JKIICE

Short Paper

한국정보통신학회논문지 Vol. 26, No. 11: 1751~1754, Nov. 2022

# 토픽모델링을 이용한 비대면 신문 기사 키워드 분석

# Non face-to-face News Articles Keyword Using Topic Modeling

**Ari Shin[1] · Jun Kwon Hwangbo[2*]**

[1]Student, My Paul School, Education administration, Goesan, 28054 Republic of Korea
[2*]Teacher, My Paul School, Education administration, Goesan, 28054 Republic of Korea

## ABSTRACT

The news articles collected with keyword “non face-to-face” were analyzed through topic modeling applied with LDA algorithm. In this study, collected articles were divided into two periods, period 1(the beginning of COVID-19 spread) and period 2(the end of COVID-19 spread), according to issued date of the articles.

The articles of period 1 showed support for non-face-to-face treatment, smart library, the beginning of the online financial era, non-face-to-face entrance exam and employment, stock investment for main topic words. And the articles of period 2 showed conversion to non face-to-face classes, increasing unmanned stores, online finance, education industry, home treatment for main topic words.

Also, further issues were discussed through visualization of topic words. These results provide evidence that education and unmanned business in non-face-to-face industries are growing.

Keywords : Topic modeling, LDA, Non face-to-face, COVID-19, Text mining

## Ⅰ. 서 론

2020년 1월 20일 국내 첫 환자를 발생시킨 코로나19는 사람들의 일상을 급속도로 변화시키고 우리 사회에 직간접적인 영향을 미쳤다. 소비 패턴의 변화부터 학교, 직장생활까지 삶의 많은 부분을 변화시켰는데 그중 한 부분은 비대면 활동의 증가이다[1].

코로나 이전에도 비대면 서비스는 존재하였으나 대중적으로 활발하게 이용되지 않았다. 코로나 초기에는 많은 사람이 비대면이라는 양식에 익숙하지 않았고 비대면 관련 인프라도 많지 않았다. 하지만 2022년, 코로나의 확산 감소 및 안정화가 된 현재는 비대면을 이용한 메타버스와 같은 산업들이 많이 성장하였다[2]. 본 연구에서는 코로나 초기부터 지금까지 ‘비대면’ 키워드에 대해 주로 다루는 이슈는 어떻게 바뀌었는지 신문 기사 텍스트 마이닝을 통해 알아보고자 하였다.

키워드 분석은 향후 비대면 사회가 어떤 방향으로 변화할 것인지 파악하는데 필수적이며 이를 기반으로 앞서 나가는 데에 도움을 줄 것이다.

해당 연구에서는 주요 이슈 및 토픽을 알아보기 위해 대중들의 인식에 많은 영향을 미치고 의견을 반영하는 뉴스 기사를 분석하였다[3, 4]. 주요 이슈를 파악하기 위해 동향 분석에 주로 사용되는 대표적인 방법인 토픽 모델링을 활용하였다[4].

## Ⅱ. 연구방법

기존 키워드 분석 연구에는 토픽모델링을 활용한 인공지능 관련 이슈 분석(2020)과 텍스트마이닝을 이용한 청소년 문제 토픽모델링(2018) 등이 있다.

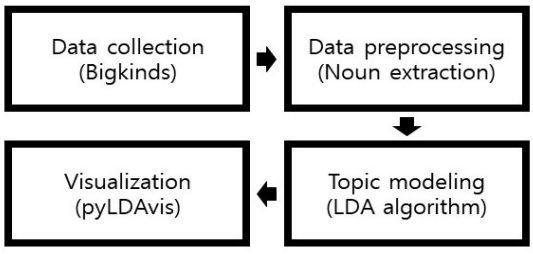


**Fig. 1** Process of research

**Received** 8 September 2022, **Revised** 22 September 2022, **Accepted** 4 October 2022

*** Corresponding Author** Jun Kwon Hwangbo(E-mail:jkhwangbo2@naver.com, Tel:+82-1661-6133)
Teacher, My Paul School, Teacher, Education administration, Goesan, 28054 Republic of Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.11.1751 print ISSN: 2234-4772 online ISSN: 2288-4165

본 연구는 데이터 수집, 데이터 전처리, 토픽모델링, 시각화 순으로 진행하였다.

뉴스 분석 서비스 빅카인즈를 이용하여 '비대면'을 키워드로 2020년 1월 1일부터 7월 1일까지의 5,020건의 뉴스 기사와 2022년 1월 1일부터 5월 9일까지의 5,000건의 뉴스 기사를 수집하였다. 동아일보, 머니투데이, 한겨레, 조선일보 등의 다양한 언론사의 뉴스를 분석하였다. '분석 제외' 옵션을 사용하여 반복되는 유사도 높은 기사와 인사, 부고, 동정, 포토 등의 내용을 담은 기사는 제외하였다.

그리고 수집한 데이터를 konlpy의 okt.noun() 함수를 사용하여 명사를 추출하였다. 연구 초기에는 konlpy의 kkma 라이브러리를 사용하였지만 약 1억 5천만 어절 규모의 KCC150을 분석하는데 kkma는 116,430초가 소요됐지만 Okt는 kkma 소요 시간의 약 1/7배인 15,916초가 소요되므로 okt라이브러리로 변경하였다[5].

토픽모델링은 gensim 라이브러리를 이용하였고, LDA 알고리즘을 사용하여 진행하였다. 마지막으로 토픽 모델링 결과는 pyLDAvis를 통해 시각화를 진행하였다.

분석은 구글에서 서비스 중인 클라우드 기반의 개발 환경인 CoLab을 이용하였다.

## Ⅲ. 연구 결과

본 연구에서는 20.01.01~20.07.01을 period 1으로 정의하고 22.01.01~22.05.09을 period 2로 정의하였다. 객관적으로 토픽 수를 선정하기 위하여 주제 응집도와 혼란도를 고려하였다. 주제 응집도란 각 토픽 내의 단어가 의미론적으로 일치하는지 수치로 계산한다. 주제 응집도의 경우 점수가 높을수록 데이터의 적정 토픽 수에 도달한다고 할 수 있다. 혼란도는 특정한 확률 모형이 실제로 관측되는 값을 얼마나 유사하게 예측하는지 평가하는 지표이다. 혼란도가 낮을수록 정교한 결과를 추출할 수 있다[6].

본 연구에서 주제 일관성 점수는 토픽의 개수가 2개에서 3개로 변할 때 높아졌으며(0.55), 혼란도는 토픽의 개수가 5개에서 6개로 변할 때 감소하였다(-7.91). 따라서 주제 일관성 점수와 혼란도, 그리고 임의로 실험한 결과를 종합적으로 고려하여 토픽 수를 5개로 결정하였다.

(1) Period 1 분석

**Table. 1** Keywords and topic names for period 1

| topic number | top 7 words | topic names |
|---|---|---|
| 1 | Medical, online, support, government, promotion, medical treatment, digital | Government support for non-face-to-face treatment |
| 2 | library, class, loan, learning, book, spread, progress | Non-face-to-face library due to the spread of Corona |
| 3 | Customer, finance, mobile, bank, product, payment, investment | The beginning of the non-face-to-face financial era |
| 4 | College, Counseling, Recruitment, School, Interview, Employment, Program | Transition to non-face-to-face entrance exam and employment |
| 5 | Class, youth, education, account, securities, lecture, stock | Stock investment after the pandemic |

토픽 1의 주요 키워드는 의료, 온라인, 지원, 정부, 추진, 진료, 디지털이며 토픽 1의 주제명은 '비대면 진료에 대한 정부의 지원 추진'으로 결정하였다. 이는 코로나19의 유행으로 인해 의료법상 불법이었던 비대면 진료가 2020년 2월부터 한시적으로 허용된 것의 영향을 받은 것으로 보인다[7].

토픽 2의 주요 키워드는 도서관, 수업, 대출, 학습, 도서, 확산, 진행이며 토픽 2의 주제명은 '코로나 확산에 따른 비대면 도서관'으로 결정하였다. 실제로 코로나 초기에 많은 도서관이 휴관하였고 이에 도서관도 스마트 도서관 등 비대면 서비스를 시작한 것의 영향을 받은 것으로 보인다[8].

토픽 3의 주요 키워드는 고객, 금융, 모바일, 은행, 상품, 결제, 투자이며 토픽 3의 주제명은 '비대면 금융 시대의 시작'으로 결정하였다. 모바일 뱅킹, 모바일 결제, 인터넷 쇼핑 산업의 성장이 이슈가 되었는데 실제로 2020년 4월 통계청에서 발표한 자료에 따르면 온라인 쇼핑 총거래액은 전년 동월 대비 24.5% 증가하였고 모바일 쇼핑 거래액은 31.1% 증가했다[9].

토픽 4의 주요 키워드는 대학, 상담, 채용, 학교, 면접, 취업, 프로그램이며 토픽 4의 주제명은 '입시와 취업의 비대면으로 전환'으로 결정하였다. 이는 비대면 채용이 등장하고 입시에서의 면접이 비대면 면접이라는 이전에 존재하지 않았던 새로운 방식으로 전환되면서 발생한 이슈들로 판단된다. 실제 취업준비생 713명을 대상

으로 '올해 채용을 대표하는 키워드'를 조사한 결과 74.6%(복수 응답)가 '비대면 채용'을 꼽았다[10].

토픽 5의 주요 키워드를 살펴보면 수업, 청소년, 교육, 계좌, 증권, 강의, 주식이며 토픽 5의 주제명은 '팬데믹 이후 주식투자 열풍'으로 결정하였다. 이는 코로나19 발생 이후 다른 재테크 방식에 비해 고수익을 얻을 수 있는 주식 투자에 많은 사람이 뛰어든 것에 영향을 받았다고 볼 수 있다. 실제 코로나 이후 생애 최초로 금융투자를 해본 사람은 전체 응답자의 19%였다[11].

(2) Period 2 분석

**Table. 2** Keywords and topic names for period 2

| topic number | top 7 words | topic names |
|---|---|---|
| 1 | Corona, class, treatment, service, university, operation, student | Recommendation for conversion to non face-to-face classes |
| 2 | service, education, business, corona, robot, operation, online | Increasing serving and unmanned stores using robots |
| 3 | service, corona, customer, finance, bank, loan, digital | The era of non face-to-face finance |
| 4 | education, business, progress, support, corona, program, participation | Non face-to-face education industry |
| 5 | Treatment, management, medical care, corona, medical treatment, home, service | Initiation of home treatment and non face-to-face treatment |

토픽 1의 주요 키워드는 코로나, 수업, 진료, 서비스, 대학, 운영, 학생이며 토픽 1의 주제명은 '대면 수업 전환 및 비대면 진료 제도화'로 결정하였다. 이는 비대면으로 진행되었던 대학 수업을 대면 수업으로 전환하기를 교육부에서 권고한 것에 영향을 받은 것으로 보인다[12].

토픽 2의 주요 키워드는 서비스, 교육, 사업, 코로나, 로봇, 운영, 온라인이며 토픽 2의 주제명은 '로봇을 이용한 서빙 및 무인점포 증가'로 결정하였다. 서빙 로봇의 대중화와 키오스크를 활용한 무인 점포의 증가에 따라 이러한 토픽이 도출된 것으로 추론된다[13,14].

토픽 3의 키워드는 서비스, 코로나, 고객, 금융, 은행, 대출, 디지털이며 토픽 3의 주제명은 '비대면 금융 시대'로 결정하였다.

토픽 4의 주요 키워드는 교육, 사업, 진행, 지원, 코로나, 프로그램, 참여이며 토픽 4의 주제명은 '비대면 교육 시장의 성장'으로 결정하였다.

토픽 5의 주요 키워드는 치료, 관리, 의료, 코로나, 진료, 재택, 서비스로 분석되었으며 토픽 5의 주제명은 '재택 치료 및 비대면 진료의 제도화'로 결정하였다. 이는 코로나19의 확산으로 인해 한시적으로 허용했던 비대면 진료를 제도화하려는 움직임의 영향을 받은 것으로 보인다[14].

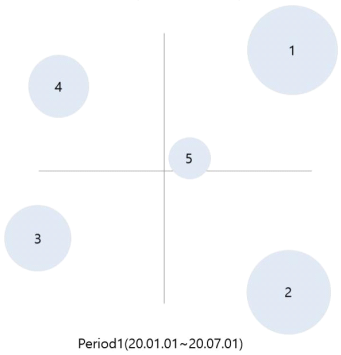


**Fig. 2** Intertopic Distance Map of period 1

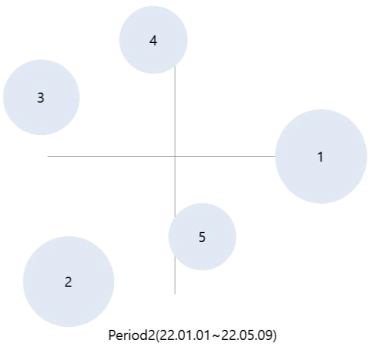


**Fig. 3** Intertopic Distance Map of period 2

LDA 시각화 라이브러리인 pyLDAvis를 통해 2차원으로 토픽을 나타내는 Intertopic Distance Map을 사용하여 결과를 분석하여 보았다. 원의 면적은 해당 토픽이 데이터 내에서 가지는 상대적 중요도를 의미하고 각 원의 위치는 다차원 척도법에 따라 유사성을 거리로 표현한 것이다. 즉 토픽 사이의 거리는 토픽 사이의 연관성을 나타낸다. 이를 통해 각 토픽의 분포 관계, 차이, 유사성 정도를 시각적으로 파악할 수 있다[15].

그림 1의 각 원의 번호는 표 1의 Topic number이고 그림 2의 각 원의 번호는 표2의 Topic number이다. 그림

1과 2에서 1번과 2번 원이 가장 큰 크기를 보이므로 주요 토픽이라고 할 수 있다. 그림 1의 오른쪽 위에 위치한 1번 원과 그림2의 가운데 아래에 위치한 5번 원을 보면 1번 원보다 5번 원의 크기가 작아진 것으로 보아 비대면 진료 관련 이슈가 확연히 감소하였다는 것을 알 수 있다. 또한 주식 열풍, 입시 및 취업 이슈가 그림 1에서는 4번 원과 5번 원으로 존재했지만 그림 2에서는 존재하지 않는 것으로 보아 관련 이슈의 언급이 줄었다는 것을 알 수 있다. 또한 그림 1에서는 존재하지 않았던 무인점포 및 로봇 서빙 토픽이 그림 2의 왼쪽 아래에 위치한 두 번째로 큰 크기의 원, 즉 상당히 중요한 이슈로 대두되었다. 반면, 비대면 금융 시대 관련 이슈는 그림 1의 왼쪽 아래에 위치한 3번 토픽과 그림 2의 왼쪽 위에 위치한 3번 토픽을 비교하여 보면 크기와 위치에 큰 변화가 없으므로 비슷한 언급량을 유지했다고 판단하였다.

## Ⅳ. 결 론

본 연구에서는 비대면 관련 뉴스 기사의 토픽을 키워드를 중심으로 정리하고 그 토픽이 나오게 된 배경에 대하여 추론해 보았다. 또한 코로나 기간 토픽의 변화를 살펴보면서 다양한 분야와 전반적인 산업에서 이슈가 되던 것이 점점 교육과 창업 쪽으로 집중된다는 결과를 도출해냈다. 이 연구는 전염병으로 인해 가속화 된 비대면 산업이 시간의 흐름에 따라 교육 및 무인사업을 향한다는 점을 텍스트마이닝 기법을 통하여 시사하였다는 점에서 의의를 가진다.

## REFERENCES


[ 1 ] Korea Policy Briefing. Masks, contactless, social distancing, daily life completely changed in one year of Corona [Internet]. Available: https://url.kr/lg4bfc.

[ 2 ] J. H. Lee, “A study on legal reform measures according to the growth of the contactless industry,” Korea Legislation Research Institute, Research Report 21-04, 2021.

[ 3 ] S. R. Lee and E. J. Choi, “Comparison of responses to issues in SNS and Traditional Media using Text Mining -Focusing on the Termination of Korea-Japan General Security of Military Information Agreement(GSOMIA)-,” *Journal of Digital Convergence*, vol. 18, no. 2, pp. 277–284, Feb. 2020.

[ 4 ] S. H. Noh, "Analysis of Issues Related to Artificial Intelligence Based on Topic Modeling," *Journal of Digital Convergence*, vol. 18, no. 5, pp. 75-87, May 2020.

[ 5 ] H. J. Won, H. Y. Lee, and S. S. Kang “A Performance Comparison of Korean Morphological Analyzers for Large-scale Text Analysis,” in *Proceedings of Communications of the Korean Institute of Information Scientists and Engineers Conference*, Online, pp. 401-403, 2020.

[ 6 ] S. U. Park, J. Y. Kang, and S. C. Jung, *The complete guide to python textmining*, Wikibooks, Paju, Korea, 2022.

[ 7 ] MEDICAL Observer.Temporarily allow non-face-to-face treatment at home by confirmed doctors [Internet]. Available: https://me2.do/5hENR9K2.

[ 8 ] Ministry of Culture, Sports and Tourism. To prevent the spread of COVID-19, borrow books from the 'Smart Library' and the 'Electronic Library' while public libraries are closed. [Internet]. Available: https://me2.do/GrqPpdO3.

[ 9 ] Statistics Korea, Sports and Tourism. February 2020 Online Shopping Trends. [Internet]. Available: https://me2.do/xi5EVQcn.

[10] Datasom. Recruitment market in 2020, many public companies hired, and ‘non-face-to-face’ is an issue. [Internet]. Available: https://me2.do/FyxLpgop.

[11] Chosun Biz. 1 in 5 Koreans enter stocks due to Corona... Half of them are 2030. [Internet]. Available: https://me2.do/5VA26UdJ.

[12] Policy briefing. Winter seasonal semester, university full-scale face-to-face class conversion pilot operation period. [Internet]. Available: https://me2.do/5KOckaIM.

[13] Science Times. Unmanned trend, global trend? [Internet]. Available: https://me2.do/58FRpgkY.

[14] DocDocDoc. “Discussing non-face-to-face treatment institutionalization, the medical community needs to be more active” [Internet]. Available: https://me2.do/xoKBEmHV.

[15] Y. H. Hong, “Topic Analysis of Software Education Policy : Focused on Busan Regional Newspapers,” *Journal of the Korean Official Statistics*, vol. 48, no. 4, pp. 235-258, Apr. 2019.