LDA 토픽 모델링을 활용한 SNS 분석

장민수[1], 임선영[2]
[1]배재대학교 컴퓨터공학과 학부생
[2]배재대학교 컴퓨터공학과 교수, 교신저자
jjkj1026@naver.com, sunnyihm@pcu.ac.kr

SNS Analysis Using LDA Topic Modeling

Min-Soo Jang, Sun-Young Ihm
Dept. of Computer Engineering, Pai Chai University

요 약

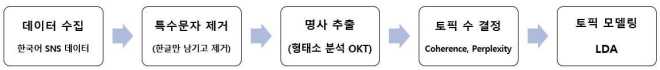
본 연구의 목적은 LDA 토픽 모델링을 활용하여 한국어 SNS데이터에 분석을 통해 우리나라의 여가활동, 일과 직업, 주거와 생활의 동향을 살펴보는 것이다. AI Hub에서 제공하는 한국어 SNS데이터를 수집하고 형태소 분석, 전처리 과정을 거친 후 coherence score을 토대로 최적의 토픽 수를 결정하여 토픽을 추출하였다. 도출한 트렌드를 바탕으로 경영, 마케팅 분야에 미치는 영향을 예측할 수 있을 것으로 기대한다.

1. 서론

최근 스마트폰의 보편화로 언제든지 메신저를 이용할 수 있는 메신저를 이용하여 일상대화를 하는 비중이 높아지고 있다. 또한 코로나(COVID-19)로 인해 비대면, 비접촉과 같은 언텍트(untact) 사회가 지속되면서 외부활동이 감소하고 실내 활동이 증가하는 등 많은 산업과 일상생활에 변화를 가져왔다[1]. 이에 따라 본 논문에서는 LDA 토픽 모델링을 기반으로 연도별로 일상생활과 밀접하게 관련된 주거와 생활, 일과 직업, 여가생활을 주제로 하는 대화를 분석하여 동향을 파악하고자 한다. 또한 주제별로 성별, 연령대별 분석도 함께 진행하여 코로나19(COVID-19) 전후의 변화 여부를 확인하고자 한다.

2. 분석 방법

데이터는 AI Hub에서 제공하는 한국어 구어체 텍스트 기반의 대화분석, 언어모델 학습 등의 자연어처리 AI기술 개발을 위해 구축된 한국인의 일상대화 SNS 데이터[2]를 사용하였다. 기간은 2017년부터 2021년까지로 선정하였고 전처리 작업을 실시하였다. 특수문자를 제거하고 한글만 추출한 후 한국어 형태소 분석기인 OKT를 사용하여 명사를 추출하였고 모든 문서에서 공통적으로 출현하는 빈도 높은 단어와 이름, 주소, 소속과 같이 개인적인 정보를 포함하는 단어들을 제거하였다.



(그림 1) LDA 기반 SNS 분석 시스템 구조

3. LDA 기반 SNS 분석 결과

3.1 주제에 따른 연도별 분석 결과

먼저 일과 직업, 그리고 여가생활에 대해 연도별로 분석 한 결과이다. 코로나 이전에는 스트레스, 퇴사, 이직과 같은 업무 스트레스 관련된 키워드들이 주를 이루었으나, 코로나를 기점으로 코로나, 재택, 혼자와 같은 키워드들이 출현하고 있는 것을 확인할 수 있다. 여가생활에 대해 연도별로 분석한 결과로는 코로나 이전에는 여행, 뮤지컬, 영화, 콘서트와 같은 외부활동 관련 키워드들에서 코로나 이후에는 넷플릭스, 코로나, 드라마, 영화, 혼자와 같이 실내 활동과 관련된 키워드들이 주로 출현하는 것을 확인할 수 있다. 또한 주거와 생활에 대해 연도별로 분석한 결과, 코로나 이전과 비교하여 코로나, 마스크, 배달과 같이 코로나와 밀접한 키워드들이 출현하는 것을 확인할 수 있다.

<표 1> 일과 직업 연도별 추이

2017	2018	2019	2020	2021
퇴사	스트레스	면접	면접	퇴사
인턴	성격	휴가	코로나	경력
야근	취업	회의	연차	코로나
회식	신입	이력서	혼자	근무
경력	회식	스트레스	스트레스	재택
칼퇴	짜증	신입	퇴사	급여
스트레스	이직	회식	재택	면접
정규직	야근	퇴사	유급	연차
퇴직금	새끼	알바	경력	혼자

<표 2> 여가생활 연도별 추이

2017	2018	2019	2020	2021
인스타	게임	제주도	운동	유튜브
배우	방탄	오버워치	드라마	방탄
티켓	야구	축구	영화	넷플릭스
콘서트	영화	여행	혼자	코로나
엑소	마블	분위기	유튜브	콘서트
영화	뮤지컬	티켓	코로나	드라마
드라마	월드컵	드라마	넷플	주식
덕질	클럽	인스타	티비	감성
뮤지컬	축구	경기	방송	덕질

<표 3> 주거와 생활 연도별 추이

2017	2018	2019	2020	2021
주차장	핸드폰	패딩	청약	편의점
청소	공항	노트북	아이폰	서울
아이폰	패딩	인터넷	카페	이사
은행	카페	아파트	핸드폰	아파트
노트북	다이소	핸드폰	코로나	코로나
삼성	지하철	은행	대출	자동차
호텔	보험	전세대출	아파트	마스크
카페	대출	월급	재난	배달
예약	롱패딩	쿠팡	마스크	가격

3.2 성별 및 주제에 따른 연령별 분석 결과

다음으로 성별 및 주제에 대해 연령별로 분석을 하였으며, 먼저 남성을 LDA를 기반으로 분석한 결과이다. 일과 생활에서는 30대를 기점으로 거래처, 미팅, 스트레스와 같이 업무 관련 키워드가, 40-50대는 직위를 나타내는 과장, 부장, 임원과 같은 키워드가 출현하는 것을 확인할 수 있다. 여가생활에 대해 분석한 결과 30대까지는 운동, 친구, 축구, 맥주, 같은 활동적인 여가생활, 40대 이후로는 캠핑, 낚시, 음식으로 여가생활의 변화를 확인할 수 있다.

다음으로 여성을 연령별로 분석한 결과이다. 먼저 일과 직업에 대해서는 30대까지는 알바, 면접, 자소서 등 같은 취업 관련 키워드가, 30대부터는 스트레스, 부장, 계약서 등 업무 관련 키워드가 출현하고 있는 것을 확인할 수 있다. 또한, 여가생활에 대해 분석한 결과로는 30대까지는 콘서트, 인스타, 아이돌 키워드가, 40대부터는 수영장, 등산, 모종, 트로트 등 여가생활의 변화를 확인할 수 있다.

<표 4> 남성-일과 직업 연령대별 추이

남성 - 일과 직업				
20대미만	20대	30대	40대	50대
알바	면접	거래처	부장님	아빠
면접	대리	미팅	담당자	임원
학원	과장	회의	프로젝트	업체
방학	회식	스트레스	과장	사업
군대	스트레스	코로나	한잔	부사장님
시급	공부	인턴	스트레스	진급
남성 - 여가생활				
20대미만	20대	30대	40대	50대
친구	축구	게임	캠핑	캠핑
운동	게임	코로나	결혼식	코로나
아이돌	넷플	주식	바다	유튜브
게임	배그	카메라	낚시	음식
노래방	유튜브	로또	유튜브	
페북	아이돌	맥주	영화	

<표 5> 여성 연령대별

여성 - 일과 직업				
20대미만	20대	30대	40대	50대
면접	연차	월급	커피	분위기
알바	월급	알바	면접	계약서
친구	퇴사	스트레스	프로젝트	엄마
카페	알바	면접	부장님	관리
뷔페	야근	코로나	면접	코로나
시급	자소서	점심	인건비	면접
여성 - 여가생활				
20대미만	20대	30대	40대	50대
유튜브	영화	드라마	수영장	드라마
콘서트	콘서트	친구	블로그	트로트
카페	카페	방탄	등산	노래
방탄	드라마	일본	드라마	방송
엑소	인스타	콘서트	모종	손흥민
티켓	추천	인스타	노래	애완견

4. 결론

본 연구에서는 LDA 토픽 모델링을 기반으로 SNS 데이터를 분석하여 코로나 전후, 성별, 연령별에 따른 키워드의 변화를 확인하였다. 분석 결과를 바탕으로 사람들의 관심, 생활의 변화를 파악하여 효율적인 마케팅, 경영의사결정에 도움이 될 것으로 기대한다. 또한, 향후 연구로는 직업군별 동향을 분석하는 연구를 진행할 예정이다.

사사문구

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1C1C2011105).

참고문헌

[1] 최동현, 송보미, 박다현, 이성우, “텍스트마이닝을 활용한 Covid-19 기간 동안의 항공산업 관련 키워드 트렌드 분석,” 한국산업정보학회논문지, 27(2), pp.115-128.

[2] 한국어 SNS 데이터, AI Hub, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=114