JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 5, pp. 1245-1256, May. 2026

Check for updates

공공데이터 개방의 주요 이슈와 정책 시기별 동향 분석: 뉴스 빅데이터와 BERTopic을 중심으로

정 준 영*
한국지능정보사회진흥원 인공지능정책실 책임연구원

Analysis of Key Issues and Trends in Open Data Policy: Focusing on News Big Data and BERTopic

Junyoung Jeong*
Principal Researcher, Office of AI Policy, National Information Society Agency, Daegu 41068, Korea

[요 약]

본 연구는 2012년부터 2025년까지 국내 언론에 보도된 뉴스 빅데이터에 BERTopic을 적용하여 공공데이터 개방 정책의 시기별 동향을 실증적으로 규명하였다. 분석 결과, 공공데이터 담론은 2012년 [토픽 2] 제도적 기반 형성을 기점으로 태동하였으며, 2020년 코로나19 극복을 위한 [토픽 3, 6] 한국판 뉴딜 및 데이터 댐 구축 이슈가 부상하며 폭발적인 증가세를 보였다. 특히 [토픽 1] 보건의료 데이터는 시기와 무관하게 지속적인 고수요를 보였고, 2022년 이후에는 [토픽 4] 디지털 플랫폼 정부 담론이 급부상하며 민관 협력 생태계로의 패러다임 전환이 확인되었다. 또한 [토픽 5, 7]을 통해 정책이 프롭테크 등 민간 산업 영역으로 확장되고 있음을 규명하였다. 본 연구는 공공데이터 개방 정책이 단순 데이터 개방에서 데이터 경제 및 플랫폼 정부로 진화해온 궤적을 실증적으로 규명하였으며, 시사점으로 AI 학습에 최적화된 질적 고도화(AI-Readiness), 지역 특화 데이터 발굴, 그리고 사회적 가치 창출을 위한 거버넌스 강화를 제언한다.

[Abstract]

This study empirically investigates the dynamic evolution of open data policy discourse in South Korea by applying BERTopic modeling to large-scale news data from 2012 to 2025. The analysis shows that public interest in open data policy began to emerge with the establishment of institutional foundations (Topic 2) around 2012 and expanded significantly in 2020, driven by the *Korean New Deal* and data dam initiatives (Topics 3 & 6) in response to the COVID-19 crisis. Notably, healthcare data (Topic 1) showed consistently high demand throughout the study period, while discourse surrounding the *Digital Platform Government* (Topic 4) increased rapidly after 2022, indicating a paradigm shift toward a public-private cooperative ecosystem. In addition, the expansion of open data into private-sector domains, such as Proptech (Topic 5) and open innovation (Topic 7), was identified. Overall, the findings suggest that open data policy has evolved from a focus on simple disclosure to a core infrastructure supporting the data economy and platform-based governance. Based on these results, this study suggests policy directions focusing on qualitative improvements for AI readiness, the development of regionally specialized data resources, and the establishment of governance frameworks to promote social value creation.

색인어 : 공공데이터, 개방형 데이터, 토픽 모델링, 뉴스 빅데이터, 정책 동향

Keyword : Open Data, BERTopic, Topic Modeling, News Big Data, Policy Trend

http://dx.doi.org/10.9728/dcs.2026.27.5.1245

Received 12 April 2026; **Revised** 27 April 2026
Accepted 30 April 2026

***Corresponding Author; Junyoung Jeong**

Tel: +82-53-230-1228
E-mail: jyjeong@nia.or.kr

Ⅰ. 서 론

OpenAI의 ChatGPT 등장 이후, 인공지능(AI) 기술은 급격한 발전을 거듭하며 사회 전반의 변화를 주도하고 있다. 인공지능은 데이터, 네트워크, 인공지능 기술과 결합하여 4차 산업혁명 시대의 핵심 동력으로 자리 잡았으며, 민간 영역을 넘어 공공 행정 분야에서도 지능형 정부 구현을 위한 필수적인 도구로 인식되고 있다[1]. 특히 복잡다단해지는 사회 문제와 한정된 예산, 인력 등의 제약 조건 속에서 행정 서비스의 효율성을 극대화하고 국민 맞춤형 정책을 수립하기 위해 데이터 기반 행정의 중요성은 그 어느 때보다 강조되고 있다. 이에 정부는 공공데이터의 적극적인 개방과 활용을 통해 공공 가치를 창출하고 데이터 경제를 활성화하기 위한 정책적 노력을 지속하고 있다[2].

이러한 흐름 속에서 공공데이터(Open Data)는 단순한 정보의 공개를 넘어 혁신의 촉매제로서 기능한다. 정부와 공공기관이 보유한 고품질의 데이터는 민간의 창의적인 아이디어와 결합하여 새로운 비즈니스 모델을 창출하고, 시민들의 국정 참여를 유도하여 정부의 투명성을 제고하는 데 기여한다[3]. 국제적으로도 공공데이터의 개방은 경제적 가치 창출뿐만 아니라 사회적 난제 해결을 위한 핵심 자원으로 간주되고 있으며, 데이터의 연결과 활용을 통한 가치 창출 메커니즘에 대한 논의가 활발히 이루어지고 있다[4]. 따라서, 변화하는 공공데이터 이슈의 흐름을 정확히 진단하고, 향후 정책 수립의 방향성을 제시하기 위해서는 관련 동향에 대한 체계적인 분석이 필수적이다.

하지만 기존의 정책 동향 분석 연구들은 주로 정량적 방법에 의존하여, 텍스트 내면에 잠재된 맥락과 세부적인 토픽의 변화를 포착하는 데 한계가 있었다. 이에 본 연구는 BERT(Bidirectional Encoder Representations from Transformers) 임베딩을 활용하여 기존 토픽 모델링 기법(LDA 등)의 한계를 극복하고, 토픽의 의미적 일관성과 해석 가능성을 높인 BERTopic 방법론을 적용하고자 한다[5]. 이를 통해 공공데이터 개방 정책과 관련된 방대한 텍스트 데이터에서 주요 이슈와 시기별 동향을 정밀하게 파악함으로써 보다 객관적이고 실증적인 정책적 시사점을 도출할 수 있을 것이다.

본 연구는 BERTopic 기반의 텍스트 마이닝 기법을 활용하여 국내 공공데이터 개방의 주요 토픽과 시계열적 동향을 분석하는 것을 목적으로 한다. 이를 달성하기 위해 설정된 구체적인 연구문제는 다음과 같다.

첫째, 공공데이터 개방 관련 핵심 토픽은 무엇인가? 선행연구 고찰을 통해 공공데이터 정책과 관련된 주요 키워드를 식별하고, 뉴스 빅데이터에서 수집된 텍스트 데이터를 바탕으로 BERTopic 모델링을 수행하여 주요 의제를 유형화한다.

둘째, 도출된 핵심 토픽의 동향은 어떻게 변화해 왔는가? 시간의 흐름에 따른 토픽의 변화 및 양상을 분석하고, 법률에 따른 공공데이터 기본계획의 정책 발표 시점별로 정책의 초점이 어떻게 이동했는지 규명한다.

셋째, 분석된 정책 동향을 바탕으로 향후 공공데이터 활성화를 위한 정책적 시사점은 무엇인가? 분석 결과를 종합하여 현재 공공데이터 개방 정책의 현주소를 진단하고, 데이터 경제 활성화 및 국민 체감형 서비스 창출을 위한 미래 지향적인 정책 수립 방향을 제언한다.

Ⅱ. 이론적 배경

2-1 공공데이터 개방에 관한 연구

공공데이터(Open Data)에 대한 정의는 국가별 법·제도와 정책적 지향점에 따라 다소 상이하나, 공통적으로 접근성, 재사용성, 그리고 기계 판독 가능성(Machine-readability)을 핵심 요소로 포함한다. 미국 연방정부의 「Open Government Data Act (2019)」는 공공데이터를 연방 기관이 생성, 수집, 보존, 배포하는 디지털 자산 중 보안상 비공개 사유가 없는 한 대중에게 개방형 표준 포맷으로 무료 제공되어야 하는 데이터로 규정하며, 기본 개방(Open by Default) 원칙을 제정하였다. 국내의 경우 「공공데이터의 제공 및 이용 활성화에 관한 법률」 제2조에서 공공기관이 법령이 정하는 목적을 위해 생성 또는 취득하여 관리하는 광(光) 또는 전자적 방식의 자료로 정의하며, 국민의 알 권리 충족과 민간 활용을 통한 삶의 질 향상 및 국민경제 발전을 그 목적으로 명시하고 있다. 이러한 개념은 2009년 미국 오바마 행정부의 'Open Government Initiative'와 'Data.gov' 플랫폼의 출범을 기점으로 투명성(Transparency), 참여(Participation), 협력(Collaboration)이라는 정부 혁신의 가치와 결합하며 전 세계적으로 확산되었다. 공공데이터 개방에 관한 선행연구는 크게 데이터 개방 및 품질 고도화, 민간 활용 생태계 조성, 사회적 가치 창출의 세 가지 차원에서 발전해 왔다.

첫째, 초기 연구들은 공공데이터의 개방 수준과 품질 제고에 초점을 맞추었는데 이는 데이터의 단순한 공개를 넘어 실질적인 재사용성을 확보하기 위해서는 기술적·관리적 표준이 필수적이기 때문이다. 개방형 데이터의 발전 단계를 5-Star 모델로 체계화하여, 단순한 웹 게재(1단계)에서부터 기계 판독이 가능한 구조화된 데이터(3단계), 그리고 데이터 간의 의미적 연결성을 보장하는 연결 데이터(Linked Data, 5단계)로의 진화를 주장한 연구를 확인할 수 있다[6]. 이는 공공데이터가 고립된 정보가 아닌 웹상에서 상호 운용 가능한 자원으로 기능해야 함을 시사한다. 다른 연구에서는 공공데이터 개방을 저해하는 요인을 분석하고, 이를 극복하기 위한 데이터 생명주기 관리의 중요성을 강조하였다[7].

둘째, 민간의 활용성 증대와 데이터 생태계 구축에 관한 연구들은 데이터 제공자와 사용자 간의 상호작용에 주목하였는데, 공공데이터가 경제적 가치를 창출하기 위해서는 정부 주도의 일방적 개방을 넘어 민간과 공공이 협력하는 생태계적 접근이 필요함을 제시하였다[8]. 이는 공공데이터가 스타트

업이나 기업의 비즈니스 모델과 결합될 때 비로소 혁신적인 서비스로 재탄생할 수 있음을 의미한다. 또한, 데이터의 품질 속성이 실제 사용자의 활용 의도에 미치는 영향을 실증적으로 분석하였다[9]. 해당 연구에서는 정확성(Accuracy), 완전성(Completeness), 시의성(Timeliness) 외에도 사용자가 데이터를 쉽게 찾고 해석할 수 있는 접근성(Accessibility), 이해가능성(Understandability)이 민간 활용의 핵심 결정 요인임을 규명하고 이를 관리하기 위한 지표 개발의 필요성을 제언하였다.

셋째, 공공데이터 개방이 가져오는 사회적 파급효과(Outcome)와 공공 가치(Public Value) 측면으로 확장되고 있다. 공공데이터 활용이 정치적 투명성 제고, 사회적 책임 강화, 경제적 성장이라는 다차원적인 이익을 제공한다고 분석하였다[10]. 특히, 공공데이터 개방을 단순한 기술적 과제가 아닌 민주주의 심화의 기제로 해석하였다[11]. 그들은 '열린 데이터'가 시민들이 사회 문제 해결 과정에 능동적으로 참여하게 하는 열린 혁신(Open Innovation)의 도구로 작동하며, 결과적으로 정부 신뢰 회복과 민주적 가치 실현에 기여함을 강조하였다.

종합하면, 공공데이터 관련 논의는 데이터 자체의 품질을 확보하는 기술적 접근에서 시작하여 경제적 활용을 위한 생태계적 접근, 그리고 최종적으로 사회 전반의 공익과 민주적 가치를 실현하는 거시적 관점으로 변화하고 있음을 알 수 있다.

2-2 Topic Modeling에 관한 연구

토픽 모델링(Topic Modeling)은 방대한 비정형 텍스트 데이터에서 의미 있는 주제를 자동으로 추출하고 문서를 분류하는 텍스트 마이닝 기법이다. 전통적으로 가장 널리 활용되어 온 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 기법은 문서를 단어들의 집합인 BoW(Bag of Words) 형태로 간주하고, 각 문서를 잠재된 주제들의 확률적 혼합물로 모델링한다. LDA는 결과의 해석이 용이하다는 장점이 있으나, 단어의 등장 순서와 문맥적 정보를 무시함으로써 단어 간의 의미적 관계를 온전히 포착하지 못한다는 본질적인 한계가 존재한다. 또한, 불용어 처리나 형태소 분석 등 복잡한 전처리에 따라 결과 품질이 크게 좌우되며, 동음이의어(Polysemy)와 같이 문맥에 따라 의미가 달라지는 단어를 구분하는 데 취약점을 보인다[12].

이러한 한계를 극복하기 위해 딥러닝 기반의 언어 모델, 특히, BERT(Bidirectional Encoder Representations from Transformers)와 같은 트랜스포머(Transformer) 아키텍처를 활용한 연구가 급부상하였다. BERT는 양방향 주의(Attention) 메커니즘을 통해 문장의 문맥을 파악하고 단어 및 문장 수준의 고차원 벡터 표현(Embedding)을 생성하는 데 탁월한 성능을 보인다. 이를 토픽 모델링에 접목한 Top2Vec은 Doc2Vec 등을 활용하여 문서와 단어를 동일한 벡터 공간에 임베딩하고, 밀도 기반 클러스터링을 수행하여 토픽을 도출하였다. 그러나 Top2Vec은 각 클러스터의 중심(Centroid)을 곧 해당 토픽으로 가정하는 경향이 있어, 실제 데이터가 구형(Spherical)이 아니거나 노이즈가 많은 경우 토픽 표현의 정확도가 저하될 수 있다는 문제점이 제기되었다. 이에 기존의 임베딩 기반 클러스터링 접근법에 클래스 기반 TF-IDF 절차를 결합한 BERTopic 방법론을 제안하였다[5]. BERTopic은 사전 훈련된 언어 모델을 활용하여 문서의 문맥적 의미를 벡터로 변환한 후, 이를 클러스터링하여 토픽을 형성하고, 최종적으로 각 클러스터를 대변하는 키워드를 추출하는 모듈형 구조를 취한다. 이는 LDA와 달리 문서 전체의 일반적인 주제를 파악하기 위해 과도한 전처리를 요구하지 않으며, 의미적으로 유사한 문맥을 가진 단어들을 효과적으로 군집화함으로써 보다 일관성 있고 해석 가능한 토픽을 도출한다는 장점이 있다. BERTopic의 구체적인 알고리즘 프로세스는 다음의 3단계로 체계화할 수 있으며, 각 단계의 알고리즘을 연구 목적과 데이터 특성에 맞게 유연하게 변경할 수 있으나, 기본적으로 SBERT, UMAP, HDBSCAN, c-TF-IDF의 조합을 따른다. 첫째, 문서 임베딩(Document Embedding) 단계이다. SBERT(Sentence-BERT) 등을 활용하여 개별 문서를 고차원의 벡터 공간으로 매핑한다. 이는 문장 단위의 의미적 유사성을 보존하는 임베딩을 생성하며, 단순한 단어의 빈도가 아닌 문맥적 정보를 반영한다[13]. 둘째, 차원 축소 및 클러스터링(Dimensionality Reduction & Clustering) 단계이다. 생성된 고차원 벡터는 희소성(Sparsity) 문제를 해결하기 위해 차원 축소 과정을 거친다. BERTopic은 UMAP(Uniform Manifold Approximation and Projection)을 사용하여 데이터의 국소적 구조와 전역적 구조를 동시에 보존하면서 차원을 축소한다[14]. 이후, 밀도 기반 클러스터링 알고리즘인 HDBSCAN (Hierarchical Density Based Spatial Clustering of Applications with Noise)을 적용하여 유사한 벡터들을 군집화한다. HDBSCAN은 K-Means와 달리 클러스터의 개수를 사전에 지정할 필요가 없으며, 어느 클러스터에도 속하지 않는 노이즈(Outlier)를 효과적으로 분리해냄으로써 토픽의 순도(Purity)를 높인다[15]. 셋째, 토픽 표현(Topic Representation) 단계이다. 도출된 각 클러스터가 어떤 주제를 의미하는지 설명하기 위해 c-TF-IDF를 적용한다. 기존의 TF-IDF가 개별 문서를 기준으로 단어의 중요도를 산출했다면, c-TF-IDF는 하나의 클러스터(토픽)에 속한 모든 문서를 하나의 거대한 문서로 간주하여 단어 빈도를 계산한다. 이 과정을 통해 각 클러스터 내에서는 빈번하게 등장하지만 다른 클러스터에서는 드물게 등장하는 단어, 즉 해당 토픽을 가장 잘 대표하는 핵심어(Keyword)를 추출할 수 있다[5]. 이러한 BERTopic의 특성은 정책 문서와 같이 전문 용어와 맥락적 의미가 중요한 데이터 분석에 있어 기존 방법론 대비 우수한 성능을 발휘하는 것으로 평가받고 있다.

Ⅲ. 연구 방법

3-1 연구 데이터

본 연구는 공공데이터 개방 정책과 인공지능 기술의 융합 동향을 실증적으로 파악하기 위해, 한국언론진흥재단이 운영하는 뉴스 빅데이터 분석 시스템인 빅카인즈(BigKinds)를 데이터를 원천 데이터로 활용하였다. 분석 대상 데이터의 수집 기간은 2012년 1월 1일부터 2025년 2월 12일까지로 설정하여, 공공데이터 개념의 도입 초기부터 현재의 생성형 AI 시대에 이르기까지의 장기적인 정책 및 기술 트렌드 변화를 포괄하고자 하였다.

데이터 수집을 위한 검색식은 '공공데이터'를 핵심 주제어로 설정하고, 이와 연계된 이슈를 포착하기 위해 구체적인 기술 용어를 결합하는 방식을 적용하였다. 구체적으로는 '공공데이터'와 '인공지능(AI)', '머신러닝(ML)', '딥러닝(DL)', '거대언어모델(LLM)', '빅데이터' 등의 키워드를 AND 조건으로 교차 검색하여, 공공데이터 개방과 지능형 기술이 결합된 뉴스 기사를 선별하였다. 상세 검색식은 ('공공데이터' OR '공공 데이터')를 기본으로 하고, 여기에 ('공공데이터 개방' OR '공공 데이터 개방') 등을 조합하는 불리언(Boolean) 로직을 적용하여 검색의 정확도와 재현율을 높였다.

본 연구의 데이터 전처리 단계에서는 한국어 텍스트의 정교한 형태소 분석을 위해 Kiwi(Korean Intelligent Word Identifier) 형태소 분석기를 활용하였다. 수집된 뉴스 데이터로부터 정책 담론의 핵심 의미를 보존하기 위해 불용어를 제거한 후 명사 키워드를 추출하여 분석에 사용하였다.

수집된 원시 데이터(Raw Data)는 뉴스 식별자, 일자, 언론사, 기고자, 제목, 통합 분류, 사건 및 사고 분류, 인물, 위치, 기관 등의 개체명(Named Entity), 키워드, 특성추출(가중치 상위 50개), 본문, URL 등 다양한 메타데이터 컬럼으로 구성되어 있다. 이 중 본 연구의 BERTopic 토픽 모델링을 위해서는 '기관', '키워드', '특성추출(가중치 상위 50개)' 컬럼을 주요 분석 변수로 활용하였다. 이와 같은 수집 및 정제 과정을 거쳐, 최종적으로 공공데이터 개방 관련 이슈를 다룬 총 5,928건의 뉴스 데이터가 분석 대상으로 확정되었다. 이 데이터셋은 시계열적 흐름에 따른 공공데이터 개방 정책의 변화 과정과 인공지능 기술 도입의 양상을 대변하는 표본으로서 본 연구의 토픽 모델링 분석에 활용되었다.

3-2 연구 절차

본 연구는 데이터의 수집부터 최종적인 토픽 분석에 이르기까지 총 3단계의 체계적인 절차를 통해 수행된다. 그림 1은 본 연구의 전체적인 분석 프레임워크를 도식화한 것이다. 연구의 초기 단계는 분석 대상이 되는 기초 데이터를 확보하는 데이터 수집(Data Collection) 과정이다. 본 연구는 한국언론진흥재단의 뉴스 빅데이터 분석 시스템인 빅카인즈(BigKinds)를 활용하여, 2012년부터 2025년까지 보도된 공공데이터 관련 뉴스 기사를 수집하였다. 확보된 데이터는 이어서 토픽 모델링(Topic Modeling) 단계로 넘어가기 전에 우선 뉴스 데이터의 품질을 높이고 분석의 효율성을 확보하기 위해 기본적인 전처리 작업이 선행된다. 특히 빅카인즈 데이터의 특성을 고려하여 키워드를 중심으로 정제 과정을 거친다. 이후 딥러닝 기반의 토픽 모델링 기법인 BERTopic 알고리즘을 적용하여 텍스트 데이터 내에 잠재된 의미 구조를 파악한다. 이는 문장의 문맥을 깊이 있게 이해하는 BERT의 임베딩 기술과 밀도 기반의 클러스터링 기법을 결합함으로써, 기존 방법론 대비 더욱 일관성 있고 해석 가능한 주제 군집을 추출하는 것을 가능하게 한다.

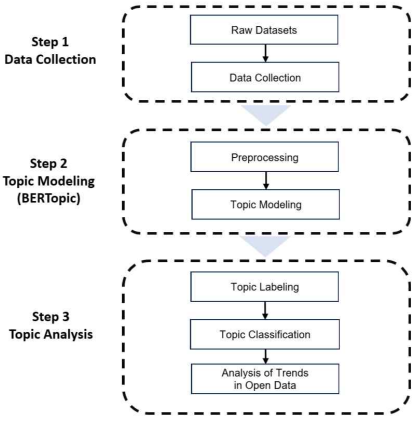


그림 1. 본 연구의 흐름도
Fig. 1. Flowchart of the research process

최종 단계인 토픽 분석(Topic Analysis)에서는 도출된 토픽에 대한 심층적인 해석이 이루어진다. 먼저 각 토픽을 대표하는 핵심 키워드를 바탕으로 해당 토픽의 의미를 가장 잘 드러낼 수 있는 명칭을 부여하는 토픽 명명(Topic Labeling) 작업이 수행된다. 이어 유사한 성격의 토픽들을 통합하거나 상위 범주로 묶는 토픽 분류(Topic Classification) 과정을 통해 주요 정책 의제들을 유형화한다. 마지막으로, 분류된 토픽들의 시기별 출현 빈도와 비중 변화를 추적하는 시계열 분석을 통해 공공데이터 개방 정책의 주요 흐름과 트렌드(Analysis of Trends)를 규명하고, 이를 바탕으로 시사점을 도출한다.

3-3 BERTopic 방법론

본 연구는 공공데이터 개방 관련 텍스트 데이터에 내재된 잠재적 주제와 맥락을 심층적으로 규명하기 위해, 최신 토픽 모델링 기법인 BERTopic 알고리즘을 분석 모형으로 채택하였다. 그림 2에 도식화된 절차를 따르는 BERTopic은 사전학습된 트랜스포머(Transformer) 기반의 언어 모델을 활용하여 문서의 문맥적 의미를 파악하는 기법이다[5]. 이는 단어의 단순 동시 등장 빈도(Co-occurrence)에 의존하는 전통적

인 잠재 디리클레 할당(LDA) 기법이 문맥 정보를 충분히 반영하지 못하는 한계를 효과적으로 보완하며, 텍스트의 의미적 구조를 보다 정밀하게 포착할 수 있다는 장점을 지닌다[12].

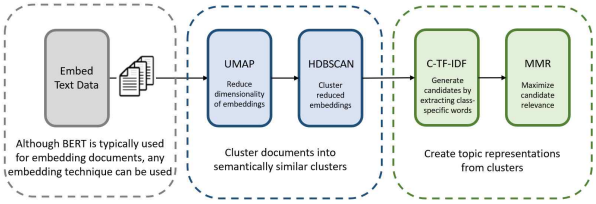


그림 2. BERTopic 방법론의 개념도
Fig. 2. Conceptual diagram of the BERTopic methodology

분석 프로세스는 문서 임베딩, 차원 축소, 클러스터링, 그리고 토픽 추출의 단계가 유기적으로 연결되어 수행된다. 우선 입력된 텍스트 데이터는 SBERT(Sentence-BERT) 프레임워크를 통해 고차원의 벡터 공간으로 변환된다. 이 과정에서 각 문서는 의미론적 정보가 보존된 밀집 벡터(Dense Vector)로 임베딩되어, 단어의 표면적 의미를 넘어선 심층적인 의미 정보를 함축하게 된다[13]. 다만, 본 연구에서 활용된 데이터는 뉴스 빅데이터의 특성상 기사 전문이 아닌 핵심 키워드 컬럼을 바탕으로 구성되었다. 문장 단위의 완전한 문맥 정보(Full Context)를 직접적으로 활용하는 데에는 제약이 있으나, BERTopic 기반 임베딩은 나열된 키워드 집합 내에서도 단어 간의 고차원적인 의미론적 관계(Semantic Relationship)를 벡터 공간상에 투영할 수 있는 장점이 있다. 따라서, 본 연구는 문장 맥락을 직접 파악하는 방식보다는 정책 관련 핵심 어휘들 사이의 잠재적 연관성을 세밀하게 분석한 어휘적 문맥 임베딩에 가깝다고 정의할 수 있다.

토픽 모델링을 위해 활용된 BERTopic 방법론의 임베딩 단계에서는 다국어 문맥 이해 능력이 검증된 'distiluse-base-multilingual-cased-v1' Sentence Transformer 모델을 채택하였다. 이를 통해 한 차례 가공된 키워드 집합 간의 잠재적 의미 관계를 고차원 벡터 공간에 투영하였다.

다음으로 생성된 고차원 벡터는 희소성(Sparsity) 문제를 완화하고 연산의 효율성을 확보하기 위해 UMAP(Uniform Manifold Approximation and Projection) 알고리즘을 거치게 된다. UMAP은 데이터의 전역적 구조와 국소적 구조를 동시에 보존하면서 고차원 임베딩을 저차원으로 효과적으로 압축하여, 후속 단계인 클러스터링의 성능을 극대화하는 역할을 수행한다[14].

차원이 축소된 벡터 공간에서는 HDBSCAN (Hierarchical Density Based Spatial Clustering of Applications with Noise)을 적용하여 의미적으로 유사한 문서들을 군집화한다. HDBSCAN은 밀도 기반 클러스터링 알고리즘으로 사전에 토픽(클러스터)의 개수를 임의로 지정해야 하는 K-Means 등의 기법과 달리 데이터의 밀도 분포에 따라 자동으로 최적의 클러스터 수를 결정한다. 또한, 특정 주제로 명확히 분류하기 어려운 데이터를 노이즈(Noise)로 처리하여 배제함으로써 도출되는 토픽의 순도(Purity)와 해석 가능성을 제고한다[15].

최종적으로 도출된 각 클러스터의 고유한 성격을 규명하고 주제를 명명(Labeling)하기 위해 c-TF-IDF(class based Term Frequency Inverse Document Frequency) 기법이 적용된다. 이는 전체 말뭉치가 아닌 특정 클러스터 내에서 상대적으로 빈번하게 등장하는 단어에 높은 가중치를 부여하는 방식이다. 이를 통해 각 클러스터를 대표하는 핵심 키워드 집합을 추출함으로써, 해당 토픽이 내포하고 있는 중심 주제를 구체적으로 설명하는 것이 가능하다[5].

Ⅳ. 연구 결과

4-1 토픽 모델링(Topic Modeling) 분석

본 연구에서 도출된 토픽 모델링의 품질을 검증하기 위해 세 가지 평가지표를 확인한 결과, Coherence Score는 0.408, Perplexity는 1.945, 그리고 군집의 응집도와 분리도를 나타내는 Silhouette Score는 0.671로 나타났다. 이는 도출된 토픽들이 의미론적으로 일관성이 있으며, 각 클러스터가 통계적으로 유의미하게 구분되고 있음을 입증한다.

BERTopic 모델링의 고도화 과정으로 문서 재할당

표 1. 토픽 Label 및 핵심 키워드
Table 1. Topic names and key keywords

토픽	토픽 Label	핵심 키워드
1	보건의료 데이터 개방 및 활용	의료, 보험, 병원, 건강, 의료 데이터, 보건, 헬스, 케어, 환자, 심사
2	공공데이터 개방 정책 및 제도적 기반	개방, 공공, 공공 데이터, 데이터, 정보, 기관, 활용, 가명, 데이터 개방
3	한국판 뉴딜 기반 데이터 일자리 창출	뉴딜, 한국판, 한국판 뉴딜, 고용, 안전망, 일자리, 투자, 경제, 에너지, 코로나
4	디지털 플랫폼 정부 구현 및 거버넌스	디지털 플랫폼, 플랫폼 정부, 디지털 플랫폼 정부, 정부, 디지털, 인수위, 위원회, 국민
5	부동산 데이터 기반 창업 지원 및 프롭테크 육성	부동산, 부동산 서비스, 창업, 서비스 산업, 대회, 창업 경진, 경진 대회, 아파트
6	디지털 뉴딜과 데이터 인프라 구축	뉴딜, 디지털, 디지털 뉴딜, 구축, 정부, 사업, 인프라, 스마트, 투자, 일자리
7	공공데이터 창업 지원 거점 및 오픈 이노베이션	창업, 오픈 스퀘어, 혁신 센터, 창조 경제 혁신, 경제 혁신, 스퀘어, 창조, 오픈
8	4차 산업혁명 대응 및 규제 혁신	혁명, 산업 혁명, 산업, 혁신, 규제, 스마트, 과학, 시대, 사회, 기술
9	스마트시티 데이터 플랫폼 구축 및 도시 문제 해결	스마트시티, 도시, 스마트, 스마트 도시, 교통, 구축, 데이터, 지능, 시티
10	지역 거점 인공지능 클러스터 조성 및 산업 육성	광주, 인공, 인공 지능, 지능, 광주시, 경기도, 협약, 인공 지능 산업, 집적단지

*Written in Korean to ensure a consistent context and coherent explanation between the main text and the results of the topic modeling analysis.

(Reallocation)을 수행하여 데이터의 의미적 구조를 명확히 한 결과, 공공데이터 관련 뉴스 담론은 특정 이슈에 매몰되지 않고 정책, 경제, 지역, 품질 등 다차원적으로 분화되어 진화하고 있음을 확인할 수 있었다. 상위 10개 토픽에 대한 토픽 Label 및 핵심 키워드는 표 1과 같이 종합할 수 있다.

토픽 1은 의료, 보험, 병원, 건강, 의료 데이터, 보건, 헬스, 케어, 환자 등과 같은 보건의료 및 건강보험 관련 키워드로 구성된다. 일반적으로 보건의료 공공데이터는 생성 및 관리 주체에 따라 국민건강보험공단의 자격 및 검진 정보, 건강보험심사평가원의 진료 내역 및 의약품 처방 정보 등으로 구분되며, 최근에는 민간 헬스케어 서비스와의 융합을 위해 데이터의 가명 처리 및 결합 방식, 활용 가이드라인 등 다방면으로 연구 성과가 축적되고 있다[16]. 특히, 2020년 코로나19 팬데믹과 데이터 3법 개정을 기점으로 보건의료 빅데이터는 감염병 예측 및 개인 맞춤형 건강 관리의 핵심 자원으로서 새롭게 주목받아 왔으며, 이에 따른 데이터의 안전한 활용과 개인정보 보호, 그리고 마이데이터 사업과의 연계 방안 관련 연구가 활발히 진행되었다[17],[18]. 한편, 연구의 주요 키워드로 헬스, 케어, 환자 등이 나타나는데 이는 단순한 정보 조회를 넘어 실질적인 의료 서비스 제공과 환자 관리에 데이터가 적극적으로 활용되고 있음을 시사한다. 이러한 맥락을 종합하여 토픽명을 보건의료 데이터 개방 및 활용으로 설정하였다.

토픽 2는 개방, 공공, 공공 데이터, 데이터, 정보, 기관, 활용, 가명 등 공공데이터 정책의 핵심 기조와 실행 전략을 포괄하는 키워드로 구성된다. 이는 정부와 공공기관이 보유한 원천 데이터를 민간에 적극적으로 제공하여 데이터 경제를 활성화하려는 국가적 차원의 거버넌스를 의미한다. 초기 공공데이터 정책이 단순한 알 권리 충족과 투명성 제고를 위한 양적 개방에 집중했다면, 최근에는 민간의 실질적 수요에 부응하기 위한 고가치 데이터 발굴과 품질 제고로 패러다임이 전환되고 있다[19]. 특히 주요 키워드로 도출된 가명은 2020년 데이터 3법(개인정보 보호법·정보통신망법·신용정보법) 개정 이후, 개인정보 침해 우려를 해소하면서도 데이터의 결합과 활용 가치를 극대화하기 위한 핵심 기제로 부상하였다. 이에 따라, 가명정보 결합 전문기관의 지정 및 안전한 활용 모델 구축에 관한 논의가 학계와 현업에서 활발히 이루어지고 있다[20]. 또한, 기관과 활용의 연계는 공공기관이 보유한 데이터가 단순히 개방되는 것에 그치지 않고, 민간의 창의적인 비즈니스 모델과 결합하여 실질적인 경제적 부가가치를 창출해야 한다는 정책적 지향점을 시사한다. 이러한 맥락을 종합하여 토픽명을 공공데이터 개방 정책 및 제도적 기반으로 설정하였다.

토픽 3은 뉴딜, 한국판, 한국판 뉴딜, 고용, 안전망, 일자리, 투자, 경제, 에너지, 코로나 등 2020년 정부가 발표한 국가발전 전략인 한국판 뉴딜과 관련된 핵심 키워드로 구성된다. 이는 코로나19 팬데믹으로 인한 경제 위기를 극복하고 추격형 경제에서 선도형 경제로 도약하기 위해 추진된 대규모 국가 프로젝트로 특히 디지털 뉴딜의 핵심 과제인 데이터 댐 구축 사업과 밀접하게 연관되어 있다. 정부는 공공데이터의 전면적인 개방과 가공을 통해 인공지능 산업의 기초를 마련하는 동시에, 이 과정에서 데이터 라벨링 등 대규모 일자리를 창출하여 고용 충격을 완화하고자 하였다[21]. 키워드에 포함된 안전망과 고용은 이러한 데이터 정책이 단순한 기술 도입을 넘어 고용 위기 극복을 위한 사회적 안전망 강화 수단으로 활용되었음을 시사한다. 이러한 정책적 맥락과 목표를 반영하여 토픽명을 한국판 뉴딜 기반 데이터 일자리 창출로 설정하였다.

토픽 4는 디지털 플랫폼, 플랫폼 정부, 디지털 플랫폼 정부, 정부, 디지털, 인수위, 위원회, 국민 등 이전 정부의 핵심 국정 과제인 디지털 플랫폼 정부의 출범 및 추진 체계와 관련된 키워드로 구성된다. 이는 기존의 전자정부가 행정 서비스의 전산화와 효율화에 집중했던 것에서 한 단계 나아가, 모든 데이터가 연결되는 디지털 플랫폼 위에서 국민, 기업, 정부가 함께 사회문제를 해결하고 새로운 가치를 창출하는 플랫폼으로서의 정부를 지향하는 패러다임의 전환을 의미한다[22]. 특히, 인수위, 위원회와 같은 키워드는 새 정부 출범 초기 대통령직인수위원회를 중심으로 국정 과제가 설계되고, 이를 전담할 민관 합동 추진 기구가 신설되는 일련의 정책 형성 과정을 반영한다. 이는 정부 주도의 일방적인 공공서비스 제공 방식에서 벗어나, 민간의 기술과 창의성을 행정에 접목하는 민관 협력 기반의 혁신 생태계를 조성하려는 거버넌스의 근본적인 변화를 시사한다[23]. 이러한 정책적 기조와 새로운 추진 체계의 특성을 종합하여 토픽명을 디지털 플랫폼 정부 구현 및 거버넌스로 설정하였다.

토픽 5는 부동산, 부동산 서비스, 창업, 부동산 서비스 산업, 서비스 산업, 대회, 창업 경진, 창업 경진 대회 등 부동산 분야의 데이터 활용과 창업 지원 활동에 관련된 핵심 키워드로 구성된다. 이는 국토교통부와 한국부동산원 등 공공기관이 보유한 부동산 실거래가, 지적도, 건축물대장 등의 고가치 공공데이터가 프롭테크(PropTech) 산업의 핵심 자원으로 활용되고 있음을 나타낸다. 특히, '창업 경진 대회'와 '대회' 같은 키워드의 빈번한 등장은 정부가 단순한 데이터 개방을 넘어, 범정부 공공데이터 활용 창업경진대회나 부동산 서비스산업 창업경진대회 등 다양한 공모전을 개최하여 민간의 창의적인 아이디어를 발굴하고 사업화를 적극적으로 지원하고 있음을 시사한다[24]. 이러한 정책은 공공데이터가 실질적인 비즈니스 모델로 연결되도록 돕는 인큐베이팅 역할을 수행하며, 전통적인 부동산 산업을 정보기술이 결합된 고부가가치 서비스 산업으로 전환하는 데 기여하고 있다[25]. 이러한 맥락을 종합하여 토픽명을 부동산 데이터 기반 창업 지원 및 프롭테크 육성으로 설정하였다.

토픽 6은 뉴딜, 디지털, 디지털 뉴딜, 구축, 정부, 사업, 인프라, 스마트, 투자, 일자리 등 한국판 뉴딜의 한 축인 디지털 뉴딜과 그 핵심 과제인 인프라 확충에 관련된 키워드로 구성된다. 이는 정부가 4차 산업혁명의 선도적 지위를 확보하기

위해 데이터, 네트워크, 인공지능 생태계를 강화하고자 추진한 대규모 국가 디지털 전환 프로젝트를 의미한다. 특히 인프라와 구축, 투자라는 키워드는 디지털 뉴딜의 대표 사업인 데이터 댐 프로젝트를 상징하는데, 이는 공공과 민간의 네트워크를 통해 생성된 데이터가 모이고 가공되어 스마트 의료, 스마트 제조 등 다양한 분야로 흐르도록 하는 기반 인프라를 만드는 것을 목표로 한다[26]. 정부는 이러한 디지털 인프라 구축 사업에 막대한 재정을 투입하여 경기 부양을 도모하고, 비대면 인프라와 SOC 디지털화를 통해 지속 가능한 일자리 창출과 산업 혁신을 견인하고자 하였다[21]. 이러한 정책적 목표와 사업의 성격을 종합하여 토픽명을 디지털 뉴딜과 데이터 인프라 구축으로 설정하였다.

토픽 7은 창업, 오픈 스퀘어, 혁신 센터, 창조 경제 혁신, 경제 혁신, 스퀘어, 창조 경제, 창조, 오픈 등 공공데이터 기반의 창업 생태계 조성을 위한 물리적 거점과 지원 정책에 관련된 키워드로 구성된다. 이는 과거 창조경제 기조 하에 설립된 창조경제혁신센터와 공공데이터 활용 원스톱 지원 센터인 오픈스퀘어-D가 데이터 기반 스타트업의 인큐베이팅 허브로 기능해 왔음을 보여준다. 정부는 이러한 혁신 거점을 통해 예비 창업자에게 공공데이터 활용 교육, 컨설팅, 입주 공간을 제공함으로써 아이디어의 사업화를 지원해 왔다[27]. 특히 '오픈 스퀘어'와 '혁신 센터' 키워드의 결합은 단순한 자금 지원을 넘어, 창업가들이 모여 정보를 교류하고 협업할 수 있는 개방형 혁신(Open Innovation) 공간의 중요성이 정책적으로 강조되었음을 시사한다. 이러한 정책적 수단과 공간적 특성을 종합하여 토픽명을 공공데이터 창업 지원 거점 및 오픈 이노베이션으로 설정하였다.

토픽 8은 혁명, 산업 혁명, 산업, 혁신, 규제, 스마트, 과학, 시대, 사회, 기술 등 4차 산업혁명의 도래와 이에 대응하기 위한 기술 및 규제 혁신에 관련된 키워드로 구성된다. 이는 인공지능, 빅데이터 등 지능정보기술이 촉발한 사회 전반의 구조적 변화인 4차 산업혁명이 공공데이터 정책의 거시적 배경으로 작용했음을 의미한다. 규제 키워드의 등장은 신기술의 산업 적용을 가로막는 기존 법·제도를 개선하기 위해 도입된 규제 샌드박스(Regulatory Sandbox)나 네거티브 규제 방식의 도입 논의를 반영한다. 정부는 데이터 경제 활성화를 위해 개인정보 규제를 합리화하고, 신산업 분야의 진입 장벽을 낮추는 등 제도적 기반을 정비하는 데 주력하였다[28]. 또한, 과학과 기술, 사회의 연결은 공공데이터와 스마트 기술이 단순한 산업 육성을 넘어 사회 문제 해결을 위한 과학기술적 수단으로 활용되어야 한다는 시대적 요구를 대변한다. 이러한 거시적 담론과 정책 방향을 종합하여 토픽명을 4차 산업혁명 대응 및 규제 혁신으로 설정하였다.

토픽 9는 스마트시티, 도시, 스마트, 스마트 도시, 교통, 구축, 데이터, 스마트시티 데이터, 지능, 시티 등 도시 문제 해결을 위한 데이터 기반의 스마트시티 구축에 관련된 키워드로 구성된다. 이는 교통 혼잡, 환경 오염 등 복잡한 도시 문제를 해결하기 위해 도시 곳곳에서 수집된 공공데이터와 IoT 센서 데이터를 융합하여 지능형 도시 관리 시스템을 구축하려는 정책적 노력을 나타낸다. 특히, 교통은 스마트시티 데이터가 가장 활발하게 적용되는 분야로 실시간 교통량 분석이나 대중교통 최적화 등에 공공데이터가 핵심적으로 활용되고 있음을 보여준다. 정부는 세종 5-1 생활권, 부산 에코델타시티 등 국가 시범도시를 지정하고, 데이터 허브 플랫폼을 구축하여 도시 데이터의 수집, 분석, 활용 체계를 표준화하려는 정책적 노력을 기울이고 있다[24]. 이는 공공데이터가 가상 공간과 물리적 공간을 연결하는 스마트시티의 혈액 역할을 수행하고 있음을 시사한다. 이러한 적용 분야와 정책 목표를 종합하여 토픽명을 스마트시티 데이터 플랫폼 구축 및 도시 문제 해결로 설정하였다.

토픽 10은 광주, 인공, 인공 지능, 지능, 광주시, 경기도, 협약, 인공 지능 산업, 집적단지 등 지역 거점 중심의 인공지능 산업 육성에 관련된 키워드로 구성된다. 특히, 광주는 정부로부터 인공지능 중심 산업융합 집적단지로 지정되어, 국가 AI 데이터 센터 구축과 AI 특화 기업 유치에 행정력을 집중하고 있음을 보여준다[29]. 키워드 중 협약은 지방자치단체와 AI 기업 간의 업무 협약(MOU) 체결 등이 보도되었음을 나타내며, 이는 기업 유치를 통해 지역 내 데이터 산업 생태계를 조성하려는 노력을 방증한다. 또한, 경기도 등의 키워드가 함께 등장한 것은 판교 테크노밸리와 같은 기존 ICT 거점과 광주 AI 클러스터 등 지역별 특화 발전 전략이 공공데이터 및 AI 정책의 중요한 축을 형성하고 있음을 의미한다. 이는 중앙 주도의 데이터 정책이 지역 균형 발전 정책과 결합하여 지역 주도의 혁신 성장 모델로 확산되고 있음을 시사한다. 이러한 지역적 특성과 산업적 목표를 종합하여 토픽명을 지역 거점 인공지능 클러스터 조성 및 산업 육성으로 설정하였다.

4-2 토픽의 계층적 구조 및 분포 분석

BERTopic 모델링을 통해 도출된 상위 10개 토픽 간의 의미적 유사성과 위계 구조를 규명하기 위해 계층적 군집 분석(Hierarchical Clustering)과 토픽 분포 분석을 수행하였다. 그림 3의 덴드로그램(Dendrogram)은 토픽들이 형성하는 군집의 양상을 시각적으로 보여주는데, 결합 거리에 따라 크게 두 개의 거시적 군집으로 구분되며, 내부적으로는 다시 세부적인 의미 연결망을 형성하고 있음을 확인할 수 있다. 우선 덴드로그램 상단에 위치한 첫 번째 거시적 군집은 국가 정책 거버넌스 및 산업적 활용 영역을 포괄한다.

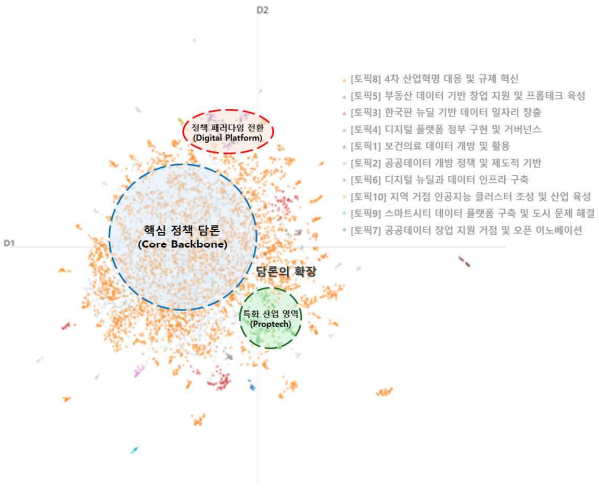
[토픽 4] 디지털 플랫폼 정부 구현 및 거버넌스와 [토픽 1] 보건의료 데이터 개방 및 활용이 가장 밀접하게 결합하며 하나의 하위 그룹을 형성한다. 이는 이전 정부의 핵심 기조인 디지털 플랫폼 정부가 추구하는 민관 협력과 데이터 연계의 가장 대표적인 선도 분야가 보건의료 데이터임을 시사한다. 즉, 디지털 플랫폼이라는 행정적 틀 안에서 헬스케어 데이터가 핵심 킬러 서비스로 기능하고 있음을 시사한다.

[토픽 3] 한국판 뉴딜 기반 데이터 일자리 창출과 [토픽 5]

*Written in Korean to ensure a consistent context and coherent explanation between the main text and the results of the topic modeling analysis.

그림 3. 토픽 간 계층적 구조

Fig. 3. Hierarchical structure of topics



*Written in Korean to ensure a consistent context and coherent explanation between the main text and the results of the topic modeling analysis.

그림 4. 토픽 분포도

Fig. 4. Topic distribution diagram

부동산 데이터 기반 창업 지원 및 프롭테크 육성이 결합하는 구조를 보인다. 이 그룹은 공공데이터 정책의 경제적 산출물(Outcome)에 초점을 맞추고 있다. 데이터 댐을 통한 일자리 창출과 프롭테크 산업 육성은 데이터의 경제적 가치 실현이라는 공통된 목표를 공유하며, 이는 혁신이 가속화되고 있음을 의미한다[30].

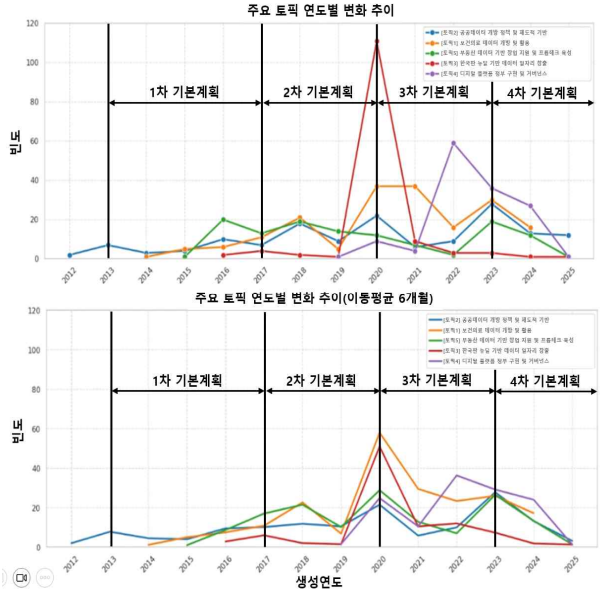
덴드로그램 하단에 붉은색 링크로 표시된 두 번째 거시적 군집은 지역 주도의 공간적 확산 및 구현 영역을 대변한다. 이 군집에서는 [토픽 10] 지역 거점 인공지능 클러스터 조성 및 산업 육성과 [토픽 9] 스마트시티 데이터 플랫폼 구축 및 도시 문제 해결이 가장 먼저 결합한다. 이는 중앙정부의 추상적인 데이터 정책이 광주 AI 집적단지나 부산 스마트시티와 같이 구체적인 물리적 공간에 투영되어 실현되고 있음을 의미한다. 여기에 [토픽 7] 공공데이터 창업 지원 거점 및 오픈 이노베이션이 결합되는 구조는 오픈스퀘어-D와 같은 창업 지원 공간이 지역 거점 및 스마트시티와 연계되어 지역 혁신 생태계의 허브 역할을 수행하고 있음을 실증적으로 보여준다[31].

그림 4의 토픽 분포도(Intertopic Distance Map)는 이러한 구조적 해석을 뒷받침하며 담론의 지형을 보여준다. [토픽 1], [토픽 2], [토픽 3], [토픽 6] 등은 2차원 평면의 중앙부에 밀집된 클러스터를 형성하며 공공데이터 담론의 핵심 축(Backbone)을 형성하고 있다. 이는 해당 의제들이 시기와 정권에 관계없이 정책 담론의 근간을 이루는 필수 요소임을 의미한다.

반면 [토픽 5]나 [토픽 4]는 중심부에서 일정 거리 떨어져 분포하는데 이러한 배치는 해당 토픽들이 일반적인 행정데이터의 개방 차원을 넘어 특정 산업 생태계를 간접적으로 대변하거나 차세대 정부의 패러다임이라는 독자적이고 전문적인 담론 영역을 형성하고 있음을 실증한다. 결과적으로 본 분포도는 공공데이터 정책이 범용적 제도에서 특화된 산업 가치 창출로 분화되고 확장되는 과정을 공간적으로 입증하고 있다.

4-3 공공데이터 정책 시기별 동향 분석

공공데이터 개방 정책이 본격화된 시점부터 현재에 이르기까지 담론의 중심 주제가 어떻게 변화했는지 규명하기 위해 주요 토픽에 대한 동적 토픽 모델링 결과를 정부의 공공데이터 제공 및 이용 활성화 기본계획의 추진 시기에 따라 시기별로 그림 5와 같이 구분하였다.



*Written in Korean to ensure a consistent context and coherent explanation between the main text and the results of the topic modeling analysis.

그림 5. 정책시기별 주요 토픽 연도별 변화 추이

Fig. 5. Yearly trends of major topics by policy period

먼저, 제1차(2013~2016) 및 제2차(2017~2019) 기본계획 추진 시기에는 [토픽 2] 공공데이터 개방 정책 및 제도적 기반이 꾸준한 발생 빈도를 보이며 담론의 기저를 형성하고

있다. 이는 2013년 「공공데이터의 제공 및 이용 활성화에 관한 법률」 제정과 함께 공공데이터 개방의 법적·제도적 기틀을 마련하고 데이터의 양적 개방을 확대하는 데 정책적 역량이 집중되었음을 시사한다. 또한, 주목할 만한 점은 2016년을 기점으로 [토픽 5] 부동산 데이터 기반 창업 지원 및 프롭테크 육성 관련 담론이 상승세를 보인다는 점인데 이는 정부가 국토교통 데이터 등 고가치 공공데이터를 중점적으로 개방하고, 이를 활용한 창업경진대회 등을 통해 프롭테크(Proptech)와 같은 실질적인 데이터 활용 산업 생태계를 조성하기 시작한 흐름을 반영한다.

제3차 기본계획(2020~2022) 추진 시기에서 가장 급격한 담론의 변화가 관측된 시점은 2020년 전후이다. 그래프에서 확인할 수 있듯이 2020년을 기점으로 [토픽 3] 한국판 뉴딜 기반 데이터 일자리 창출이 폭발적으로 급증하여 정점을 기록하였다. 이는 팬데믹으로 인한 경제 위기 극복을 위해 정부가 한국판 뉴딜의 일환으로 데이터 댐 구축 등 대규모 재정이 투입된 데이터 일자리 사업을 추진함에 따라, 공공데이터 정책이 국가 경제 회복의 핵심 아젠다로 부상했기 때문이다. 동일 시기에 [토픽 1] 보건의료 데이터 개방 및 활용 또한 동반 상승하는 양상을 보이는데, 이는 팬데믹 상황에서 공적 마스크 재고 데이터, 확진자 동선 정보 등 보건의료 데이터의 개방이 사회적 안전망 구축에 결정적인 역할을 수행했음을 확인할 수 있다.

빈도 데이터와 6개월 이동평균 데이터를 비교 분석한 결과, 2020년 원본 데이터에서는 [토픽 3] 한국판 뉴딜 기반 데이터 일자리 창출의 빈도가 압도적으로 높게 나타났으나, 이동평균 그래프에서는 [토픽 1] 보건의료 데이터 개방 및 활용이 [토픽 3]과 유사하거나 오히려 상회하며 장기간 지속되는 현상이 관찰된다. 이는 한국판 뉴딜 관련 이슈가 정부의 정책 발표나 특정 이벤트 시점에 따라 단기적으로 급증하는 높은 휘발성을 띠는 반면, 의료 및 보건 관련 이슈는 국민의 생명과 직결된 생활 밀착형 주제로서 상대적으로 언론의 보도가 꾸준하고 지속적으로 이루어졌음을 의미한다. 즉, 강력한 정책적 드라이브에 의한 토픽은 단기간에 강한 파급력을 가지지만, 지속성 측면에서는 사회적 실수요에 기반한 토픽이 더 견고하게 유지됨을 알 수 있다.

제4차 기본계획(2023~2025) 추진 시기인 2022년 하반기부터는 새로운 담론의 전환이 확인된다. [토픽 3] 한국판 뉴딜 기반 데이터 일자리 창출의 비중이 소멸 수준으로 감소하고, 그 자리를 [토픽 4] 디지털 플랫폼 정부 구현 및 거버넌스가 대체하며 가파른 상승세를 보이고 있다. 이는 공공데이터 정책의 패러다임이 정부 주도의 일방적 개방과 데이터 댐 구축을 넘어, 데이터를 매개로 정부와 민간, 국민이 유기적으로 연결되는 디지털 플랫폼 정부 구현으로 연계됨을 보여준다. 과거의 정책이 데이터의 축적에 방점을 두었다면, 현재는 민관 협력을 통한 데이터의 연계와 융합 서비스 창출이 핵심 의제로 부상했다는 점을 시사한다.

V. 결 론

5-1 연구의 결론 및 시사점

본 연구는 2012년부터 2025년까지 축적된 뉴스 빅데이터에 BERTopic 기반의 토픽 모델링을 적용하여, 공공데이터 개방의 주요 이슈와 정책 시기별 동향 분석을 실증적으로 규명하였다. 분석 결과를 종합해 볼 때, 공공데이터 정책은 단순한 행정 정보의 공개 차원을 넘어 데이터 경제 활성화와 지능형 정부 구현을 위한 핵심 인프라로서 그 위상과 역할이 지속적으로 확대되어 왔음을 알 수 있다.

연구의 주요 시사점으로 공공데이터 관련 주요 의제는 [토픽 2] 공공데이터 개방 정책 및 제도적 기반이 형성되던 2012년을 전후하여 본격적으로 태동하였다. 이후 2020년 [토픽 3] 한국판 뉴딜 기반 데이터 일자리 창출과 [토픽 6] 디지털 뉴딜과 데이터 인프라 구축 이슈가 부상하며 관련 보도가 폭발적으로 증가하였는데, 이는 팬데믹 위기 극복을 위한 데이터 댐 구축 사업이 공공데이터 정책의 양적 확산을 주도했음을 시사한다. 특히, [토픽 1] 보건의료 데이터 개방 및 활용은 시기와 무관하게 꾸준히 높은 비중을 차지하며, 국민 건강과 직결된 고수요 데이터로서 인공지능 헬스케어 서비스의 원천 자원으로 활발히 소비되고 있음을 입증하였다. 또한, 2022년을 기점으로 [토픽 4] 디지털 플랫폼 정부 구현 및 거버넌스 관련 담론이 급부상한 점은 정책의 지향점이 데이터 구축을 넘어 민관이 협력하는 플랫폼 기반의 생태계 조성으로 진화하고 있음을 보여준다. [토픽 5] 부동산 데이터 기반 창업 지원 및 프롭테크 육성, [토픽 7] 공공데이터 창업 지원 거점 및 오픈 이노베이션의 등장은 정부가 경진대회나 창업 지원 센터 설립 등 행정력을 집중하여 데이터 활용의 저변을 민간 산업 영역으로 확장해 왔음을 방증한다[10].

이러한 분석 결과를 바탕으로 도출한 정책점 시사점으로 공공데이터의 개방 패러다임을 양적 확대에서 인공지능 학습에 최적화된 질적 고도화(AI-Readiness)로 전환해야 한다. 생성형 AI 시대의 경쟁력은 기계가 학습하고 추론할 수 있는 고품질 데이터셋 확보에 달려 있다. 따라서 비정형 데이터의 표준화와 메타데이터의 정교화를 통해 개발자가 별도의 전처리 과정 없이 즉시 활용 가능한 수준으로 데이터의 품질을 제고하기 위한 정책적 노력이 필요하다[32].

다음으로 지역 특화 데이터의 발굴을 통한 지역 균형 발전 전략의 적극적인 추진이 필요하다. 연구 결과, [토픽 9] 스마트시티 데이터 플랫폼 구축 및 도시 문제 해결과 [토픽 10] 지역 거점 인공지능 클러스터 조성 및 산업 육성이 독립적인 군집을 형성하고 있음이 확인되었다. 이는 공공데이터가 중앙정부의 전유물이 아니라 광주의 AI 집적단지나 부산의 스마트시티와 같이 지역 특화 산업을 육성하는 동력으로 작용하고 있음을 의미한다. 따라서, 각 지자체의 산업적 특성과 연계된 특화 데이터를 전략적으로 개방하고, 이를 활용하는 지역 스타트업을 육성함으로써 데이터 경제의 낙수 효과를 지역

사회로 확산시키는 정책적 노력이 요구된다.

마지막으로, 데이터 활용의 성과가 사회적 가치 창출로 이어지는 선순환 구조를 확립해야 한다. 분석된 토픽들은 프롭테크와 같은 경제적 가치뿐만 아니라 [토픽 8] 4차 산업혁명 대응 및 규제 혁신과 연계되어 사회 문제 해결을 지향하고 있다. 공공데이터 개방의 궁극적인 목적은 시민 삶의 질 향상에 있으므로, 민간의 창의적인 아이디어가 공공 서비스 혁신으로 이어질 수 있도록 민관 협력 거버넌스를 강화하고, 데이터 활용 과정에서의 개인정보 보호 및 인공지능 윤리 가이드라인을 정립하는 제도적 보완이 병행되어야 한다[7].

본 연구는 뉴스 빅데이터 기반의 텍스트 마이닝을 통해 공공데이터 개방 정책이 지난 10여 년간 단순한 데이터 개방에서 데이터 경제의 핵심 인프라로, 그리고 디지털 플랫폼 정부의 근간으로 진화해 온 궤적을 실증했다는 점에서 학술적 의의를 지닌다.

5-2 연구의 한계점

본 연구는 BERTopic 기반의 뉴스 빅데이터 분석을 통해 공공데이터 개방 관련 담론의 동향과 의미적 구조를 실증적으로 규명하였으나, 연구 데이터 원천과 분석 방법론적 측면에서 여러 한계점을 내포하고 있다. 분석 대상과 관련하여 한국언론진흥재단의 빅카인즈에서 제공하는 뉴스 데이터로 한정하였는데, 뉴스 미디어는 정책 이슈를 사회적 의제로 확산시키는 중요한 매개체이나, 언론사의 편집 방향이나 보도 관행에 따라 특정 이슈가 과대 포장되거나 축소되는 프레이밍 효과가 발생할 수 있다. 따라서, 토픽의 빈도 급증이 실제 정부의 행정력 투입이나 예산 집행 규모와 반드시 정비례한다고 단정하기는 어렵다. 이는 언론 보도가 정책의 실체를 기계적으로 반영하기보다는 언론이 중요하다고 판단한 이슈를 선별적으로 강조하는 경향이 있다는 선행 연구들의 지적과 맥락을 같이한다[33]. 이는 본 연구의 분석 결과가 정책 변화를 단정적으로 결론지을 수 없다는 것을 의미한다. 또한, 본 연구에서 도출된 'AI-Readiness'로의 정책 패러다임 전환 제언은 분석 설계 단계에서 AI 관련 키워드를 포함함에 따라 일정 선행적으로 영향을 미친 측면이 있다. 이는 공공데이터 담론 전체의 보편적 현상으로 일반화하기에는 무리가 있을 수 있으며 기술 진보와 도입에 따른 정책적 수요 변화를 특정 관점에서 분석한 결과로 해석되어야 한다.

데이터의 구체적인 형식과 관련하여, 저작권 문제로 인해 뉴스 기사의 전체 본문이 아닌 키워드 및 특성 추출 데이터를 중심으로 임베딩을 수행하였다는 점이다. 뉴스 데이터의 핵심 정보를 함축한 키워드를 활용하는 것이 연산 효율성을 높이고 노이즈를 줄이는 데 효과적일 수 있으나, 문장 내에 존재하는 미세한 문맥 정보나 서사적 구조가 일부 소실될 가능성을 배제할 수 없다. 이러한 점은 텍스트 마이닝 분야에서 지속적으로 제기되어 온 과제이다[34]. 따라서, 향후 연구에서는 전문 확보가 가능한 데이터를 활용하거나, 요약문과 본문을 결합하는 방식 등을 통해 분석의 정확도를 제고할 필요가 있다.

마지막으로 BERTopic 알고리즘이 기존 토픽 모델링 기법에 비해 해석 가능성이 뛰어나지만, 도출된 토픽의 명명(Labeling) 및 해석 과정에서 연구자의 주관적 판단이 개입될 여지가 여전히 존재한다. 본 연구에서는 [토픽 1]부터 [토픽 10]까지의 명칭을 부여함에 있어 주요 키워드와 문헌을 참고하였으나, 비지도 학습(Unsupervised Learning)의 특성상 도출된 군집이 기계적으로는 타당하더라도 사회과학적으로 유의미한 해석을 도출하는 데에는 한계가 따를 수 있다. 이는 텍스트 분석 결과의 타당성을 확보하기 위해 정성적 내용 분석(Qualitative Content Analysis)을 병행해야 한다는 연구[35]와 같이, 향후 연구에서는 도메인 전문가 그룹의 델파이 조사나 검토 과정을 거쳐 토픽 해석의 객관성을 보완하는 절차가 요구된다고 볼 수 있다.

참고문헌

[1] W. Sung, "A Study on the Improvement of Big Data Policy in the Public Sector," *The Korea Association for Policy Studies*, Vol. 25, No. 2, pp. 125-150, 2016.

[2] J. Lim and G. Choi, "The Influence of Open Data Policies on Public Innovation," *Journal of Korean Institute of Industrial Engineers*, Vol. 43, No. 1, pp. 19-29, February 2017. https://doi.org/10.7232/jkiie.2017.43.1.019

[3] H. Seo, "An Empirical Study on Open Government Data: Focusing on ODB and OUR Index," *Informatization Policy*, Vol. 24, No. 1, pp. 48-78, March 2017.

[4] Y. Gao and M. Janssen, "The Open Data Canvas: Analyzing Value Creation from Open Data," *Digital Government: Research And Practice*, Vol. 3, No. 2, pp. 1-15, 2022.

[5] M. Grootendorst, "BERTopic: Neural Topic Modeling With A Class-Based TF-IDF Procedure," arXiv:2203.05794, March 2022. https://doi.org/10.48550/arXiv.2203.05794

[6] C. Bizer, T. Heath, and T. Berners-Lee, Linked Data, in *Semantic Services, Interoperability and Web Applications: Emerging Concepts,* pp. 205-227, 2011.

[7] A. Zuiderwijk and M. Janssen, "Open Data Policies, Their Implementation and Impact: A Framework for Comparison," *Government Information Quarterly*, Vol. 31, No. 1, pp. 17-29, Jan 2014.

[8] E. Kalampokis, E. Tambouris, and K. Tarabanis, "A Classification Scheme for Open Government Data: Review and Inspection of National Portals," *International Journal of Web Engineering and Technology*, Vol. 6, No. 3, pp. 266-285, 2011.

[9] J. Crusoe, A. Simonofski, A. Clarinval, and E. Gebka, "The Impact of Impediments on Open Government Data Use:

Insights from Users," in *Proceeding of the 2019 13th International Conference on Research Challenges in Information Science*, Brussels, Belgium, pp. 297-290, 2019.

[10] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, Adoption Barriers and Myths of Open Data and Open Government," *Information Systems Management*, Vol. 29, No. 4, pp. 258-268, October 2012. https://doi.org/10.1080/10580530.2012.716740

[11] E. Ruijer and A. Meijer, "Open Government Data as an Innovation Process: Lessons from a Living Lab Experiment," *Public Performance & Management Review*, Vol. 43, No. 3, pp. 613-635, February 2020. https://doi.org/10.1080/15309576.2019.1568884

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceeding of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, pp. 4171-4186, June 2019.

[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in *Proceeding of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, pp. 3982-3992, 2019.

[14] L. McInnes, J. Healy, and N. Saul, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," arXiv:1802.03426, 2018. https://doi.org/10.48550/arXiv.1802.03426

[15] L. McInnes, J. Healy, and S. Astels, "Hdbscan: Hierarchical Density Based Clustering," *Journal of Open Source Software*, Vol. 2, No. 11, 205, 2017. https://doi.org/10.21105/joss.00205

[16] H. J. Kim and S. J. Kim, "Healthcare Data in the Intelligent Information Age How to Activate Use," *Journal of Korean Association for Regional Information Society*, Vol. 25, No. 3, pp. 1-21, September 2022.

[17] S. Cho and B. Choi, "The Overview of the Public Opinion Survey and Emerging Ethical Challenges in the Healthcare Big Data Research," *Journal of KAIRB*, Vol. 4, No. 1, pp. 16-22, 2022.

[18] E. Choi, "Utilization of Medical Data and Policy Implications," *KIET Industrial Economy*, No. 296, pp. 23-33, 2023.

[19] K. Yoon, Enhancing Data-Driven Public Administration: Focused on Public Data Integration, Korea Institute of Public Administration, Seoul, December 2019.

[20] S. Kim, "Significance and Future Tasks of the Amendment to the Three Data Acts," *Journal of Convergence Security*, Vol. 20, No. 2, pp. 59-68, January 2020.

[21] Korean Government, Korean New Deal Comprehensive Plan: National Briefing, Sejong, 2020.

[22] Digital Platform Government Committee, Digital Platform Government Implementation Plan, Seoul/Sejong, 2023.

[23] H. Ju, H. Choi, and Y. Choi, "Digital Platform Government and Government Innovation: Focucing on REdefining Roles and Functions of Government," *The Korean Journal of Local Government Studies*, Vol. 26, No. 3, pp. 307-327, November 2022.

[24] S. Kim and H. Jang, "A Study of the Capacity Enhancement and Countermeasure of the Real Estate Industry in the Prop-Tech Era," *SH Urban Research & Insight*, Vol. 10, No. 3, pp. 97-117, December 2020.

[25] Y. Heo and S. Kim, PropTech Companies: A New Future for the Real Estate Industry, Construction & Economy Research Institute of Korea, Seoul, Issue Focus 2019-01, pp. 1-38, 2019.

[26] M. Park, "Korean New Deal, Data, Network, and AI-Based Data Dam for National Digital Transformation," *Korea Information Processing Society Review*, Vol. 27, No. 2, pp. 13-20, 2020.

[27] C. Park, K. Choi, J. Choi, M. Yoon, and J. Lee, Employment Effects of Public Data Release and Utilization Support Policies, Korea Labor Institute, Sejong, Policy Research Report 2023-04, December 2023.

[28] W. Lee, J. Kim, S. Lee, D. Kim, and K. Jung, A Study on Legal Issues of ICT and Its Policy Implications in the Era of the Fourth Industrial Revolution, Korea Legislation Research Institute, Sejong, November 2016.

[29] H. Na and B. Lee, "Analyzing Technology Mining and Issue Scanning for Exploring Regional Artificial Intelligence Industry Development Strategies," *Journal of the Korea Contents Association*, Vol. 25, No. 3, pp. 71-87, March 2025.

[30] E. Kang and H. Park, "Strategic Innovation in the Era of Proptech : A Case Study of Zigbang's Digital Transformation," *Journal of Cultural Industry Studies*, Vol. 25, No. 3, pp. 51-62, 2025.

[31] E. Kim and T. Lee, "A Study on the Establishment of Smart City Platform According to the Fourth Industrial Revolution; Focusing on the Case of Pohang Smart City," *International Commerce and Information Review*, Vol. 21, No. 2, pp. 205-229, June 2019.

[32] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A Systematic Review of Open Government Data Initiatives," *Government Information Quarterly*, Vol. 32, No. 4, pp. 399-418, October 2015.

[33] R. M. Entman, “Framing: Toward Clarification of a Fractured Paradigm,” *Journal of Communication*, Vol. 43, No. 4, pp. 51-58, 1993.

[34] L. Hong and B. D. Davison, “Empirical Study of Topic Modeling in Twitter,” in *Proceedings of the First Workshop on Social Media Analytics*, Washington, DC, pp. 80-88, July 2010.

[35] J. Grimmer and B. M. Stewart, “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts,” *Political Analysis*, Vol. 21, No. 3, pp. 267-297, 2013. https://doi.org/10.1093/pan/mps028

정준영(Junyoung Jeong)

2025년 : Ph.D. in Technology Management, Sungkyunkwan University

2017년～현 재: Principal Researcher, National Information Society Agency (NIA)

※관심분야 : Artificial Intelligence, Data Policy, Open Data, Policy Research, Science and Technology Information