경영정보학연구
제25권 제3호, 2023년 8월 http://dx.doi.org/10.14329/isr.2023.25.3.163

토픽모델링과 시계열 회귀분석을 활용한 헬스케어 분야의 뉴스 빅데이터 분석 연구

Big Data News Analysis in Healthcare Using Topic Modeling and Time Series Regression Analysis

김 은 정 (Eun-Jung Kim) 한양대학교 경영학과 박사과정
장 석 권 (Suk-Gwon Chang) 한양대학교 경영대학 명예교수, 교신저자
이 상 용 (Sang-Yong Tom Lee) 한양대학교 경영대학 교수

요 약

본 연구는 디지털 헬스케어 산업 활성화를 위한 정책적 접근으로서, 주요 의제 도출 및 정책적 시사점을 제시하는데 목적이 있다. 본 연구에서는 10년(2013년~2022년) 간의 헬스케어와 관련된 뉴스 빅데이터 총 91,873건을 수집하여 토픽모델링 분석, 다차원척도 분석 및 시계열 회귀분석을 수행하였다. 토픽모델링 분석 및 다차원척도법을 통해 총 20개의 토픽을 도출하여 2차원선상에 토픽들의 군집 형태를 파악하였고, 시계열 회귀분석을 통해, 상승 추세를 나타내는 4개의 Hot topic(건강관리, 바이오제약, 기업매출・전망, 정부・정책)과 하향 추세를 나타내는 3개의 Cold topic(스마트기기, 주식・투자, 도시・건설)을 도출되었다. 본 연구의 결과는 우리나라 정책을 수립하는 정부 기관에 중요한 기초 자료로 활용될 수 있을 것이다.

키워드 : 디지털 헬스케어, 토픽모델링, LDA, 시계열회귀분석, 데이터마이닝

Ⅰ. 서 론

최근 COVID-19로 인해 전 세계적으로 질병 예방 및 건강 관리에 대한 관심이 증가하면서 헬스케어 시장은 미래 유망 융합 산업으로 크게 각광받고 있다(IRS Global, 2020). 다보스 세계경제포럼(WEF)에서 발표한 2021년 10대 미래 유망기술 가운데 바이오・헬스케어 관련 기술이 큰 비중을 차지함에 따라, ICT와 바이오기술의 융합이 건강에 대한 접근과 삶의 방식을 크게 변화시킬 것이며, 막대한 사회・경제적 효과를 창출할 것으로 전망하였다(WEF, 2021). 또한 주요국들은 안정적인 인프라 및 우호적인 정책을 바탕으로 디지털 헬스케어 시장의 경쟁적 투자를 확대하고, 다양한 의료 서비스 확장에 집중하며 시장 선점에 힘쓰고 있다(김유진, 2022). 최근 우리 정부도 디지털 헬스케어 산업을 활성화하기 위해 다양한 전략을 수립하고, 생태계 조성・지원, 규제 합리화를 위한 정책을 추진하고 있으며 관련 산업계, 학계, 연구계에서도 산업 육성을 위한 R&D, 제품・서비스 개발, 정책 개발, 규제 완화 등을 위한 다방면의 활동을 추진하고 있다. 그러나, 헬스케어 시장의

활성화 필요성이 전 세계적으로 부각됨에도 불구하고 우리나라는 개인정보 문제, 생체 데이터 수집, 빅데이터 취급, 의료 사고의 책임 소재, 다양한 이해관계들 간의 법적인 문제들로 인해 시장의 즉각적인 활성화와 성장을 기대하기 어려운 상황이다(이경은, 2021). 즉, 우리나라의 헬스케어 산업 생태계 구축 및 활성화를 위해서는 면밀한 규제 · 제도 개선이 시급한 상황이라고 할 수 있다. 특히, 팬데믹 시기를 지나오면서 우리나라는 질병 예방에 대한 욕구가 크게 증대하였고, 초고령 사회 진입을 앞두고 건강관리에 대한 인식과 니즈도 점차 증가하여, 기존의 병원 · 치료 중심의 의료 산업 체제의 본격적인 대전환이 불가피한 상황이다. 따라서, 새로운 의료 모델의 본격적인 확장이 필요하며, 이를 위해 우선적으로 도입이 가능한 분야, 수요가 있는 분야에 대한 적극적인 정책적 지원과 더불어 다양한 사례를 상당수 구축하는 것이 중요할 것이다(이경은, 2021). 또한 디지털 헬스케어를 둘러싼 최근 이슈, 트렌드, 소비자들의 인식 정도 등의 검토를 통해 우리나라에서 현실적으로 추진 가능한 발전 방향에 대한 논의가 필요하다.

이에 본 연구에서는 헬스케어 산업의 활성화를 위한 정책적 접근으로서, 최근 사회적으로 형성되고 있는 헬스케어와 관련한 주요 이슈가 무엇이며, 도출된 주제들이 정책적으로 어떠한 시사점을 제공할 수 있는지에 대해 분석해보고자 한다. 이를 위해 뉴스 빅데이터를 수집하여 텍스트마이닝 기법 중 토픽모델링 분석을 수행하여 국내 주요 이슈들을 파악하고, 도출된 토픽의 유형별 군집화를 위해 다차원척도법을 활용하여 시각화 분석을 수행하였다. 최종적으로 Hot/Cold 토픽을 선별하기 위해 시계열 선형회귀분석을 수행하였으며, 연구결과를 토대로 주요 토픽별 정책적 시사점을 제안하였다. 이에 본 연구의 연구문제는 다음과 같다.

연구문제 1: 헬스케어 관련 뉴스 빅데이터에 나타난 주요 토픽과 키워드는 무엇인가?

연구문제 2: 헬스케어 관련 뉴스 빅데이터에 도출된 주요 토픽이 지니는 의미와 시사점은 무엇인가?

Ⅱ. 선행연구 고찰

2.1 헬스케어 관련 연구동향

디지털 헬스케어 분야는 디지털 전환 시대를 맞이하여, 첨단기술과의 융합을 통해 병원 및 치료 중심의 기존 의료 체계에서 건강 관리와 질병 예방과 관련한 새롭고 다양한 의료 모델의 확장을 이루어 내고 있다(IRS Global, 2020).

최근 헬스케어와 관련한 국내 · 외 동향 분석, 정책 현황 분석, 발전 방향 수립 등의 분석보고서들이 전문기관별로 꾸준히 발표되고 있으며, 다양한 분석을 바탕으로 디지털 헬스케어 분야의 육성 필요성을 제안하는 정책 연구들이 늘어나고 있다.

안정민(2021)의 연구에서 디지털 헬스케어 산업의 각종 유발효과를 분석하였는데, 생산 측면, 부가가치 측면, 근로자 가계 측면, 기업의 이익 측면에서는 전체 산업의 평균보다 디지털 헬스케어 산업의 유발계수가 크다고 분석하였다. 반면, 정부의 재정 수입과 연계된 생산세유발계수는 전체 평균보다 작게 나타났는데, 이는 산업의 시스템 및 인프라 구축, 정보의 표준화 등의 수준이 미흡하기 때문이라고 분석하였다. 권승수(2021)는 헬스케어 분야에 빅데이터 활용을 위해서는 보안 및 법적 문제를 해소하기 위한 규제와 정책의 전반적인 개선이 필요하며, 무엇보다 의료정보 서비스 패러다임의 변화가 필요하다고 주장하였다. 권기대(2022)의 연구에서는 디지털 헬스케어 분야의 인식 조사를 통해 연령별로 선호하는 디지털 헬스케어 제품 · 서비스가 다르고, 인식의 차이가 있다고 밝혔다. 결과적으로 헬스케어 제품에 대한 접근성을 강화하고 새로운 경험을 제공하면 헬스케어 시장은 지금보다 빠르게 성장할 수 있을 것이라고 시사하였다.

이처럼 최근 연구들은 헬스케어 시장의 성장 가능성, 육성 필요성을 주장하면서, 의료 영역의 디지털 전환에 따른 규제와 정책 마련을 비롯해 건강 데이터 활용 및 보건의료 데이터 플랫폼 구축 등에 대한 대책 마련의 시급성을 함께 주장하고 있다. 이 밖에도 디지털 헬스케어 연구 중 서비스의 분류 연구(하소희 등, 2020), 특허 기반 주요 토픽을 도출한 연구(김은정 등, 2022), 디지털 헬스케어 연구 동향을 분석한 연구(이택균, 2020) 등 다양한 연구들이 발표되고 있는데, 공통적으로 빅데이터, 인공지능, 블록체인, 플랫폼 등과 같이 바이오와 ICT의 기술 융합에 대한 동향 분석 등 헬스케어 분야의 미래 발전 방안에 대한 시사점을 제시하고 있다.

2.2 뉴스 빅데이터 분석 동향

뉴스 기사는 사회 전반에서 발생되는 다양한 이슈들을 반영하고 있으며, 무엇보다 객관적인 사실과 중요한 정보를 정확하게 전달하는 기능에 초점을 두고 있어 정책 의제를 도출하는데 중요한 역할을 할 수 있다. 이뿐만 아니라, 뉴스는 전문성이 높은 정보를 빠르게 전달하는 매체로서 미래신호를 탐색하여, 사회적 이슈를 파악하는데 중요한 연구 데이터가 될 수 있다(나경식 등, 2018).

뉴스 데이터는 서로 완전히 독립된 형태가 아닌 시간의 흐름에 따라 데이터간 영향을 받는 시계열 데이터이다. 즉, 과거의 뉴스가 현재의 뉴스에 영향을 주기 때문에 사회적인 흐름, 트렌드, 이슈를 파악하고 미래를 예측하는 데 중요한 역할을 할 수 있다. 특히 최근에는 빅데이터를 활용한 미래 예측 연구가 증가하면서 데이터베이스가 잘 구축된 뉴스를 이용한 연구들이 지속적으로 늘어나고 있는 추세이다(박대민, 2016).

노희경(2022)은 뉴스 빅데이터 분석을 통해 메타버스를 관광 분야의 공공 및 민간 비즈니스 영역에 어떻게 적용해야 하는지에 대한 방향성을 제시하였다. 최한별(2022)의 연구에서는 뉴스 빅데이터를 활용하여 정보 프라이버시와 관련한 사회적 이슈를 도출하고 이를 시기별로 나누어 그 변화를 확인하였다. 김태종(2022)은 세계적으로 큰 관심을 갖고 있는 디지털 전환 현상을 거시적 차원에서 분석하고, 이에 대한 시사점을 제안하기 위해 뉴스 빅데이터를 활용하여 토픽모델링을 수행하였다. 이 밖에도 탄소중립, ESG, 코로나 19 등 최근 이슈화 되고 있는 사회적인 문제와 이슈에 대해 분석하는 뉴스 빅데이터 연구들이 증가하고 있다(고민규 등, 2023; 안지연 등, 2021; 최은경 등, 2022). 이처럼 뉴스 빅데이터를 활용한 연구들은 전반적으로 사회 전반에 걸쳐 불확실하고 복잡한 이슈에 대한 사회적인 논의 구조 및 담론을 도출하여 중요하면서 다양한 시사점을 제시한다.

Ⅲ. 연구방법

3.1 분석데이터

본 빅데이터 분석 서비스인 '빅카인즈(www.bigkinds.or.kr)'를 활용하여 분석 데이터를 수집하였다. '빅카인즈'는 한국언론진흥재단에서 개발한 뉴스 분석 서비스로, 종합일간지, 경제지, 지역일간지, 방송사 등 다양한 언론사로부터 뉴스를 수집하여 다양한 빅데이터 분석 결과를 제공한다. 본 연구에서는 연구 목적에 맞게 언론보도기사 본문에 '헬스케어'가 검색되는 뉴스를 수집하여 분석을 실시하였고, 총 10년(2013.01.01. ~ 2022.12.31.) 간의 데이터를 활용하였다. 전국일간지 11개, 경제일간지 8개, 지역일간지 28개, 방송사 5개, 전문지 2개 총 54개 언론사를 대상으로 뉴스를 수집하였다.

데이터의 신뢰성 확보를 위해 전문가 협의를 통해 같은 내용을 신문사마다 중복 게시한 기사와 기사 본문에 키워드 '헬스케어'가 존재하지만 기사의 핵심 내용은 헬스케어와 관련이 없고, '헬스케어' 키워드가 1회만 언급된 기사의 경우 분석 대상에서 제외하였다. 결론적으로 총 91,873개의

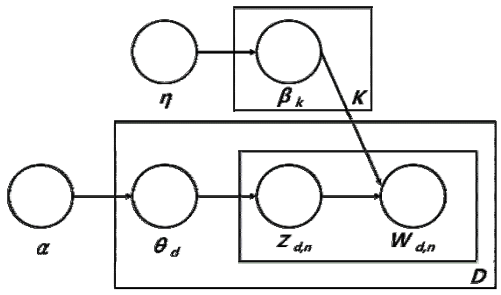
뉴스 데이터를 분석에 활용하였다.

3.2 분석방법

본 연구에서는 파이썬(Python)을 활용하여 토픽 모델링 분석, 다차원척도 분석 및 시계열 회귀분석을 수행하였다. 본격적인 분석에 앞서 비정형화된 데이터를 분석에 알맞은 형태로 변환하고, 불필요한 정보를 필터링하기 위해 전처리 과정을 수행한 후, 파이썬 한글 형태소 분석 패키지인 'KoNLP'로 보통 및 고유명사만을 추출하였다.

텍스트마이닝 분석 방법론 중 잠재적 의미와 주제를 찾아 내주는 '토픽모델링(Topic Modeling)' 기법은 대량의 문서 집단에서 단어 리스트를 구성하여 문서 내에 존재하는 유의미한 토픽을 추출한다. 토픽모델링 분석기법은 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)기법으로, 확률적 방법으로 각 토픽에 해당될 가능성이 높은 단어들을 집합 형태로 추출하는 알고리즘이다(David *et al.*, 2003).

David(2003)에 의해 소개된 LDA 모델은 관찰된



α: 문서별 토픽 k의 Dirichlet prior weight, θ값을 결정하는 파라미터
θ_d : 문서별 토픽의 비율
$Z_{d,n}$: 문서 d의 n번째 단어의 토픽(index)
$W_{d,n}$: 문서 d의 n번째 단어(문서에 관측되는 변수)
D: 문서 집합
η: 토픽별 단어 w의 Dirichlet prior weight
β값을 결정하는 파라미터
β_k: 토픽별 단어 w의 생성확률
K: 토픽의 개수

〈그림 1〉 LDA 모델

변수인 문서, 단어 등을 통해 전체 문서집합의 주제(β), 각 문서별 주제 비율(θ), 각 단어들이 각 주제에 포함될 확률(Z) 등을 문서의 구조에서 보이지 않는 변수를 파악하여 분석한다. θ는 디리클레(Dirichlet) 분포를 따르며, θ을 따라 문서 내에 존재하는 단어들의 주제인 Z가 결정된다. 또한 Z와 θ값에 따라 단어 W가 결정된다. 어떤 문서에 대해 파라미터 θ(주제 벡터)가 있고, 앞에서부터 단어를 하나씩 채울 때마다 θ로부터 하나의 주제를 선택하고, 다시 그 주제로부터 단어를 선택하는 방식으로 문서 생성 과정을 모델링하는 것이다.

본 연구에서 활용될 두 번째 분석 방법론은 다차원척도법(Multidimensional Scaling, MDS)으로 케이스 간의 거리를 분석하여 이들 간의 관계 구조와 연관성을 시각적으로 표현하는 통계 데이터 분석기법이다. MDS 기법은 다차원 공간상에서 유사성이 큰 대상들은 가깝게, 유사성이 작은 대상들은 상대적으로 멀게 위치시켜 그 결과를 얻는다(Borg *et al.*, 2005). 본 연구에서는 MDS 분석기법을 통해 도출된 토픽들의 연관성을 분석하여 좌표상에 토픽들 간의 위치를 시각화하였다.

마지막으로 시계열 회귀분석은 시간의 흐름에 따라 관측된 시계열 변수들 간의 함수관계를 통계적으로 접근하는 방법론으로서, 선형적인 상관성을 가진 변수 간의 인과관계를 증명한다. 본 연구에서는 선형회귀모델을 통해 시계열 형태의 독립변수(x)와 종속변수(y) 사이의 선형적 관계를 파악하고, 추정된 회귀식을 활용하여 독립변수가 종속변수에 미치는 영향을 분석한다. 이는 아래의 식(1)과 같이 표현할 수 있다.

$$Y_t = \beta_o + \beta_1 X_t + \epsilon_t \quad (1)$$

Y_t = 종속변수, 반응변수
X_t = 독립변수, 예측변수, 설명변수
β_0 = 절편
β_1 = 회귀계수
ϵ = 오차항

본 연구에서는 세 가지 분석 방법론을 활용하여, 헬스케어와 관련한 뉴스 빅데이터를 분석하였다. 분석 절차는 다음 <그림 2>와 같다.

Step 1	데이터 수집 및 선정 (총 91,873개 데이터)
⇓	
Step 2	데이터 전처리 · 정제
⇓	
Step 3	토픽모델링(LDA) 분석, 다차원척도법(MDS) 분석
⇓	
Step 4	시계열 회귀분석
⇓	
Step 5	Hot/Cold topic 도출
Step 6	정책적 시사점 제시

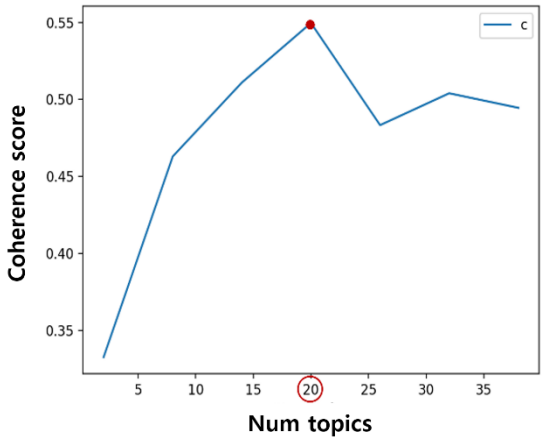
〈그림 2〉 분석 절차

Ⅳ. 연구결과

4.1 토픽모델링(LDA) 분석 결과

LDA분석을 수행하기 전에 토픽의 최적의 개수를 판단하기 위해, ‘Coherence score’ 평가를 수행하였다. ‘Coherence Score’은 추출된 각각의 토픽들이 얼마나 높은 유사도를 나타내는지를 평가하는 방법으로, Coherence Score가 높을수록 각각의 토픽들이 의미를 갖는 유사한 단어들로 분류되었다고 해석한다(Frank *et al.*, 2013). 본 연구의 데이터들의 ‘Coherence Score’를 측정한 결과 최적의 토픽 수는 20개임(score: 0.55)을 확인하였다(<그림 3>).

토픽모델링 분석결과로 30개의 핵심 키워드로 구성된 20개 토픽이 도출되었으며, 각 토픽별로 높은 빈도를 나타내는 상위 10개 키워드만 <표 1>과 같이 정리한 후 토픽의 키워드 및 뉴스 원문을 검토하여 종합 토픽명을 부여하였다.



〈그림 3〉 Coherence Score 측정 결과

최종적으로 ‘병원 · 진료’, ‘기업성장’, ‘보험 · 금융’, ‘도시 · 건설’, ‘스마트기술’, ‘정부 · 정책’, ‘기업전망 · 매출’, ‘제품출시’, ‘주식 · 투자’, ‘스마트기기’, ‘바이오제약’, ‘주식시장’, ‘건강관리’, ‘질병 · 치료’, ‘대학 · 연구’, ‘지역조성’, ‘헬스케어제품’, ‘해외진출’, ‘기업경영’, ‘창업 · 스타트업’과 같이 헬스케어와 관련한 다양한 토픽이 도출되었다. 토픽의 비중은 ‘지역조성(9.32%)’, ‘스마트기술(7.85%)’, ‘주식시장(7.73%)’, ‘병원 · 진료(7.58)’, ‘창업 · 스타트업(6.67%)’, ‘바이오제약(5.88%)’, ‘헬스케어제품(5.87%)’, ‘스마트기기(5.56%)’, ‘정부 · 정책(4.58%)’, ‘기업성장(4.45%)’, ‘주식 · 투자(4.0%)’, ‘제품출시(3.94%)’, ‘대학 · 연구(3.91%)’, ‘기업경영(3.69%)’, ‘보험 · 금융(3.63%)’, ‘기업전망 · 매출(3.54%)’, ‘건강관리(3.23%)’, ‘도시 · 건설(3.02%)’, ‘해외진출(3.02%)’, ‘질병 · 치료(2.54%)’ 순으로 높게 나타났다.

본 연구의 5장에서 다룰 정책적 시사점 도출을 위해 각 토픽에 속하는 기사들을 검토한 후 토픽의 주요 주제와 핵심 내용을 아래 <표 2>와 같이 정리하였다.

〈표 1〉 토픽모델링(LDA) 분석 결과

구분	Topics									
	1	2	3	4	5	6	7	8	9	10
	병원 · 진료	기업성장	보험 · 금융	도시 · 건설	스마트 기술	정부 · 정책	기업전망 · 매출	제품출시	주식 · 투자	스마트 기기
1	병원	그룹	보험	도시	데이터	정부	증가	제품	펀드	스마트
2	환자	회사	금융	시설	인터넷	규제	매출	식품	수익	기기
3	진료	상장	상품	스마트	인공지능	정책	성장	특허	주식	삼성전자
4	진단	인수	고객	주택	혁명	경제	이익	피부	운용	스마트폰
5	기기	규모	생명	건설	미래	대통령	대비	건강	자산	제품
6	검사	업체	가입	시티	통신	국민	전망	인증	상품	웨어러블
7	원격	합병	은행	제주	플랫폼	제도	영업	화장품	투자자	전자
8	디지털	지분	보장	조성	디지털	후보	기록	사용	해외	애플
9	기관	설립	카드	센터	사물	위원회	예상	생산	증권	센서
10	솔루션	성장	관리	지역	스마트	일자리	실적	출시	금융	모바일
비중 (개)	7.58% (6,965)	4.45% (4,087)	3.63% (3,332)	3.02% (2,778)	7.85% (7,215)	4.58% (4,205)	3.54% (3,250)	3.94% (3,622)	4.0% (3,676)	5.56% (5,112)

구분	11	12	13	14	15	16	17	18	19	20
	바이오 제약	주식시장	건강관리	질병 · 치료	대학 · 연구	지역조성	헬스케어 제품	해외진출	기업경영	창업 · 스타트업
1	바이오	지수	건강	치료	교육	지역	고객	중국	회장	스타트업
2	제약	상승	관리	환자	대학	센터	제품	해외	경영	창업
3	의약품	증시	모바일	진단	교수	추진	바디	수출	사회	벤처
4	약품	업종	운동	질환	과학	육성	안마	진출	사장	행사
5	제약사	하락	헬스	질병	연구	구축	소비자	일본	직원	혁신
6	생산	종목	개인	왓슨	융합	경제	상품	현지	본부	개최
7	치료제	코스닥	생활	수술	공학	조성	구매	경제	회사	참여
8	임상	금리	검진	고령	학생	미래	디자인	유럽	그룹	참가
9	신약	포인트	프로그램	노인	인력	혁신	스포츠	업체	부회장	프로그램
10	계약	코스피	상담	재활	기술	활성화	판매	협력	사람	포럼
비중 (개)	5.88% (5,400)	7.73% (7,104)	3.23% (2,971)	2.54% (2,329)	3.91% (3,589)	9.32% (8,558)	5.87% (5,392)	3.02% (2,770)	3.69% (3,394)	6.67% (6,124)

〈표 2〉 토픽모델링(LDA) 분석 결과 요약

토픽	토픽명	설명
1	병원 · 진료	· (주요뉴스) AI의료기기 산업, 진단 솔루션, 원격의료, 의료기기, 디지털 의료서비스, 질병 예측 솔루션 등 관련 뉴스 · (종합) 의료산업이 전통적인 하드웨어 산업에서 디지털(소프트웨어) 산업으로 변모
2	기업성장	· (주요뉴스) 기업 인수합병, 바이오 기업 성장, 코스닥 상장, 헬스케어 사업 확대, 유상증자, 투자유치, 기업 성장성 등 관련 뉴스 · (종합) 바이오, 디지털 헬스케어 분야 기업의 미래 기업가치 증대와 성장 가능성 기대
3	보험 · 금융	· (주요뉴스) 생명보험산업, 금리 하락, 금융산업, 보험제도 개선 등 관련 뉴스 · (종합) 인슈어테크의 활성화, 헬스케어 서비스와 보험의 결합 등 미래 성장동력 발굴
4	도시 · 건설	· (주요뉴스) 헬스케어 클러스터, 바이오 융합기술단지, 스마트시티, 미래도시 조성, 정책 활성화, 협력체 마련 등에 대한 뉴스 · (종합) 바이오 · 의료 클러스터의 조성과 활성화, 바이오 · 의료 산업에 특화된 중개연구 활성화와 투자 유치, 지역 클러스터와의 협력과 연계 강화 추진
5	스마트기술	· (주요뉴스) 신기술 전시회, 기술동향, AI 플랫폼개발, 신사업 진출, 스마트기술 접목, 빅데이터 분석 등 관련 뉴스 · (종합) 헬스케어 산업에 스마트기술 접목하여 미래 신산업 개척, 신성장 동력 발굴
6	정부 · 정책	· (주요뉴스) 정부 규제, 의료데이터 확보 · 활용, 데이터3법, 의료서비스 혁신 방안에 대한 뉴스 · (종합) 정부 규제로 미래 의료시장에서 국가 경쟁력 퇴보, 시급한 정부 정책 규제 개혁 필요
7	기업전망 · 매출	· (주요뉴스) 수출 증가, 헬스케어 분야 투자 확대 및 수익창출 기대, 기업의 성장률, 헬스케어 영역 확대 및 매출 증가 등 관련 뉴스 · (종합) 헬스케어 분야 진출 기업 성장세는 빠르게 증가하는 추세로 투자 규모 확대
8	제품출시	· (주요뉴스) 헬스케어 신제품, 건강보조식품, 의료기기, 웨어러블 기기, 화장품, 웰빙, 기능성 제품 등 헬스케어 관련 신제품 홍보 뉴스 · (종합) 건강에 대한 관심의 증가로 건강관리를 위한 다양한 제품 출시
9	주식 · 투자	· (주요뉴스) 주식매수, 주식투자, 유망투자상품, 투자유치 기업 해외진출, 헬스케어 펀드, 제약회사의 수익률, 유망 스타트업 투자와 관련한 뉴스 · (종합) 4차 산업혁명 관련 분야에서 경쟁력을 보이는 바이오 혁신기업에 투자 증대
10	스마트기기	· (주요뉴스) AI기반 가전제품, 헬스케어 앱, 생체신호 측정 센서개발 및 웨어러블 기기, 스마트워치, 헬스케어 로봇, 반려동물 보조기기 관련 연구개발 및 제품 등에 관한 뉴스 · (종합) 혁신 스마트 기술 기반 헬스케어 분야 제품 다양화로 헬스케어 산업 활성화 기대
11	바이오제약	· (주요뉴스) 바이오 신약개발, 임상 성공, 글로벌 제약사와 MoU, AI 기반 신약개발, 연구개발 투자, 치료제 개발 등과 관련한 뉴스 · (종합) 지속적인 연구개발 투자와 품질 혁신 등으로 제약 · 바이오산업 활성화 추진
12	주식시장	· (주요뉴스) 코스닥 주식 시장, 시장경제 동향, 주식 시장 상승 · 하락과 관련한 전망에 관한 뉴스 · (종합) 수익을 내기 어려운 바이오 관련 주식 시장의 둔화로 장기 투자 전략 강조
13	건강관리	· (주요뉴스) 모바일 헬스케어 앱, 건강관리 플랫폼 · 서비스 출시 등 관련 뉴스 · (종합) 의료서비스의 패러다임의 변화로 질병 예방 및 건강 관리 관련 서비스 지속 확대 추세
14	질병 · 치료	· (주요뉴스) 디지털치료제, 수술용 로봇, 치매치료, 고령화 관련 질병에 대한 치료제 관련 뉴스 · (종합) 헬스케어 디지털 솔루션과 콘텐츠의 지속적 연구개발로 디지털 헬스케어 시장 확대 기대
15	대학 · 연구	· (주요뉴스) 헬스케어 사업단, 융합전공, 연구개발의 성과, AI기반 교육 활성화 관련 뉴스 · (종합) AI 중심의 새로운 융합 연구를 추진, 헬스케어 분야의 연구 활성화 및 인재 육성 강화
16	지역조성	· (주요뉴스) 지역활성화, 헬스케어 산업 및 투자 활성화, 헬스케어 플랫폼 구축 관련 뉴스 · (종합) 지역 중심 헬스케어 산업 육성 및 기업 지원 역할 강화
17	헬스케어제품	· (주요뉴스) 건강관리와 관련한 기기(안마의자, 눈, 운동 등 바디 중심) 관련 뉴스 · (종합) 건강에 대한 관심의 증가로 건강관리를 위한 다양한 제품 출시
18	해외진출	· (주요뉴스) 자유무역, 경제 정책, 글로벌 협력, 해외 진출 등 관련 뉴스 · (종합) 헬스케어 산업, 글로벌 협력 확대로 유망기업의 해외 진출 증대
19	기업경영	· (주요뉴스) 기업경영 환경, 신사업 추진, 대표이사 선임, 조직개편, 혁신경영 등에 관한 뉴스 · (종합) 디지털 전환을 위한 기업의 경영체계 전환
20	창업 · 스타트업	· (주요뉴스) 벤처투자 증대, 액셀러레이팅 강화, 전시회 참석, 유망 스타트업 등에 관한 뉴스 · (종합) 대기업에 비해 유연하고 혁신의 속도가 빠른 유망 스타트업의 투자 강화

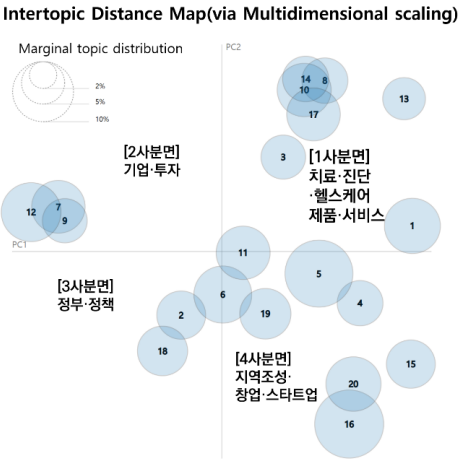
4.2 다차원척도법(MDS) 분석 결과

파이썬의 토픽모델링 시각화 도구인 LDAvis 패키지는 다차원척도법(MDS) 알고리즘을 사용하여 토픽들의 군집 형태를 2차원 척도로 시각화한다. MDS 기법은 차원 축소를 통해 조금 더 종합적인 해석을 가능하게 한다. 따라서, MDS 분석을 통해 앞서 파악한 20개의 토픽을 크게 4개의 종합 주제로 분류해봄으로써 헬스케어와 관련한 우리나라의 중요한 종합 이슈를 파악할 수 있었다.

MDS 분석을 수행한 결과, 1사분면(38.23%)에 가장 많은 토픽들이 배치되었다. 하위 토픽으로는 '병원 · 진료', '바이오제약', '헬스케어제품', '스마트 기기', '제품출시', '보험 · 금융', '건강관리', '질병 · 치료'가 있으며, 배치된 토픽들의 특성을 고려하여 '치료 · 진단 · 헬스케어 · 제품 · 서비스'로 종합 토픽명을 부여하였다. 그 다음으로 가장 높은 비중을 차지한 4사분면은 '지역조성', '스마트기술', '창업 · 스타트업', '대학 · 연구', '기업경영', '도시 · 건설' 관련 하위 토픽들로 구성되어 '지역 · 조성 · 창업 · 스타트업(34.64%)'으로 종합 토픽명을 부여하였다. 2사분면은 '주식시장', '주식 · 투자', '기업전망 · 매출'과 관련한 토픽들로 구성되어, 종합 토픽명을 '기업 · 투자(15.27%)'로 부여하였다. 마지막으로 가장 낮은 비중을 차지한 3사분면은 '정부 · 정책', '기업성장', '해외진출' 관련 하위 토픽이 군집화 되어 종합 토픽명을 '정부 · 정책(12.05%)'으로 명명하였다.

〈표 3〉 MDS 분석 결과 상세분류

구분	주제	Topic 분류
1사분면	치료 · 진단 · 헬스케어 제품 · 서비스 (38.23%)	토픽1(7.58%): 병원 · 진료 토픽11(5.88%): 바이오제약 토픽17(5.87%): 헬스케어제품 토픽10(5.56%): 스마트기기 토픽8(3.94%): 제품출시 토픽3(3.63%): 보험 · 금융 토픽13(3.23%): 건강관리 토픽14(2.54%): 질병 · 치료
2사분면	기업 · 투자 (15.27%)	토픽12(7.73): 주식시장 토픽9(4.0%): 주식 · 투자 토픽7(3.54%): 기업전망 · 매출
3사분면	정부 · 정책 (12.05%)	토픽6(4.58%): 정부 · 정책 토픽2(4.45%): 기업성장 토픽18(3.02%): 해외진출
4사분면	지역조성 · 창업 · 스타트업 (34.46%)	토픽16(9.32%): 지역조성 토픽5(7.85%): 스마트기술 토픽20(6.67%): 창업 · 스타트업 토픽15(3.91%): 대학 · 연구 토픽19(3.69%): 기업경영 토픽4(3.02%): 도시 · 건설



〈그림 4〉 다차원척도 분석(MDS) 결과

4.3 시계열 회귀분석 분석 결과

본 연구에서는 시계열 뉴스 데이터를 기반으로 선형회귀모델을 활용하여 Hot topic과 Cold topic을 선별하였다. 뉴스의 연도별로 토픽들의 평균 비중을 y로, 시계열 데이터인 연도를 x로 설정한 후, 유의확률 값이 유의미하면서, 회귀계수가 양수(+)로 나타나는 것을 Hot topic, 음수(-)의 회귀계수로 나타나는 것은 Cold topic으로 구분하였다. Hot topic은 시간이 지남에 따라 상승 추세를, Cold topic은 시간이 흐름에 따라 하향 추세를 나타낸다

〈표 4〉 시계열 회귀 분석 결과

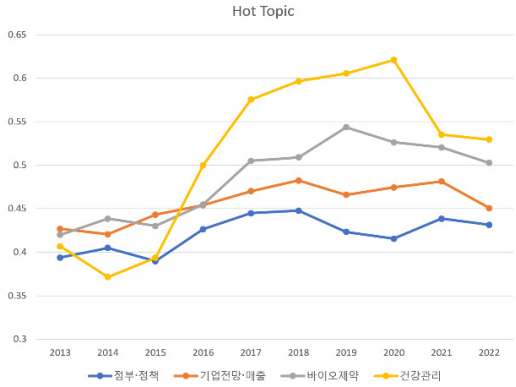
주제	토픽	p-value	회귀계수	Hot/Cold
치료・진단・헬스케어 제품・서비스 (38.23%)	토픽1(7.58%): 병원・진료	0.5764	-0.0008	-
	토픽11(5.88%): 바이오제약	**0.0019****	0.0125	**Hot**
	토픽17(5.87%): 헬스케어제품	0.1454	-0.0044	-
	토픽10(5.56%): 스마트기기	0.0162*	-0.0073	**Cold**
	토픽8(3.94%): 제품출시	0.3359	-0.0024	-
	토픽3(3.63%): 보험・금융	0.8861	-0.0004	-
	토픽13(3.23%): 건강관리	**0.0152***	**0.0226**	**Hot**
	토픽14(2.54%): 질병・치료	0.1724	-0.0029	-
기업・투자 (15.27%)	토픽12(7.73): 주식시장	0.9941	0	-
	토픽9(4.0%): 주식・투자	0.0072*	-0.0076	**Cold**
	토픽7(3.54%): 기업전망・매출	**0.0208***	**0.0051**	**Hot**
정부・정책 (12.05%)	토픽6(4.58%): 정부・정책	**0.0498***	**0.0042**	**Hot**
	토픽2(4.45%): 기업성장	0.3755	0.0029	-
	토픽18(3.02%): 해외진출	0.5412	-0.0009	-
지역조성・창업・스타트업 (34.46%)	토픽16(9.32%): 지역조성	0.086	0.0047	-
	토픽5(7.85%): 스마트기술	0.6434	-0.001	-
	토픽20(6.67%): 창업・스타트업	0.1205	0.0044	-
	토픽15(3.91%): 대학・연구	0.2445	-0.002	-
	토픽19(3.69%): 기업경영	0.8173	-0.0003	-
	토픽4(3.02%): 도시・건설	0.0381*	-0.0041	**Cold**

*p<0.05, **p<0.01.

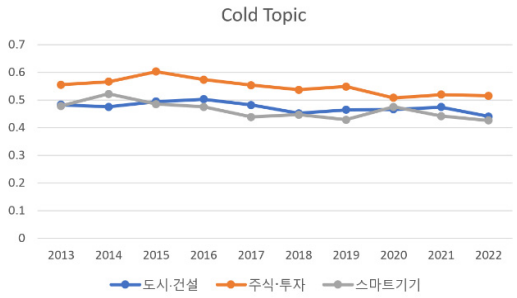
고 판단할 수 있으며, 유의수준이 유의미하지 않은 토픽은 중립 토픽으로 구분하였다. 시계열 회귀분석 결과는 <표 4>와 같다.

시계열 회귀분석 결과, 회귀계수가 양수(+)로 나타나는 Hot topic은 '바이오제약', '건강관리', '기업전망・매출', '정부・정책' 총 4개로 도출되었으며, 회귀계수가 음수(-)인 Cold topic은 '스마트기기', '주식・투자', '도시・건설' 총 3개 그리고 나머지 13개 토픽은 중립 토픽으로 도출되었다.

Hot topic에 해당하는 4개 토픽에 대해 10년간의 연도별 토픽 비중의 변화 추이를 확인하였다(<그림 5>). '건강관리'에 관련한 토픽의 비중이 2015년을 기점으로 가파르게 상승하다가 2020년부터 하향하고, '바이오제약'과 관련한 토픽의 비중은 2016년을 기점으로 상승하다 2019년에 하향한다. '기업전망・매출' 관련 토픽 의 비중은 2018년 한해 하향하다 회복하며, '정부・정책' 관련 토픽의 비중은 2018년과 2020년 간 하향하다 회복하는 것을 확인할 수 있다. 3개 Cold topic은 10년간 지속적으로 소폭 하향하는 추세를 보인다(<그림 6>).



〈그림 5〉 Hot topic 연도별 토픽 비중 변화



〈그림 6〉 Cold topic 연도별 토픽 비중 변화

Ⅴ. 토의 및 시사점

본 연구에서는 Hot topic으로 도출된 4가지(건강관리, 바이오제약, 기업전망 · 매출, 정부 · 정책) 토픽을 중심으로 정책적 시사점을 제시하고자 한다.

5.1 Hot topic: 건강관리

최근 정부에서는 디지털 헬스케어 산업 육성을 위해 첫 번째 핵심 과제로 '의료 · 건강 · 돌봄 서비스'를 선정하여 세부적인 추진 및 투자 계획을 수립하였다. 본 연구에서 Hot topic으로 도출된 '건강관리' 분야가 가파르게 상승 추세를 보인 것을 보더라도 현재 우리나라에서는 정책적으로나 사회적으로 해당 토픽은 크게 이슈화되고 있다는 것을 유추해 볼 수 있다(보건복지부, 2023).

개인의 건강 관리를 위한 서비스 개발을 위해서는 웨어러블 기기, 빅데이터 수집, AI 예측 기술, 어플리케이션 등 다양한 기술의 접목이 필요하고, 아울러 산업 생태계 패러다임의 변화도 필요하다. 기존에는 병원, 제약회사, 의료기기 업체, 보험회사, 환자를 중심으로 의료 생태계가 구성되었다면, 변화된 패러다임에서는 수요자 중심의 '자가 건강 측정 · 관리' 트렌드가 확산됨에 따라 건강관리 전문서비스사, 어플리케이션 개발사, 정보통신서비스 기업, 웨어러블 기기 제조사, 플랫폼 서비스 개발사, 콘텐츠 개발사 등 다양한 이해관계자가 모인 새로운 생태계 구축이 필요하다. 즉, 헬스케어 산업은 타 산업과 달리 생산자와 소비자 간의 구도가 복잡하기 때문에 이해관계자들 간의 협업 및 상생이 중요하다. 이에 정부 차원에서 디지털 헬스케어 생태계 조성을 위한 공동연구, 협약, MOU, 인수 · 합병 등이 이루어질 수 있도록 '협업의 장'을 마련하는 기업 상생 지원 정책을 확대해야 할 것이다(정일영 등, 2021).

5.2 Hot topic: 바이오제약, 기업전망 · 매출

최근 정부에서 제4차 산업혁명과 맞물려 국가 신약개발사업에 가장 큰 규모의 R&D 지원을 할 것이라고 발표했다(메디컬타임즈, 2023). 두 번째와 세 번째 Hot topic으로 도출된 '바이오제약', '기업전망 · 매출' 토픽에 해당하는 기사들을 세부적으로 살펴보면, 신약과 제약과 관련한 뉴스가 다수 분류된 것을 확인할 수 있었다. 이는 여전히 헬스케어의 중심엔 신약개발의 중요성과 필요성이 크게 자리 잡고 있다고 유추해볼 수 있다.

최근 국내 대형 제약 바이오 기업들은 디지털 의료기기, 건강관리 서비스, AI솔루션 개발을 위해 파트너사와 MoU, M&A, 투자 활동을 확대하면서 첨단기술을 활용한 디지털 헬스케어 분야의 신사업 진출에 박차를 가하고 있다(데일리팜, 2023). 또한 국내 벤처캐피탈(Venture Capital)들은 차세대 바이오 신약을 개발하는 기업들에 대한 투자를 2015년 22.9% 규모에서 2019년 58.4% 수준까지 확대하면서, 바이오 신약개발 기업들의 성장성에 크게 주목하고 있다(홍미영 등, 2021). 다만, 많은 투자기관들이 신약분야는 유망 시장분야임이 틀림없고, 타 산업분야에 비해 민간 혁신 역량의 성장 속도가 빠를 것이라고 예측하고 있지만, 장기적 관점에서 바라봤을 때 성공 가능성은 확답할 수 없다고 지적하고 있다. 바이오 기업들이 장기적 경쟁력을 갖기 위해서는 장기적인 연구개발 투자, 임상 파이프라인을 늘리는 등 지속적인 자금 · 투자금이 필요하다. 하지만 우리나라 바이오

기업들은 신약개발 성공률이 2019년 기준으로 지난 10년간 평균의 절반 수준으로 감소하였다. 이는 국내 바이오 기업들이 해외 바이오 기업들에 비해 경쟁력을 갖추기가 어려운 환경에 놓여있다고 판단할 수 있다(홍미영 등, 2021). 이에, 우리나라 신약개발 산업 규모의 성장을 위해서는 바이오 기업의 상장, 주식, 투자와 관련한 민간 투자생태계의 성장이 반드시 필요할 것으로 판단되며, 정부 차원에서 바이오 기업에 대한 민간투자가 확대될 수 있도록 지원 정책을 확대해야 할 것이다(김영국, 2022; 김종란 등, 2022).

5.3 Hot topic: 정부 · 정책

마지막으로 '정부 · 정책'에 관련한 정책적 시사점은 다음과 같다. 디지털 헬스케어 시장은 코로나19 사태의 영향으로 향후 대규모의 시장 형성이 예상되는 바이나, 우리나라는 주요 선진국 대비 장기적 관점의 정책 지원 전략과 규제 개선이 미흡한 상황으로 디지털 전환 등 패러다임 전환에 대한 신속한 대응이 어려운 실정이다. 국내 의료산업의 디지털 전환의 가속화를 위해서 의료 산업의 법 · 제도의 완화를 비롯하여 디지털 전환을 위한 인프라 · 시스템 구축이 우선시되어야 할 것으로 보인다. 주요 선진국들은 디지털 헬스케어 산업을 국가 발전의 핵심 동력으로 강조하며, 일원화된 정책 지원 체제로 바이오 헬스 혁신 기술 · 제품 개발 및 산업 성장을 견인하고 있으며, 글로벌 ICT기업 중심으로 생체 데이터 수집 · 분석에 집중 투자를 하고 있는데 비해, 우리나라는 웨어러블 기기를 통해 수집된 데이터 활용을 위한 산업육성 방안을 논의하는 단계(김유진, 2022; 이경은, 2021)이다. 의료 현장의 디지털 전환을 가속화하기 위해 개인정보 문제, 생체 데이터 수집, 빅데이터 취급, 의료 사고의 책임 소재, 다양한 이해관계자들 간의 법적인 문제에 대한 신속한 개선이 요구된다. 또한 산업 내에 혁신을 이끌 유망 스타트업 및 중소기업을 발굴 · 육성하고, 기술패권 경쟁 시대에 대응하기 위한 글로벌 네트워크 강화 등 글로벌 경쟁력 제고를 위한 다양한 기업 지원 정책을 강화해야 할 것이다.

Ⅵ. 결 론

본 연구에서는 뉴스 빅데이터를 활용하여 '헬스케어'와 관련한 사회적 주요 토픽을 도출하였다. 디지털 헬스케어 산업은 전 세계적으로 고부가가치 시장으로 부각되고 있음에도 불구하고 아직 태동기에 놓여있다는 점을 고려하여 산업 활성화를 위한 정책적 접근으로서, 주요 의제 도출 및 정책적 시사점을 제시하고자 본 연구를 수행하였다. 이를 위하여 2013년 1월 1일부터 2022년 12월 31일까지(10년간) 총 91,873개의 '헬스케어' 관련 기사를 수집하였으며, 토픽모델링(LDA), 다차원척도법(MDS) 및 시계열 회귀분석을 통해 연구를 수행하였다.

토픽모델링 분석 결과, '병원 · 진료', '기업성장', '보험 · 금융', '도시 · 건설', '스마트기술', '정부 · 정책', '기업전망 · 매출', '제품출시', '주식 · 투자', '스마트기기', '바이오제약', '주식시장', '건강관리', '질병 · 치료', '대학 · 연구', '지역조성', '헬스케어제품', '해외진출', '기업경영', '창업 · 스타트업' 순으로 토픽이 구성되었다. 다차원척도법 시각화 분석 결과, 1사분면은 '치료 · 진단 · 헬스케어제품 · 서비스', 2사분면은 '기업 · 투자', 3사분면은 '정부 · 정책', 4사분면은 '지역조성 · 창업'에 관한 토픽들로 분류되었다. 도출된 20개의 토픽을 통해 헬스케어 관련 분야가 단순 질병 진단 · 치료에 국한되지 않고, 스마트기술이 융합된 새로운 의료 비즈니스 모델로 확대되고 있음을 확인할 수 있었다. 시계열 회귀분석 결과, 유의확률 값이 유의미하면서, 회귀계수가 양수(+)로 나타난 Hot topic은 '바이오제약', '건강관리', '기업전망 · 매출', '정부정책' 총 4개로 도출되었으며 회귀계수가 음수(-)인 Cold topic은 '스마트기기', '주식 · 투자', '도시 · 건설' 총 3개, 나머

지 13개 토픽은 중립 토픽으로 도출되었다.

본 연구는 뉴스 빅데이터를 대상으로 디지털 헬스케어 분야의 사회적 현안과 트렌드를 분석하여 정책적 시사점 및 방향성을 제시했다는 점에서 정책적 의의가 있다. 또한 기술 · 시장 · 서비스 동향 분석 등 특정 분야의 연구 중심으로 논의되었던 헬스케어 연구의 영역을 확장하여 거시적이고 종합적인 차원에서 사회적 논의를 분석했다는 점에서 학술적 의의를 가진다.

본 연구의 한계점은 다음과 같다. 언론보도기사는 사회현상 및 이슈를 다루는 객관적인 데이터로 유용하게 활용되나 정서를 다루는 단어의 출현빈도가 적기 때문에 해석에 한계점이 존재한다. 또한 텍스트마이닝 기법은 양적 분석방법론을 기반으로 분석하기 때문에 기사의 심층, 질적 분석이 어렵고, 인과관계에 대한 해석이 불가능하다. 이러한 한계점이 존재함에도 불구하고 본 연구는 헬스케어 분야의 다양한 측면의 이슈를 도출하여, 사회적으로 형성되어 있는 현안에 대해 되짚어보며 관련 정책 개선의 시급성 및 관련 시사점을 도출하였다는 점에서 의의가 있다. 본 연구는 우리나라 정부 기관이 정책을 수립하는데 중요한 기초자료로 활용될 수 있을 것이다.

참 고 문 헌

[1] 고민규, 김태종, “LDA 기반 ESG 이슈 분석: 2009~2022년 뉴스 빅데이터를 중심으로”, *디지털콘텐츠학회논문지*, 제3권, 제24호, 2023, pp. 517-530.

[2] 권기대, “디지털 헬스케어에 대한 국민 인식 조사”, *디지털콘텐트학회논문지*, 제23권, 제3호, 2022, pp. 551-558.

[3] 권승수, “4차 산업혁명시대 헬스케어의 의료정보 활용화 과제”, *문화산업연구*, 제21권, 제2호, 2021, pp. 119-124.

[4] 김영국, “디지털 헬스케어의 나아갈 방향”, *상사법연구*, 제41권, 제3호, 2022, pp. 221-258.

[5] 김유진, “디지털화로 확장되는 헬스케어 생태계”, *하나금융경영연구소 Bi-Weekly Hana Financial Focus*, 제12권, 제13호, 2022.

[6] 김은정, 최희진, “토픽모델링과 네트워크분석을 활용한 헬스케어 분야의 핵심기술과 기술융합 분석 연구: 특허정보를 중심으로”, *한국정보통신학회논문지*, 제26권, 제5호, 2022, pp. 763-778.

[7] 김종란, 강유진, 홍미영, “바이오헬스 정책 · 투자동향”, *KISTEP 브리프*, 제6호, 2022.

[8] 김태종, 이원철, 하소현, 박혜진, 이유리, 강혜진, 안부영, “토픽 모델링 기반 디지털 전환(Digital Transformation) 동향 분석: 1994~2021년 뉴스 빅데이터를 중심으로”, *디지털콘텐츠학회논문지*, 제23권, 제5호, 2022, pp. 929-942.

[9] 나경식, 이지수, “신문 빅데이터를 바탕으로 본 국내 정보화의 경향과 도서관의 역할”, *한국콘텐츠학회논문지*, 제18권, 제9호, 2018, pp. 14-33.

[10] 노희경, “뉴스 빅데이터를 활용한 관광분야 메타버스관련 이슈 분석”, *관광레저연구*, 제34권, 제2호, 2022, pp. 151-166.

[11] 문성호, 바이오부터 디지털헬스케어까지 2700억원 과제 관심 집중”, 메디컬타임즈, 2023.01.26. Available at https://www.medicaltimes.com/Main/News/NewsView.html?ID=1151763.

[12] 박대민, “장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석의 활용 가능성: 100만 건 기사의 정보원과 주제로 본 신문 26년”, 한국언론학보, 제60권, 제5호, 2016, pp. 353-407.

[13] 보건복지부, “바이오헬스 신시장 창출 전략 발표”, 보건복지부 보도자료, 2023.

[14] 안정민, “디지털 헬스케어 산업와 원격의료 산업의 경제적 파급효과 비교분석”, *e-비즈니스연구*, 제22권, 제4호, 2021, pp. 15-25.

[15] 안지연, 이윤정, 이복임, “텍스트 마이닝과 토픽모델링 분석을 활용한 코로나 19와 간호사에 대한 언론기사 분석”, *지역사회간호학회지*,

제32권, 제4호, 2021, pp. 467-476.

[16] 이경은, “국내 디지털 헬스케어의 발전방향”, *AI Trend Watch*, 2021-4호, 2021.

[17] 이택균, “소셜미디어 데이터에 기반한 디지털 헬스케어 연구 동향”, *한국콘텐츠학회논문지*, 제20권, 제3호, 2020, pp. 515-526.

[18] 정일영, 최병삼, 송명진, 김지은, “헬스케어 데이터 공공플랫폼의 활성화를 위한 통합적 전략 연구”, *과학기술정책연구원 정책연구*, 2021-6호, 2021.

[19] 천승현, “제약, 디지털헬스케어에 꽂히다. 새 먹거리 발굴 총력”, 데일리팜, 2022.10.06, Available at http://www.dailypharm.com/Users/News/NewsView.html?ID=292505.

[20] 최은경, 안부영, 김태종, “뉴스 빅데이터 기반 탄소중립 토픽 분석: 2006~2022년 국내 언론보도를 중심으로”, *디지털콘텐츠학회논문지*, 제23호, 제7권, 2022, pp. 1213-1226.

[21] 최한별, 장윤혁, 김성철, “텍스트 마이닝을 활용한 정보 프라이버시 의제 분석: 1990~2021년 뉴스 빅데이터를 중심으로”, *한국정보사회학회지*, 제23권, 2호, 2022, pp. 69-113.

[22] 하소희, 금영정, “네트워크 분석을 이용한 애플리케이션 서비스 하위 카테고리 분류: 헬스케어 어플리케이션 중심으로”, *한국전자거래학회지*, 제25권, 제3호, pp. 15-40, 2020.

[23] 홍미영, 김주원, “바이오헬스 산업 성장가속화를 위한 정부R&D의 역할 및 예산배분 전략,” 한국과학기술기획평가원, 2021-09, 제309호

[24] Borg, I. and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.), New York: Springer-Verlag, 2005, pp. 207-212.

[25] David, M. B., A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation”, *Journal of Machine Learning Research*, Vol.3, 2003, pp. 993-1022.

[26] Frank. R., A. Hinneburg, M. Roder, M. Nettling, and A. Both, “Evaluating topic coherence measures”, *Conference: Neural Information Processing Systems Foundation (NIPS 2013)* - Topic Models Workshop, 2013.

[27] IRS Global, “포스트 코로나 시대 디지털 헬스케어 산업 동향”, IRS Global, 2020.

[28] World Economic Forum, “Top 10 Emerging Technologies of 2021”, World Economic Forum Insight Report, 2021.

Information Systems Review
Volume 25 Number 3
August 2023

Big Data News Analysis in Healthcare Using Topic Modeling and Time Series Regression Analysis

Eun-Jung Kim* · Suk-Gwon Chang** · Sang-Yong Tom Lee***

Abstract

This research aims to identify key initiatives and a policy approach to support the industrialization of the sector. The research collected a total of 91,873 news data points relating to healthcare between 2013 to 2022. A total of 20 topics were derived through topic modeling analysis, and as a result of time series regression analysis, 4 hot topics (Healthcare, Biopharmaceuticals, Corporate outlook · Sales, Government · Policy), 3 cold topics (Smart devices, Stocks · Investment, Urban development · Construction) derived a significant topic. The research findings will serve as an important data source for government institutions that are engaged in the formulation and implementation of Korea's policies.

Keywords: Digital healthcare, Topic Modeling, LDA, Time Series Regression, Data Mining

* Ph.D. Candidate, Business School, Hanyang University
** Corresponding Author, Emeritus Professor, Business School, Hanyang University
*** Professor, Business School, Hanyang University

◐ 저 자 소 개 ◑

김 은 정 (eunjungkim@etri.re.kr)

현재 한양대학교 경영학과 박사과정 중에 있으며, 한국전자통신연구원 기술사업화부서에 재직 중이다. 주요 관심 분야는 기술사업화, 기술경영, 빅데이터 분석, 데이터마이닝 분석 등이다.

장 석 권 (changsg@hanyang.ac.kr)

현재 한양대학교 경영대학 명예교수로 재직 중이다. 주요 관심 분야는 정보통신정책, ICT 정책, IT경쟁전략, 디지털 컨버전스 비즈니스 모델, 디지털 생태계 전략이다. 관련 연구들을 IEEE Transactions on Communications, Telecommunication Systems, Telecommunications Policy, Information Economics and Policy, Operations Research, Decision Support Systems, Journal of Knowledge Management 등 다수의 저널에 논문을 게재하였다.

이 상 용 (tomlee@hanyang.ac.kr)

현재 한양대학교 경영대학 교수로 재직 중이다. 주요 관심 분야는 정보경제, 개인정보보호, 보안, 소셜미디어, 빅데이터애널리틱스 등이다. 관련 연구들을 MIS Quarterly, Management Science, Journal of Management Information Systems 등을 비롯한 다수의 저널에 관련 논문을 게재하였다.

논문접수일 : 2023년 05월 02일 **게재확정일 : 2023년 07월 06일**
1차 수정일 : 2023년 06월 20일