등분산 검정 결과 기각 불가. **등분산성을 만족**
      LeveneResult(statistic=2.3015630155226177, pvalue=0.12928673879342328)
    2) Bartlett test를 통해 등분산 검정 결과 기각 불가. **등분산성을 만족**
      BartlettResult(statistic=3.400306095990987, pvalue=0.0651843218847816)
  c. 독립표본 t-Test 결과 : **실험군과 대조군은 통계적으로 유의한 차이가 있음**(p-value〈0.05)
    Ttest_indResult(statistic=5.4529857492139024, **p-Value=5.1135150342091636e-08**)

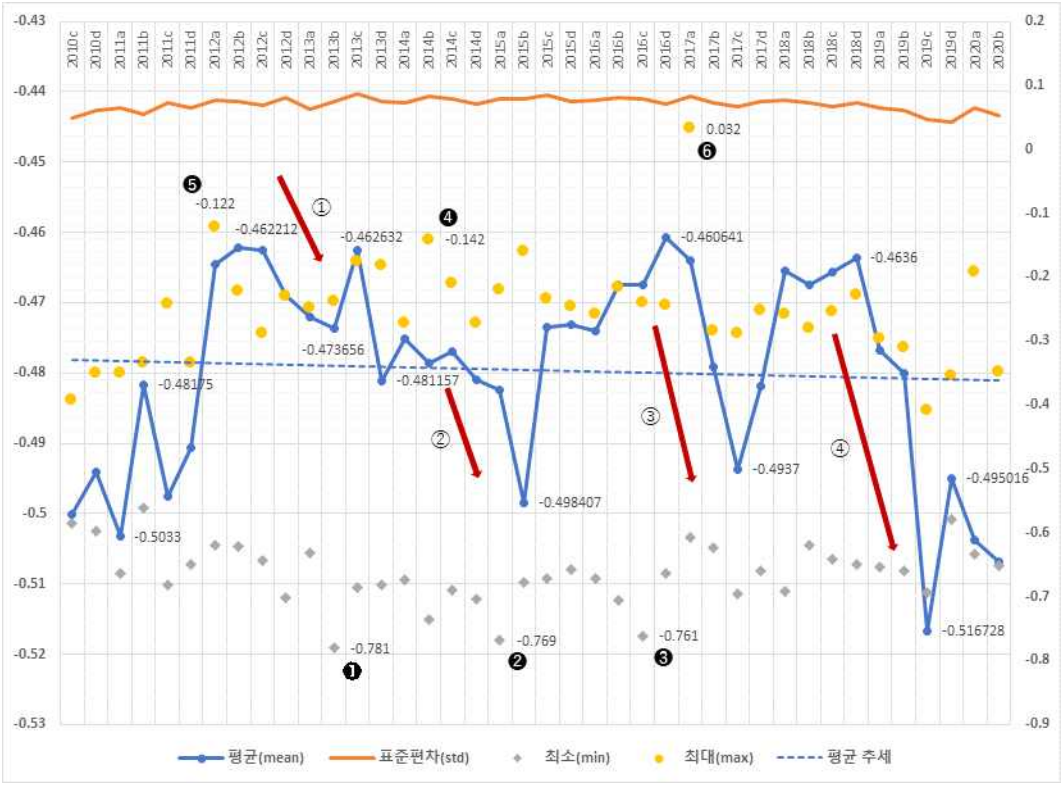
4.3.2 감정극성치 추이 분석

앞서 「AJ 일본 뉴스 코퍼스」의 전체 코퍼스에 대한 통계적 검정을 실시했으나, 본 절에서는 10년간의 텍스트 데이터를 연도 및 분기로 분할하고, 각각의 코퍼스에 대한 통계 데이터에 대한 분석을 수행했다. 이를 통해 각 기사별 감정극성치에 대한 타당도 및 토픽 등에 대한 상세 분석, 그리고 10년간의 변화의 추이에 대한 조망에 이르기까지 종합적인 분석을 수행하고자 한다. 우선, 각 연도 및 분기별 뉴스 코퍼스에 대한 통계 데이터는 아래의 표10, 그림10과 같다.

〈표10〉 「AJ 일본 뉴스 코퍼스」의 연도 및 분기별 감정극성치 통계 변화 추이

연도 year	분기 quarter	개수 count	평균 mean	표준편차 std	최소값 min	최대값 max	샤피로 Statistics	샤피로 p-value	귀무가설
2010	3	50	-0.5002	0.048494	-0.585	-0.393	0.978	0.472	fail to reject
2010	4	46	-0.494087	0.060493	-0.597	-0.351	0.961	0.129	fail to reject
2011	1	50	-0.5033	0.065631	-0.664	-0.35	0.992	0.976	fail to reject
2011	2	48	-0.48175	0.055191	-0.561	-0.335	0.945	0.026	fail to reject
2011	3	67	-0.497522	0.073445	-0.681	-0.243	0.966	0.064	fail to reject
2011	4	94	-0.490713	0.064543	-0.65	-0.335	0.99	0.704	fail to reject
2012	1	146	-0.464507	0.077105	-0.62	-0.122	0.961	0	reject H0
2012	2	203	-0.462212	0.07407	-0.622	-0.223	0.994	0.529	fail to reject
2012	3	212	-0.462505	0.067807	-0.644	-0.288	0.995	0.666	fail to reject
2012	4	214	-0.468925	0.080938	-0.701	-0.231	0.993	0.375	fail to reject
2013	1	195	-0.472067	0.063536	-0.632	-0.248	0.991	0.266	fail to reject
2013	2	224	-0.473656	0.075161	-0.781	-0.239	0.989	0.08	fail to reject
2013	3	220	-0.462632	0.086023	-0.686	-0.176	0.965	0	reject H0
2013	4	216	-0.481157	0.07407	-0.682	-0.182	0.991	0.223	fail to reject
2014	1	197	-0.475152	0.073664	-0.673	-0.273	0.987	0.059	fail to reject
2014	2	218	-0.478661	0.082311	-0.736	-0.142	0.985	0.022	fail to reject
2014	3	222	-0.476865	0.07807	-0.689	-0.21	0.993	0.436	fail to reject
2014	4	201	-0.48101	0.07111	-0.703	-0.273	0.987	0.072	fail to reject
2015	1	205	-0.482463	0.079384	-0.769	-0.221	0.988	0.094	fail to reject
2015	2	204	-0.498407	0.079068	-0.677	-0.159	0.953	0	reject H0
2015	3	204	-0.473525	0.085599	-0.671	-0.235	0.994	0.53	fail to reject
2015	4	196	-0.473133	0.074213	-0.658	-0.247	0.993	0.467	fail to reject
2016	1	208	-0.474034	0.077475	-0.672	-0.258	0.997	0.958	fail to reject
2016	2	219	-0.467379	0.081135	-0.705	-0.217	0.989	0.108	fail to reject
2016	3	233	-0.467391	0.079293	-0.761	-0.24	0.992	0.227	fail to reject
2016	4	234	-0.460641	0.070902	-0.664	-0.245	0.988	0.044	fail to reject
2017	1	214	-0.464033	0.083595	-0.607	0.032	0.875	0	reject H0
2017	2	201	-0.479154	0.073202	-0.623	-0.284	0.98	0.006	reject H0
2017	3	240	-0.4937	0.067571	-0.696	-0.288	0.992	0.236	fail to reject
2017	4	230	-0.481848	0.074366	-0.66	-0.253	0.993	0.354	fail to reject
2018	1	249	-0.465382	0.076392	-0.691	-0.258	0.998	0.975	fail to reject
2018	2	272	-0.467522	0.072253	-0.62	-0.281	0.988	0.026	fail to reject
2018	3	256	-0.465699	0.066288	-0.642	-0.255	0.996	0.833	fail to reject
2018	4	190	-0.4636	0.072455	-0.65	-0.229	0.981	0.012	fail to reject
2019	1	175	-0.476686	0.064254	-0.653	-0.297	0.985	0.054	fail to reject
2019	2	159	-0.480088	0.061187	-0.66	-0.311	0.997	0.98	fail to reject
2019	3	81	-0.516728	0.045764	-0.693	-0.409	0.963	0.019	fail to reject
2019	4	63	-0.495016	0.042228	-0.58	-0.355	0.963	0.055	fail to reject
2020	1	80	-0.5038	0.064071	-0.634	-0.192	0.913	0	reject H0
2020	2	163	-0.506896	0.052998	-0.652	-0.348	0.99	0.297	fail to reject
합 계		7142			최소값	최대값			
평 균			-0.478173		-0.781	0.032			

우선, 아래의 그림10의 '평균추세'선을 통해 최근 10년간 한국에 관한 일본 미디어의 뉴스 기사의 감정은 지속적으로 부정적인 방향으로 변해 왔음을 알 수 있다. 특히 2010년 사사분기(2010d0), 2011년 삼사분기(2011c), 2015년 이사분기(2015b), 2017년 삼사분기(2017c), 2019년 사사분기(2019d) 이후 등의 시기의 뉴스 기사의 감정극성치가 눈에 띄게 낮은데, 이를 (10)과 같은 한일 갈등의 사례와 함께 구체적인 기사를 통해 확인해 보고자 한다.



〈그림12〉 「AJ 일본 뉴스 코퍼스」 의 감정극성치, 연도 및 분기별 추이

범례) ❶~❻와 같이 극단적 감정극성치를 보이는 뉴스 기사는 모두 한류 관련 연예, 미식 관련 기사로, 해당 수치는 드라마의 내용이나 미식에 대한 묘사 등의 영향으로 인한 것.

(10) 2010년 이후, 한일 갈등 주요 일지

a. 2012년 8월 10일 - 2012b : 이명박 대통령 독도 방문

b. 2014년 1월 15일 ~ 16일 - 2014a : 미국 상하원 2007년 일본군 '위안부' 결의 준수촉구 법안 통과

c. 2015년 12월 28일 - 2016d : 한일외교장관회담, 한일위안부 합의

d. 2016년 12월 28일 - 2016d : 부산 일본총영사관 앞 평화의 소녀상[6] 설치관련 일본 정부 반발

6) 2011년 한국정신대문제대책협의회의 주도로 1000번째 수요집회를 기려 종로구 일본대사관 앞에 평화비(평화의 소녀상)라는 이름으로 처음 설치되기 시작하여 국내외 곳곳에 설치되기 시작했다.

e. 2019년 1월 2일 - 2019a : 강제징용 피해자 신일철주금 한국자산 강제집행신청
f. 2019년 7월 1일 - 2019c : 아베 정부 대한 수출규제 발표

(10)a의 2012b 시기 이후 즉, 이명박 대통령이 한국 대통령 최초로 독도를 방문한 이후, 그림10의 ①과 같이 일본 미디어의 한국 관련 뉴스 기사의 감정극성치가 지속적으로 하락하여 부정적 뉴스 기사가 증가하고 있음을 확인할 수 있다. 실제로 아래 (11)과 같이 부정적 감정극성치를 가진, 특히 영토 분쟁이나 과거사 관련 토픽의 뉴스 기사를 다수 찾아볼 수 있는 점이 이를 뒷받침하고 있다.

(11) 그림10-①시기, 감정극성치 하위 주요 기사
a. -0.546(2012b-하위25)「日本のネットで「竹島も買ってしまえ」　うわさをもとに韓国で警戒感広がる」J-CASTニュース, 2012年4月18日 https://news.livedoor.com/article/detail/6480915
b. -0.582(2012c-하위8)「自衛隊ＶＳ韓国軍　「竹島戦争」完全シミュレーション（2）」, アサ芸プラス, 2012年8月29日 https://news.livedoor.com/article/detail/6899188
c. -0.557(2012c-하위17)「竹島も尖閣も日本人が開拓した歴史ある島…なのに？」新刊JPニュース, 2012年8月21日, https://news.livedoor.com/article/detail/6874694

이어 일본군 '위안부' 갈등이 본격화 된 (10)b의 2012b 시기 이후에는 악화 되었던 한국 관련 뉴스 기사의 감정극성치가 그림10의 ②와 같이 더욱 더 부정적으로 악화되었는데, 이와 같은 주제는 (10)c 한일외교장관회담, 한일위안부 합의가 이루어진 이후에야 감정극성치가 개선되는 모습을 보이고 있다. 이를 통해 이 시기 감정극성치의 악화에는 일본군 '위안부' 갈등이 그 원인으로 작용했음을 미루어 짐작할 수 있게 한다. 그림10-② 시기의 한국 관련 뉴스 기사의 경우 부정적 감정극성치를 가진 기사 속에서 (12)d와 같이 일본군 '위안부'와 관련된 직접적인 기사뿐만 아니라, (12)a・b와 같이 한국전쟁과 베트남 전쟁의 한국군의 예를 들어 일본군 '위안부'에 대한 논점을 흐리고자 하는 의도를 가진 뉴스 기사가 특히 눈에 띈다는 특징을 보인다. 마찬가지로 일본군 '위안부' 갈등을 언급하면서도 (12)c・d와 같이, 한국에 의해 역사 전반의 왜곡이 이루어지고 있으며 한국의 역사의식이 부족하다는 점을 들어 마찬가지로 한국이 주장을 평가절하 하고자 하는 의도가 엿보인다.

(12) 그림10-②시기, 감정극성치 하위 주요 기사
a. -0.617(2015b-하위8)「ベトナム人女性「韓国軍に2晩の間何度も強姦された」と証言」NEWSポストセブン, 2015年6月29日 https://news.livedoor.com/article/detail/10287450
b. -0.547(2015b-하위52)「北から連行された女性たちが韓国兵の「性奴隷」になった過去」NEWSポストセブン, 2015年5月11日 https://news.livedoor.com/article/detail/10098980
c. -0.547(2015b-하위54)「佳子さまに暴言韓国人筆者　　世界の歴史家共通認識からズレる」NEWSポストセブン, 2015年5月25日 https://news.livedoor.com/article/detail/10149935
d. -0.537(2015b-하위62)「韓国を訪問中の米下院議員団が朴大統領らと会談、韓国は「慰安婦問題の早期解決」を強調―韓国英字紙」Record China, 2015年4月3日 https://news.livedoor.com/article/detail/9964572

(10)d 시기 이후에는 일본군 '위안부'를 기억하기 위한 '평화의 소녀상'의 한국뿐만 아니라 세계 각지에 설치되는 과정 속에서 일본의 반발이 노골적으로 뉴스 기사에 나타나고 있다. 이와 같은 흐름 속에서 그림10의

③과 같이 한국 관련 뉴스기사의 감정극성치가 다시금 급격하게 악화되는데, 대표적인 기사가 아래 (13)과 같다. 그림10-③ 시기의 경우도 그림10-② 시기와 마찬가지로 (13)d와 같은 '평화의 소녀상'에 대한 직접적인 비판 기사는 물론이거니와 (13)a・c와 같은 베트남전의 한국군의 사례를 들어 논점을 흐리거나, 2015년의 한일위안부 합의에 대한 불이행에 대한 국제적 신의의 문제를 들어 한국을 비난하는 기사가 감정극성치 하위권에서 많이 눈에 띈다는 점이 특징이라고 할 수 있다.

(13) 그림10-③시기, 감정극성치 하위 주요 기사

a. -0.541(2017b-하위41)「慰安婦像撤去拒否ならハノイ韓国大使館前にライダイハン像を」NEWSポストセブン, 2017年4月8日 https://news.livedoor.com/article/detail/12907405

b. -0.519(2017b-하위62)「ケント氏「韓国には嘘が恥ずかしいという概念がないのか」」NEWSポストセブン, 2017年5月22日 https://news.livedoor.com/article/detail/13094486

c. -0.55(2017c-하위41)「百田尚樹氏 「文在寅大統領に「巨大なブーメラン」 ライダイハン像の建立計画も - ベトナムの韓国大使館前に「ライダイハン母子像」建立計画」NEWSポストセブン, 2017年9月26日, https://news.livedoor.com/article/detail/13663753

d. -0.55(2017c-하위42)「韓国、慰安婦像をさらに10体増設へ―中国メディア」Record China, 2017年8月12日, https://news.livedoor.com/article/detail/13464302

e. -0.545(2017c-하위42)「なぜ韓国は国と国との約束が守れない 漢字廃止で思考能力が低下？ - 韓国人がおかしなことを鵜呑みにするのは漢字廃止が影響か」「文在寅大統領の暴走が加速している。慰安婦問題の日韓合意を反故にするような発言を繰り返すとともに、…… 본문 이하 생략」NEWSポストセブン, 2017年9月4日 https://news.livedoor.com/article/detail/13562772

마지막으로 (10)e・f 시기의 경우 그림10의 ④와 같이 최근 10년간에 걸쳐 일본 미디어의 한국 관련 기사의 감정극성치가 가장 극적으로 부정적으로 내려앉은 시기라고 할 수 있다. 이 시기의 경우 (14)a・b와 같이 강제징용 피해자 관련 소송은 무엇이며 이에 대한 대처 방안에 관하여 다룬 뉴스 기사, 그리고 그 대처 방안으로 일본 정부에 의해서 시행된 수출규제의 영향으로 한국이 입게 된 혹은 예상되는 피해에 대하여 다룬 (14)c・e와 같은 기사, 한국의 강제징용 피해자의 소송으로 인한 일본 기업의 한국 내 자산 처분이 얼마나 무모한 일인지에 대해 다룬 (14)d와 같은 기사, 마지막으로 일본의 수출규제에 대한 반발로 일어난 '노노재팬 운동'과 같은 일본 제품 보이콧의 움직임의 의미를 퇴색시키거나 불가능하다는 취지의 (14)d・f와 같은 뉴스기사가 매우 낮은 감정극성치를 나타냈다. 그림10의 ④에서 보이는 극단적인 부정적 흐름은 이와 같은 뉴스기사에 의한 영향으로 판단된다.

(14) 그림10-④시기, 감정극성치 하위 주요 기사

a. -0.546(2019b-하위20)「日韓「対話重要」だけど「徴用工」の解決は.識者討論、なお埋まらない溝」J-CASTニュース, 2019年6月25日 https://news.livedoor.com/article/detail/16676330

b. -0.527(2019b-하위33)「韓国「日本企業の資産差し押さえ」 有効な対抗策とは」NEWSポストセブン, 2019年4月4日 https://news.livedoor.com/article/detail/16263884

c. -0.506(2019b-하위55)「詰んだ韓国。サムスン営業利益60％減の衝撃と文大統領の暗い命運」まぐまぐニュース, 2019年4月11日 https://news.livedoor.com/article/detail/16297959

d. -0.579(2019c-하위5)「徴用工の真実を明かした韓国人、「塩酸まくぞ」と脅迫される」NEWS

ポストセブン, 2019年8月19日 https://news.livedoor.com/article/detail/16947239

e. -0.572(2019c-하위4)「韓国のガソリンスタンドが日本車への給油を拒絶、でも「被害者」は日本政府でなく韓国市民という矛盾」サーチナ, 2019年7月25日
https://news.livedoor.com/article/detail/16829395

f. -0.579(2019c-하위5)「輸出規制に激しく反応する韓国、それを気にしない日本…両国の「実力の差」が浮き彫りに＝中国メディア」サーチナ, 2019年7月16日
https://news.livedoor.com/article/detail/16778985

(15)「감정극성」분석 및 통계 프로그램」AJ-Semantic Orientations of Words Ver.1.2007

a. 개발환경 : Python 3.8.3, Jupyter Notebook 6.0.3

b. 주 사용 패키지 : Janome, pandas, scipy, seaborn, pyplot 등

c. 프로그램 구동 및 분석 순서
수집된 코퍼스 전처리(정제) → 형태소 분석 → 감정극성치 분석 → 결과 출력 → 통계용 전처리(정제) → 기술통계, 샤피로 검정, 앤더슨 검정, 정규분포 검정 → 정규분포 히스토그램, QQ 플롯, 박스 플롯 작성 → 레빈 검정 및 바틀렛 검정 → 독립표본 t-Test → 분석 결과 정성분석

5. 맺음말

본 연구에서는 독자적으로 일본 미디어의 한국에 대한 뉴스 기사를 대규모로 수집 및 정제하여 코퍼스를 구축하고, 이렇게 구축한 코퍼스의 뉴스 기사의 '감정극성'을 머신러닝에 기반을 둔 텍스트마이닝 기법을 통해 분석했다. 그 결과 「단어감정극성대응표(単語感情極性対応表)」를 활용한 텍스트의 '감정극성' 분석은 텍스트의 '감정' 분석에 효과적이라는 점을 확인할 수 있었다. 물론, 텍스트의 제목과 본문의 가중치 문제 등 부족한 부분을 찾아볼 수 있었으나, 이는 금후의 과제로 삼고 싶다. 또한, 최근 10년간 일본 미디어의 한국에 대한 '감정'은 지속적으로 악화되고 있다는 점, 그리고 특히 한일 관계의 이슈에 따라 한국에 대한 일본 미디어의 뉴스 기사의 감정극성치가 적극적으로 반영되어 나타나는 점 또한 확인할 수 있었다. 저자는 금후의 과제로, '감정극성' 분석 기법을 개선하고 각각의 뉴스 기사에 대한 토픽 모델링 분석을 동시에 실시하는 것을 통해 더욱 다각도로 텍스트 데이터에 대한 분석이 가능한 텍스트마이닝 기법을 확립하는 것 등을 들고자 한다.

【참고문헌】

이병욱(2020)「바이오 연구데이터 현황과 활용방안」『BioINPro-바이오연구데이터 동향과 시사점』74호 생명공학정책연구센터 pp.1-17

折本伸之・渥美雅保(2018)「Twitter連携ニュースフィルタリングのためのトピックモデルに基づくユーザの興味の学習」『第80回全国大会講演論文集』情報処理学会 pp.289-290

金原直也・幸谷智紀(2018)「Twitterユーザーの感情分析」『第80回全国大会講演論文集』2018-1号 情報処理学会 pp.637-638

狩野達哉・柏熊宏幸・佐瀬圭祐・山口崇志・河野義広・マッキンケネスジェームス(2012)「ソーシャルメディアにおける感情極性を用いた文章の適性判定」『日本知能情報ファジィ学会　ファジィシステムシンポジウム講演論文集』 日本知能情報ファジィ学会 pp.719-722

黒田絢香(2017)「小説テクストの計量的分析－アーサー・コナン・ドイルの作品から」『言語文化共同研究プロジェクト』大阪大学大学院言語文化研究科 pp.23-41

胡碩・鄭立儀・高橋佑介・小池大地・牧田健作・宇津呂武仁・吉岡真治(2012)「日中時系列ニュース・ブログにおけるトピックモデルの推定と分析」『電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション』112(196), 電子情報通信学会 pp.25-30

武内達哉・萩原将文(2019)「単語の持つ感情推定法の提案と単語感情辞書の構築」『日本感性工学会論文誌』18巻4号 情報処理学会 pp.273-278

高村大也・乾孝司・奥村学(2006)「スピンモデルによる単語の感情極性抽出」『情報処理学会論文誌ジャーナル』Vol.47 No.02 pp.627-637

冨田純平・山田光穂・石井英里子・星野祐子(2019)「TwitterデータをいたLDAとBTMのトピック抽出の結果の比較」『パーソナルコンピュータ利用技術学会全国大会講演論文集』14, 電子情報通信学会 pp.13-16

中岡伊織・中本晋太郎(2013)「自己組織化マップおよび対応分析によるロコミ情報を用いた地域性特徴分析」『第29回ファジィシステムシンポジウム講演論文集』2018-1号日本知能情報ファジィ学会 pp.891-894.

森田晋也・白井靖人(2017)「単語間の類似度に基づいた単語感情極性の判定」『第79回全国大会講演論文集』2017-1号 情報処理学会 pp.601-602

森田晋也・白井靖人(2018)「分野別単語感情極性辞書の作成及び評価」『第80回全国大会講演論文集』2018-1号 情報処理学会 pp.317-318

Blei, David M., Andrew Y. Ng, and Michael I. Jordan, 2003, Latent dirichlet allocation, The Journal of machine Learning research 3 pp.993-1022

Hiroya Takamura, Takashi Inui, Manabu Okumura, 2005, Extracting Semantic Orientations of Words using Spin Model, In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics(ACL2005), pp.133-140

金曙泳「AJ」- All about Japanese Study「My Japanese Study - CORPUS(コーパス)」
http://www.japanese.or.kr/JapaneseStudy_corpus.aspx(검색일: 2020.06.30)

東京工業大学　奥村・高村・船越研究室「単語感情極性対応表」
http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html(검색일: 2020.06.30)

RONDHUIT「livedoor ニュースコーパス」
https://www.rondhuit.com/download.html#ldcc(검색일: 2020.06.30.)

SimilarWeb「上位ウェブサイトランキング」
https://www.similarweb.com/ja/top-websites/japan(검색일: 2020.06.30.)

＜要 旨＞

テキストマイニングを用いた日本のメディアの韓国ニュースにおける感情の推移に対する分析-Pythonを用いた「単語感情極性対応表」の分析を活用して-

ビックデータの時代には今までの手作業だけでは膨大なデータの中で必要な情報を効果的に取集・精製し，意味のある解釈を遂行することが不可能にひとしくなっている。よって本稿では，日本語学の分野でもテキストマイニングの手法の導入が必要である点，またその実際の手法自体に対する検証も必要であるという認識のもとで，独自的に大規模のテキストデータ(コーパス)を構築し，これに対するテキストマイニングの分析を行った。その結果，「単語感情極性対応表」を活用したテキストの「感情極性」の分析はテキストの「感情」の分析に効果的である点を確認できた。また，この10年間における日本のメディアの韓国に対する「感情」は持続的に悪化している点，また何より韓日関係の出来事によって韓国に対する日本のメディアのニュース記事の「感情極性」が積極的に反映されて現れている点もまた確認できた。

論文分野：語彙論

キーワード：データマイニング，単語感情極性対応表，マシンラーニング，日本のメディアの韓国ニュース，パイソン

▣ 김유영(金曙泳)

동덕여자대학교 부교수

yuiyu1004@dongduk.ac.kr

■投稿日 : 2020년 7월 11일

■審査開始 : 2020년 7월 20일

■審査完了 : 2020년 8월 2일

■掲載確定 : 2020년 8월 21일