토픽모델링과 분포 기반 클러스터링을 활용한 생성형 인공지능의 사회적 반응 분석

김민영*, 윤선영*, 김수현*
*경북대학교 데이터사이언스대학원
e-mail: tt08@naver.com, tjsdud6900@naver.com, suhyeonkim@knu.ac.kr

Analysis of Social Reaction for Generative Artificial Intelligence based on Topic Modeling and Density-based Clustering

Min-Young Kim*, Sun-Young Yoon*, Suhyeon Kim*
*Graduate School of Data Science, Kyungpook National University

요 약

ChatGPT의 등장과 함께 전세계적으로 생성형 인공지능에 대한 관심과 사용이 급증하고 있다. 본 연구에서는 생성형 AI 키워드가 포함된 뉴스 기사 데이터 기반 토픽모델링을 수행하여 생성형 AI에 대한 사회적인 관심의 변화를 파악하였다. 더하여, 월별 검색량 데이터 분석을 통해 특정 키워드의 검색량이 급증한 날짜를 식별하여 생성형 AI에 대한 사회적 이슈와 트렌드를 파악하였다. 이를 통해 본 연구는 생성형 AI를 활용한 서비스 개발에 필요한 기관에 다양한 인사이트와 정보를 제공하고 그에 따른 중요한 자원으로 활용될 수 있을 것으로 기대된다.

1. 서론

ICT 기술의 발전은 데이터의 저장과 활용에 편의성을 높이면서 우리 사회를 더욱 지능화하고 있다. 최근 인공지능 기술 중 생성형 인공지능(Artificial Intelligence, AI) 기술이 큰 관심을 받고 있으며, 기존에는 사람이 생성하던 문장, 이미지, 음성, 비디오 등을 인공지능이 자동으로 생성할 수 있게 되었다[1].

이에 본 연구에서는 소셜 미디어 데이터를 분석하여 생성형 AI와 관련된 사용자들의 주요 관심사와 기술 및 사회적 이슈를 파악하고자 한다. 첫째로, 뉴스기사 데이터에 Dynamic Topic Model (DTM) 기반 토픽모델링을 수행하여 생성형 AI 관련된 핵심 키워드를 추출함으로써 사회적 관심에 대한 동향을 분석하고자 한다. 또한, 키워드 검색량 데이터를 활용하여 검색량이 급증한 특정 날짜에 대한 기술 분야 및 사회적 이슈를 파악하고자 한다.

2. 연구 방법

2.1 분석 데이터

월별 이슈 및 동향을 분석하기 위해 빅카인즈에서 2023년도 1월부터 2024년 2월까지 약 2,100건의 생성형 AI 관련 뉴스 데이터를 수집하여 분석에 활용하였다. 또한, 검색량이 급증한 시기를 파악하기 위해 네이버 데이터 랩에서 동일한 기간 내에 생성형 AI, ChatGPT로 검색된 검색량 데이터를 수집하였다. 이후 검색량 급증 날짜를 파악하고, 해당 기간에 게시된 뉴스를 추출하여 분석에 사용하였다.

2.2 분석 방법

본 연구에서는 DTM 을 활용한 월별 토픽 분석을 통해 생성형 AI 관련 이슈와 동향을 분석하고자 한다. 먼저, 생성형 AI 관련 뉴스 데이터의 제목과 본문 텍스트에 대한 전처리를 진행하였다. 이 과정에서 유의미하지 않은 단어들인 stop words를 제거하고, 상위 10%의 단어를 제거하여 빈번하게 중복되는 단어를 분석에서 배재하였다. 전처리된 데이터를 바탕으로 DTM을 적용하여 월별 주요 토픽을 추출하였다. DTM이란 데이터 내에 대표 키워드를 추출하고 이를 바탕으로 주제를 정의하는 방식으로, 시간 변수의 추가로 시간에 따른 주제의 변화를 파악할 수 있다[2].

또한, 본 연구에서는 생성형 AI와 ChatGPT의 검색량 급증을 분석하여 어떤 기술 및 사회적 이슈가 이에 영향을 미치는지 파악하고자 하였다. 검색량이 급증하는 날짜를 찾기 위해 네이버 데이터 랩의 검색량 데이터를 활용하여 시계열 그래프의 기울기를 계산하고, 기울기의 상위 5%에 해당하는 임계값을 설정하여 특정 날짜를 식별하였다. 이후, 해당 날짜에 게시된 뉴스의 제목과 본문에서 추출한 키워드 기반 임베딩을

생성하고, Density Based Spatial Clustering of Applications with Noise (DBSCAN)을 활용하여 클러스터링을 수행하였다. DBSCAN은 데이터의 밀도를 기준으로 클러스터를 형성하는 알고리즘으로, 클러스터의 개수를 사전에 정하지 않아도 자동으로 클러스터가 생성된다는 장점이 있다[3]. 분석 결과로 사회적 관심을 끌고 있는 기관, 기술 분야 및 사회 이슈를 식별하고자 하였다.

3. 연구 결과

3.1 DTM 기반 키워드 분석

본 연구에서는 DTM을 활용하여 월별로 변화하는 주요 단어들을 추출하였다. 추출된 상위 10개의 단어들을 기반으로 각 달의 토픽명을 정하였고, 2023년 1월부터 2024년 2월까지 생성형 AI 관련 뉴스 키워드를 통해 실제 어떤 이슈가 있었는지 파악하였다. 분석 결과는 [표 1]에 기술되어 있다. 2023년도 1월에는 많은 국내외 기업들이 AI 개발을 위해 경쟁에 가세하고 있다는 것을 파악할 수 있다. 2023년도 2월에는 챗GPT의 영향으로 국내외 기업에서 검색엔진 개발 열풍이 불었던 것을 알 수 있었다. 2023년 3월에는 교육과 관련된 키워드의 등장으로 생성형 AI로 교육 효과를 극대화하고자 하는 기업들이 등장하고 있음을 확인할 수 있었다. 또한, GPT-4 키워드가 나타남으로 GPT-3.5의 업그레이드 버전이 출시했음을 확인할 수 있다. 2023년 4월 키워드에는 GPT-4의 출시로 국내외 기업들이 생성형 AI 개발에 사투를 벌이고 있는 것을 확인할 수 있었다.

2023년 5월에는 우려, 리스크와 같은 키워드가 등장하는 것을 보아, 생성형 AI 서비스가 많이 개발될 때의 리스크 문제에 대한 관심이 높아짐을 파악할 수 있었다. 2023년 6월에는 콘텐츠, 마케팅, 경쟁력 등의 키워드가 나오는 것을 보아 많은 기업들이 생성형 AI 기술을 콘텐츠 생산 및 마케팅 전략에 도입하여 경쟁력을 강화하고 있음을 시사한다. 2023년 7월에는 글로벌, ETF, 상장 등의 키워드의 등장으로 생성형 AI 개발 기업 투자와 관련된 내용이 등장한 것을 알 수 있었다. 2023년 8월에는 네이버의 초대규모 AI 하이퍼클로바 X의 공개와 함께 네이버가 선도적인 기업의 역할을 수행하고 있음을 확인할 수 있었다. 2023년 9월에는 OTT, 디지털, 초거대 AI와 같은 키워드가 새롭게 나오는 것으로 보아, 미디어 산업에서도 생성형 AI 기술 도입과 발전이 더욱 촉진될 것으로 기대된다. 2023년 10월에는 삼성전자, 가우스, 삼성 스마트 싱스, 콘퍼런스 등의 키워드를 보아, 10월 중 개최되었던 삼성 개발자 콘퍼런스 2024를 통해 삼성전자가 AI 고도화 시대를 주도하고 있음을 알 수 있었다. 2023년 11월에는 CES, 혁신상, 한국 기업 등의 키워드가 등장하여, 한국 기업들이 전자전시회 'CES 2023'에서 활약한 것을 알 수 있다. 2023년 12월은 소상공인진흥공단이 생성형 AI를 도입하여 공공기관의 업무 효율화를 위해 추진하고 있음을 알 수 있다.

2024년 1월, 삼성, 스마트폰 등의 키워드는 삼성전자가 AI를 탑재한 스마트폰을 출시했다는 사실을 나타내기도 하였다. 2024년 2월에는 AX, 인공지능 등의 키워드가 등장하는 것을 보아 AX의 시대가 도래하여 새로운 패러다임이 변화할 것으로 예상된다.

[표 1] 토픽 내 뉴스 단어 분포

날짜	토픽명	토픽 내 Top-10 키워드
202301	AI 기업 전쟁	AI 전쟁, 한국기업, 챗GPT, 시나리오, 참전, 선결과제, 데이터법, 차별화, 논문, 아티피셜소사이어티
202302	검색엔진 개발 열풍	검색엔진, 열풍, AI 칩, 구글, 바드, 실리콘밸리, 점유율, 어니봇, 중국, 챗GPT
202303	에듀테크 생성형 AI 접목	메타버스, 교육, 에듀테크, 웅진씽크빅, 챗봇, 경쟁, 경기GPT, 가이드라인, AI윤리법제포럼, GPT-4
202304	생성형 AI 개발 경쟁	규제, 가상인간, 클라우드, 경쟁, 챗GPT, 실리콘밸리, 소프트웨어, 개발자, 사투, 인공지능
202305	생성형 AI 리스크	생태계, 우려, 주가, 사람, 물류, 인간대체, 반도체, 한국어, 생성형AI리스크, 메타버스
202306	콘텐츠 내 생성형 AI 도입 및 보안	사이버, 수출, 반도체기업, 마케팅, 메타버스, 키즈토피아, 아이디어, 경쟁력, 콘텐츠, 보안
202307	생성형 AI 산업 투자 ETF 상장	차별화된, 플랫폼, AWS, 물적분할, 글로벌, 웹오피스, ETF, 열풍, 하반기, 상장
202308	네이버의 초대규모 AI 하이퍼클로바 X	스타트업, 결제, 네이버페이, Iaae, 하이퍼클로바X, 대항마, 네이버, 저작권, 침해, 클라우드
202309	초거대 AI 기반 디지털 생태계 조성	프라이빗AI, 생태계, 디지털, 카카오모빌리티, 채용, 데이터, 초거대AI, OTT, 인공지능, 규제
202310	삼성전자 생성형 AI '가우스' 공개	클라우드, 삼성전자, 삼성 스마트 싱스, 스타트업, 인텔, 가우스, 프롬프트, 콘퍼런스, 규제, AI활용
202311	CES	ces, 혁신상, 이커머스, 삼성전자, 반도체, ETF, 유치, 사우디, UAE, 한국기업
202312	공공기관 생성형 AI 챗봇 도입	영상제작, 클라우드, 크리스마스, 서비스, 내년, ETF, 소상공인진흥공단, 김계약주임, 업무, 개발본격화
202401	세계최초 삼성 AI 스마트폰 출시	삼성, 스마트폰, 신년기획, 자동화, 생성형, 사진, 비플라이소프트, 데이터, CES, 저작권
202402	DX를 넘어 AX시대	총선, 세무조사, 고도화, 조직개편, 인공지능, 휴메인, MWC, AX, 글로벌, nipa

3.2 검색량 급증과 관련된 이슈 분석

검색량 급증 분석을 통해 벡터 간의 유사도를 기준으로 선정된 클러스터 3개는 '기관', '분야 및 기술', '사회 및 경제'를 대표한다. 각 클러스터의 주요 단어 Top10은 [표 2]와 같

다.

[표 2] DBSCAN 클러스터별 주요 단어 Top10

Cluster1 (기관)	Cluster2 (분야 및 기술)	Cluster3 (사회 및 경제)
구글	반도체	투자
MS	디지털	학습
애플	콘텐츠	사업
정부	교육	달러
네이버	메타버스	출시
엔비디아	예술	CEO
삼성전자	챗GPT	혁신
과학기술정보통신부	클라우드	추진
카카오	오픈AI	고객
국토안보부	인공지능	업무

Cluster 1은 대표적인 기업 및 정부 기관으로 구성되어 있다. 구글, MS, 애플, 네이버, 삼성전자 등과 같은 기업뿐만 아니라 과학기술정보통신부, 국토안보부 등 정부 기관에서도 생성형 AI와 관련된 활발한 관심을 보이고 있음을 시사한다.

Cluster 2는 다양한 분야 및 기술을 대표한다. 반도체, 디지털, 콘텐츠, 교육, 메타버스, ChatGPT 등과 같은 다양한 주제와 기술이 포함되어 있어 생성형 AI와 관련된 다양한 영역에서의 관심과 연구가 진행되고 있음을 시사한다.

Cluster 3은 주로 사회 및 경제와 관련된 주제로 구성되어 있다. 투자, 학습, 사업, 달러, 출시, CEO 등이 포함되어 있으며, 생성형 AI가 사회와 경제에 미치는 영향을 확인할 수 있었다.

4. 결론

본 논문에서는 2023년 1월부터 2024년 2월까지의 뉴스 데이터를 활용하여 키워드 기반 생성형 AI 와 관련된 사회적 관심에 대한 동향을 분석하였으며, 검색량 급증과 관련된 이슈를 분석하였다. 이를 통해 전 세계의 빅테크 기업과 정부 기관이 생성형 AI에 큰 관심을 갖고 있고, 다양한 분야에서 생성형 AI 도입이 진행되고 있는 것을 파악할 수 있다.

본 연구의 시사점은 다음과 같다. 생성형 AI 서비스를 개발하고 연구하기 전에 서비스 동향과 잠재적인 리스크를 먼저 파악하는 것이 중요하다. 또한, 경쟁사가 어떤 기술을 개발하고 있는지, 어떤 기술이 점차적으로 나타나고 있는지에 대한 분석도 필요하다. 본 논문은 기업이 생성형 AI를 활용한 서비스 개발에 앞서 중요한 인사이트와 정보를 제공하여 중요한 자원으로 활용될 것을 기대한다.

참고문헌

[1] 양지훈, 양성병, 윤상혁, "생성형 AI 서비스의 성공요인에 대한 탐색적 연구: 텍스트 마이닝과 ChatGPT 를 활용하여", 경영정보학연구, 제 25권 2호, pp. 125-144, 5월, 2023년.

[2] 김동훈, 오찬희, 주영준, "다이나믹 토픽 델링 및 네트워크 분석 기법을 통한블록체인 관련 국내 연구 동향 분석", 정보관리학회지, 제 38권 3호, pp. 23-39, 9월, 2021년.

[3] 김민규, 최한수, 이민규, 박진수, 김재웅, “DBSCAN을 이용한 GPS 데이터 클러스터링”, 한국정보기술학회종합학술대회논문지, pp. 541-543, 10월, 2020년.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00242528) and by the Ministry of Education (No. RS-2023-00245529).