LDA 기반 텍스트마이닝 접근을 통한 AI와 탄소발자국 연구의 이중성 분석

박소연*, 최승일*
*국립공주대학교 산업공학과
e-mail:sichoi@kongju.ac.kr

An LDA-Based Text Mining Analysis of the Dual Role of AI and Carbon Footprint Research

So-Yeon Park*, Seungil Choi*
*Dept. of Industrial Engineering, Kongju National University

요 약

본 연구는 인공지능(AI)과 탄소발자국 연구의 이중적 관계를 규명하기 위해 2000-2024년 논문을 수집하고, 특히 변곡점 이후인 2022-2024년 3개년을 중점적으로 분석하였다. 이를 위해 LDA 토픽모델링을 적용하고 Perplexity와 Coherence 지표를 활용하여 최적의 토픽 수와 하이퍼파라미터를 도출하였다. 분석 결과, 연구는 탄소배출 유발, 산업·에너지 효율화, 지속가능성, 정책 담론의 네 주제로 구조화되었으며, 연도별 비중 변화를 통해 유발 연구의 감소와 저감·지속가능성 연구의 확대를 확인하였다.

1. 서론

인공지능(AI)은 산업 전반에서 혁신을 이끌었으나, 대규모 연산과 데이터센터 운영으로 인한 막대한 에너지 소비와 탄소배출을 초래하며 환경적 부담을 가중시켰다[1]. 반면, AI는 에너지 예측, 재생에너지 최적화, 산업 공정 효율화 등을 통해 탄소저감에도 기여할 수 있다[2]. 본 연구는 이러한 AI의 유발과 저감의 이중성에 주목하여, COP26(2021)과 ChatGPT 공개(2022)를 계기로 급격히 확산된 2022-2024년 논문을 중점 분석하였다. 이를 통해 AI-탄소 연구의 주제 구조와 시계열적 변화를 규명하고, 향후 학문적·정책적 시사점을 제시하고자 한다. 이를 위해 LDA 토픽모델링과 연도별 토픽 비중 분석을 적용하였다.

2. 데이터 수집

본 연구는 인공지능(AI)과 탄소발자국(carbon footprint) 간의 관계를 계량적으로 규명하기 위하여, Google Scholar를 주요 데이터베이스로 설정하고 학술정보 수집 도구인 Publish or Perish(Version 8.1)을 활용하였다. 검색식은 “AI AND carbon”과 “Artificial Intelligence AND carbon”을 논문 제목(title)에 포함하는 조건으로 설정하여, 주제 적합성이 높은 문헌을 확보하였다. 특허(patent)와 단순 인용문헌(citation entry)은 제외하여 학술 논문만을 대상으로 하였다.

이 과정을 통해 2000-2024년 사이 발표된 총 531편의 원자료를 수집하였다. 이후 출판연도가 누락된 22편과 중복으로 확인된 14편을 제거하고, 최종적으로 495편을 분석 표본으로 확정하였다. 이 가운데 2022-2024년 발표된 413편은 변곡점 이후 연구 동향을 집중적으로 규명하기 위한 핵심 분석 대상으로 활용되었다.

3. 데이터 전처리

분석 텍스트는 각 논문의 제목(title)과 초록(abstract)을 병합하여 하나의 분석 단위로 구축하였다. Python 기반 자연어처리 도구 spaCy와 NLTK를 활용하여 소문자 변환, 특수문자 및 숫자 제거, 토큰화(tokenization), 불용어(stopwords) 제거, 표제어 추출(lemmatization) 절차를 수행하였다. 불용어 사전은 spaCy 기본 불용어를 기반으로, study, result, effect, data, paper, research, method, model, analysis 등 연구 전반에 빈번하게 등장하나 분석적 구분력이 낮은 단어를 추가하였다. 특히 데이터 수집 과정에서 사용된 기저 검색어(ai, artificial, intelligence, carbon)는 토픽모델링 분석 결과에 편향을 유발할 수 있으므로 전처리 단계에서 제거하였다. 최종적으로 전처리된 텍스트는 불필요한 잡음을 최소화하면서 핵심 개념을 보존하였으며, 이를 기반으로 토픽모델링을 수행하였다.

4. 토픽 모델링

변곡점 이후(2022-2024)에 발표된 413편 논문을 대상으로 LDA 토픽모델링을 수행하였다. 토픽 수(K)는 3~8 범위에서 탐색하였으며, Perplexity와 Coherence 지표를 기준으로 성능을 비교하였다. <표 1>은 K값 변화에 따른 지표 산출 결과를 나타낸다.

[표 1] 토픽 수(K)별 Perplexity 및 Coherence 지표

K	Perplexity	Log-Perplexity	Coherence
3	154.48	−7.27	0.247
4	154.83	−7.27	0.288
5	156.15	−7.29	0.243
6	159.91	−7.32	0.246
7	159.17	−7.31	0.313
8	161.75	−7.34	0.268

분석 결과, K=7에서 가장 높은 Coherence 값을 기록하였으나 토픽 간 중복이 발생하여 주제 해석력이 저하되었다. 반면 K=4는 수치적 안정성과 주제 구분의 명확성이 확보되어 최적 모형으로 선정되었다. 이어 α(alpha)와 η(eta)의 조합을 검토한 결과는 <표 2>와 같다.

[표 2] α·η 조합에 따른 Coherence 값

α	η	Coherence
0.01	0.01	0.320
0.5	0.50	0.319
0.5	0.10	0.311
0.5	0.01	0.304
0.1	0.01	0.292
0.1	0.10	0.288
0.1	0.50	0.269
0.01	0.50	0.265

수치적으로는 α=0.01, η=0.01이 가장 높은 값을 보였으나, 지나치게 희소한 분포를 형성하였다. 반면 α=0.5, η=0.50은 유사한 성능을 유지하면서도 토픽별 키워드 구성이 균형적으로 분화되었다. 따라서 본 연구는 K=4, α=0.5, η=0.50을 최적 모형으로 확정하였다.

확정 모형(K=4, α=0.5, η=0.50)에서 도출된 토픽별 주요 키워드와 해석은 <표 3>과 같다.

[표 3] 토픽별 주요 키워드와 주제 해석

구분	주요 키워드	비중(%)	주제 해석
토픽 1	emission, capture, dioxide, storage, prediction	30.8	온실가스 배출과 CCS(포집·저장) 및 배출량 예측(유발 측면)
토픽 2	low, soil, energy, footprint, design, reduce	20.6	산업·에너지 효율화 및 저탄소 설계·응용(저감 측면)
토픽 3	emission, footprint, technology, reduce, green, sustainable	34.1	AI 기반 탄소저감 기술 및 지속가능성 연구(저감 측면)
토픽 4	emission, neutrality, system, transition, reduction	14.5	탄소중립·에너지 전환 및 정책적 담론(정책적 대응)

토픽 1은 주로 AI 학습·데이터센터 운영으로 인한 탄소배출의 유발을 진단하고 관리하는 연구에 해당한다. 토픽 2와 3은 AI를 산업 효율화와 저감 기술에 활용하는 흐름으로, 저감의 도구로서 AI의 역할을 보여준다. 토픽 4는 국제적 탄소중립 담론과 연결되어, 정책·제도적 대응 차원에서 AI가 편입되는 연구군을 나타낸다. 즉, 네 개 토픽의 분포는 AI가 "탄소배출을 유발하는 동시에 저감의 도구로 기능한다"는 이중성을 계량적으로 뒷받침한다. <표 4>는 2022-2024년 기간 동안 토픽별 비중 변화를 제시한다.

[표 4] 연도별 토픽 비중 변화(2022-2024)

연도	토픽 1	토픽 2	토픽 3	토픽 4
2022	33.1%	24.5%	27.1%	15.3%
2023	29.8%	21.7%	32.5%	16.1%
2024	27.9%	19.6%	36.5%	15.9%

분석 결과, 토픽 1(유발 연구)은 꾸준히 감소한 반면, 토픽 3(저감 및 지속가능성 연구)은 뚜렷한 증가세를 보이며 2024년 가장 높은 비중을 차지하였다. 토픽 2(산업 효율화 연구)는 점차 축소되었으며, 토픽 4(정책 담론 연구)는 낮은 수준이지만 안정적으로 유지되었다. 이를 통해 변곡점 이후 연구의 중심이 유발 연구에서 저감·지속가능성 연구로 이동했음을 확인할 수 있다. 이는 AI의 이중적 성격이 시간에 따라 어떻게 재구성되는지를 보여주는 결과라 할 수 있다.

5. 결론

본 연구는 2000-2024년 논문 데이터를 기반으로, 변곡점 이후 2022-2024년 413편을 대상으로 LDA 토픽모델링과 연도별 비중 분석을 수행하였다. 그 결과, 연구는 탄소배출 유발, 저감, 지속가능성, 정책 담론의 네 주제로 구조화되었으며, 유발 연구는 감소하고 저감·지속가능성 연구가 확대되는 흐름을 보였다. 이는 AI가 탄소배출의 원인인 동시에 해결의 도구로 기능하는 이중성을 실증적으로 보여주며, 향후 Green AI와 Sustainable AI 전략의 필요성을 시사한다.

참고문헌

[1] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," Proc. 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 3645–3650, 2019.

[2] Y. Wang, X. Li, and Z. Li, "Artificial intelligence for carbon neutrality: Opportunities and challenges," Renewable and Sustainable Energy Reviews, vol. 185, 113635, 2024.