JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 2, pp. 231-239, February 2026
https://doi.org/10.9708/jksci.2026.31.02.231

A Comparative Analysis of BERTopic and LDA for Topic Modeling of Korean Sleep Health Discourse on Social Media

JongHwi Song*

*Research Professor, Institute of AI Convergence Science, Yonsei University, Wonju, Korea

[Abstract]

This study compared the performance of BERTopic and Latent Dirichlet Allocation (LDA) for topic modeling of Korean sleep health-related social media text. A total of 8,002 blog posts were collected from Naver using nine sleep-related keywords between March and October 2025. Both methods were applied to the same dataset, and their performance was evaluated using metrics including the number of topics, noise ratio, distribution entropy, and topic coherence. The results indicated that BERTopic identified 9 topics with a noise ratio of 22.8%, whereas LDA yielded 6 effective topics with a significantly lower noise ratio of 0.9%. BERTopic demonstrated higher distribution uniformity (0.852) compared to LDA (0.804), indicating more balanced topic assignments. LDA achieved a coherence score (C_V) of 0.5287. The cross-tabulation analysis revealed that BERTopic's "Melatonin/Hormone" topic showed 84.1% concentration in LDA's "Insomnia General" topic, demonstrating high consistency for well-defined topics. This study provides methodological insights for researchers selecting topic modeling approaches for Korean health-related text analysis.

▸ **Key words**: Topic Modeling, BERTopic, LDA, Sleep Health, Social Media Analysis, Text Mining

[요 약]

본 연구는 한국어 수면 건강 관련 소셜미디어 텍스트의 토픽 모델링을 위해 BERTopic과 잠재 디리클레 할당(LDA)의 성능을 비교 분석하였다. 2025년 3월부터 10월까지 네이버에서 9개의 수면 관련 키워드로 총 8,002개의 블로그 게시물을 수집하였다. 동일한 데이터셋에 두 방법론을 적용하고, 토픽 수, 노이즈 비율, 분포 엔트로피, 토픽 일관성 등의 지표로 성능을 평가하였다. 분석 결과, BERTopic은 9개의 토픽을 도출하며 22.8%의 노이즈 비율을 보인 반면, LDA는 6개의 유효 토픽을 도출하며 0.9%의 낮은 노이즈 비율을 나타냈다. BERTopic은 LDA(0.804)보다 높은 분포 균등성(0.852)을 보여 더 균형 잡힌 토픽 할당을 수행하였다. LDA의 일관성 점수(C_V)는 0.5287이었다. 교차분석 결과, BERTopic의 '멜라토닌/호르몬' 토픽은 LDA의 해당 토픽과 84.1%의 일치율을 보여 잘 정의된 주제에서 높은 일관성을 나타냈다. 본 연구는 한국어 건강 관련 텍스트 분석을 위한 토픽 모델링 방법론 선택에 실질적인 지침을 제공한다.

▸ **주제어**: 토픽 모델링, BERTopic, LDA, 수면 건강, 소셜미디어 분석, 텍스트 마이닝

• First Author: JongHwi Song, Corresponding Author: JongHwi Song
*JongHwi Song (jh_song@yonsei.ac.kr), Institute of AI Convergence Science, Yonsei University
• Received: 2025. 12. 30, Revised: 2026. 01. 25, Accepted: 2026. 01. 27.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

수면은 인간의 신체적, 정신적 건강을 유지하는 데 필수적인 생리적 과정이다. 세계보건기구(WHO)에 따르면, 전 세계 성인의 약 30%가 불면증 증상을 경험하고 있으며, 이는 삶의 질 저하와 다양한 건강 문제로 이어질 수 있다[1]. 한국에서도 수면 장애 환자 수가 지속적으로 증가하고 있으며, 이에 따라 수면 건강에 대한 대중의 관심이 높아지고 있다[2].

소셜미디어는 건강 정보 탐색과 경험 공유의 주요 플랫폼으로 자리 잡았다. 특히 블로그와 같은 플랫폼에서는 개인의 수면 문제 경험, 약물 사용 후기, 생활습관 개선 시도 등 다양한 주제가 활발하게 논의되고 있다[3]. 이러한 비정형 텍스트 데이터에서 의미 있는 주제를 추출하기 위해 토픽 모델링 기법이 널리 활용되고 있다[4].

토픽 모델링의 대표적인 방법론인 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)은 문서 내 단어의 동시출현 패턴을 기반으로 잠재 토픽을 추출하는 확률적 생성 모델이다[5]. LDA는 해석 가능성이 높고 계산 효율성이 우수하여 다양한 분야에서 활용되어 왔다. 그러나 단어의 의미적 관계를 충분히 반영하지 못하고, 짧은 텍스트에서 성능이 저하되는 한계가 있다[6].

최근 BERT(Bidirectional Encoder Representations from Transformers) 기반의 문장 임베딩을 활용한 BERTopic이 제안되었다[7]. BERTopic은 사전 학습된 언어 모델을 통해 문서의 의미적 유사성을 포착하고, HDBSCAN 클러스터링과 c-TF-IDF를 결합하여 토픽을 추출한다. 이 방법은 의미론적으로 일관된 토픽을 생성하고, 노이즈 문서를 명시적으로 분리할 수 있다는 장점이 있다[8].

그러나 한국어 건강 관련 텍스트에서 두 방법론의 성능을 체계적으로 비교한 연구는 부족한 실정이다. 특히 수면 건강과 같이 의학적 용어와 일상적 표현이 혼재된 텍스트에서 각 방법론의 특성을 파악하는 것은 연구자들의 방법론 선택에 중요한 지침이 될 수 있다.

따라서 본 연구는 네이버 블로그의 수면 건강 관련 게시물을 대상으로 BERTopic과 LDA의 토픽 모델링 성능을 비교 분석하고, 각 방법론의 특성과 적용 시사점을 도출하고자 한다. 구체적인 연구 문제는 다음과 같다.

첫째, BERTopic과 LDA는 동일한 데이터셋에서 어떻게 다른 토픽 구조를 생성하는가?

둘째, 두 방법론의 토픽 할당 결과는 얼마나 일치하는가?

셋째, 각 방법론은 어떤 상황에서 더 적합한가?

II. Related Works

1. LDA-based Topic Modeling

LDA는 Blei 등[5]이 제안한 확률적 생성 모델로, 문서가 여러 토픽의 혼합으로 구성되고, 각 토픽은 단어들의 확률 분포로 표현된다고 가정한다. LDA의 생성 과정은 다음과 같다. 먼저, 각 문서에 대해 토픽 분포 θ를 디리클레 분포에서 샘플링한다. 다음으로, 문서의 각 단어에 대해 토픽 z를 다항 분포에서 샘플링하고, 해당 토픽의 단어 분포 φ에서 단어를 샘플링한다[9].

LDA는 비지도 학습 방식으로 대규모 텍스트 코퍼스에서 잠재 주제를 발견할 수 있어, 뉴스 분석[10], 학술 문헌 분석[11], 소셜미디어 분석[12] 등 다양한 분야에서 활용되어 왔다. 건강 분야에서도 LDA를 활용한 연구가 활발히 수행되었다. Kim 등[13]은 온라인 건강 커뮤니티의 게시물을 LDA로 분석하여 환자들의 주요 관심사를 도출하였고, Park과 Lee[14]는 의료 리뷰 텍스트에서 서비스 품질 요인을 추출하였다.

그러나 LDA는 단어 간의 의미적 관계를 고려하지 않고 단순히 동시출현 빈도에 의존하기 때문에, 동의어나 다의어 처리에 한계가 있다. 또한, 짧은 텍스트에서는 단어 동시출현 정보가 부족하여 토픽 품질이 저하될 수 있다[6].

2. BERTopic

BERTopic은 Grootendorst[7]가 개발한 신경망 기반 토픽 모델링 기법이다. 이 방법은 세 단계로 구성된다. 첫째, 사전 학습된 언어 모델(BERT, Sentence-BERT 등)을 사용하여 문서를 고차원 임베딩 벡터로 변환한다. 이 과정에서 문서의 의미적 특성이 벡터 공간에 인코딩된다[15].

둘째, UMAP(Uniform Manifold Approximation and Projection)을 사용하여 고차원 임베딩을 저차원으로 축소하고, HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise)으로 클러스터링을 수행한다. HDBSCAN은 밀도 기반 클러스터링 알고리즘으로, 클러스터에 속하지 않는 노이즈 포인트를 명시적으로 식별할 수 있다[16].

셋째, c-TF-IDF(class-based Term Frequency-Inverse Document Frequency)를 통해 각 클러스터의 대표 키워드를 추출한다. c-TF-IDF는 전통적인 TF-IDF를 클래스(토픽) 수준으로 확장한 것으로, 각 토픽을 가장 잘 대표하는 단어를 식별한다[7].

BERTopic의 주요 장점은 의미적으로 유사한 문서들을 효과적으로 군집화하고, 클러스터링 과정에서 어떤 토픽에

도 할당되지 않는 노이즈 문서를 명시적으로 분리한다는 점이다[8]. 이러한 특성은 소셜미디어 텍스트와 같이 노이즈가 많은 데이터에서 유용하다.

3. A Comparative Study on Topic Modeling Methodology

토픽 모델링 방법론 간 비교 연구도 수행되어 왔다. Egger와 Yu[17]는 호텔 리뷰 데이터에서 LDA, NMF, BERTopic을 비교하여 BERTopic이 더 해석 가능한 토픽을 생성함을 보였다. Sia 등[18]은 다양한 데이터셋에서 전통적 토픽 모델과 신경망 기반 모델의 성능을 비교하였다. Abuzayed와 Al-Khalifa[19]는 아랍어 트위터 데이터에서 BERTopic과 LDA를 비교하여 BERTopic이 더 일관된 토픽을 생성함을 확인하였다.

그러나 한국어, 특히 건강 관련 소셜미디어 텍스트에서 두 방법론을 비교한 연구는 부족하다. 한국어는 교착어적 특성과 띄어쓰기 불규칙성으로 인해 영어와 다른 전처리가 필요하며[20], 이러한 언어적 특성이 토픽 모델링 성능에 미치는 영향을 파악할 필요가 있다.

III. Methodology

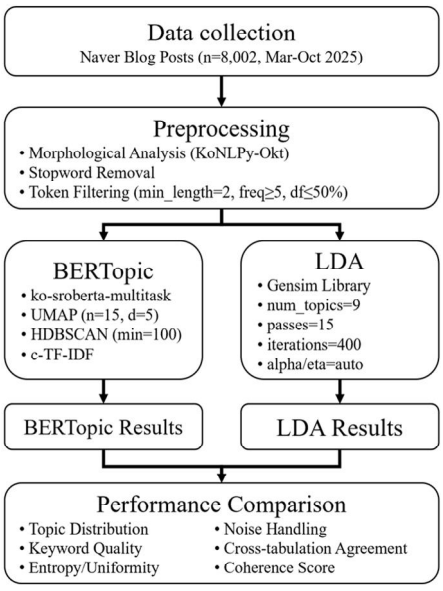

Fig. 1. Research Framework

1. Data Collection

본 연구는 2025년 3월부터 10월까지 네이버 블로그에서 수면 관련 게시물을 수집하였다. 네이버는 한국에서 가장 높은 점유율을 보이는 검색 엔진이며, 블로그 서비스는 개인의 건강 경험과 정보가 활발히 공유되는 플랫폼이다[21].

네이버 검색 API를 활용하여 9개의 키워드로 검색을 수행하였다. 키워드는 수면 문제 관련(불면증, 수면장애, 불면증극복, 잠이안와), 약물 관련(수면제, 스틸녹스, 멜라토닌), 비약물 개입 관련(수면루틴, 수면위생)으로 구분된다. 중복 게시물을 제거한 후 총 8,002개의 게시물이 분석에 사용되었다.

Table 1. Dataset Overview

Item	Value
Data Source	Naver Blog
Collection Period	March-October 2025
Search Keywords	9 (sleep-related terms)
Total Documents	8,002
Valid Documents (after filtering)	8,002
Average Tokens per Document	24.9
Vocabulary Size (after filtering)	3,450
Minimum Token Frequency	5
Maximum Document Frequency	50%

2. Text Preprocessing

수집된 텍스트는 다음과 같은 전처리 과정을 거쳤다. 첫째, HTML 태그, URL, 이메일 주소, 특수문자를 정규표현식을 사용하여 제거하였다. 둘째, KoNLPy 라이브러리의 Okt(Open Korean Text) 형태소 분석기를 사용하여 형태소 분석을 수행하고, 명사, 동사, 형용사를 추출하였다[22].

셋째, 분석의 정확도를 높이기 위해 불용어를 제거하였다. 불용어 목록에는 일반적인 한국어 불용어(이, 그, 저, 것, 등)와 함께 수면 관련 텍스트에서 과도하게 출현하는 일반 단어(오늘, 정말, 너무, 진짜 등)가 포함되었다. 넷째, 2글자 미만의 단어를 제거하여 의미 있는 단어만 분석에 포함하였다.

전처리 후 문서당 평균 24.9개의 토큰이 추출되었으며, 전체 코퍼스의 고유 단어 수는 3,450개였다.

3. LDA Modeling

LDA 분석은 Python의 Gensim 라이브러리를 사용하여 수행하였다[23]. 최적의 토픽 수를 결정하기 위해 토픽 수를 3부터 15까지 변화시키며 일관성 점수(Coherence Score, C_V)를 계산하였다. C_V는 토픽 내 상위 키워드들 간의 의미적 일관성을 측정하는 지표로, 값이 높을수록 토픽의 해석 가능성이 높음을 의미한다[24].

분석 결과, 토픽 수 5에서 일관성 점수가 0.5826으로 최대값을 보였으나, 토픽의 해석 가능성과 세분화 수준을 고려하여 0.5287의 점수를 보인 토픽 수 9를 최종 선정하였다. 모델의 하이퍼파라미터는 alpha='auto', eta='auto'로 설정하여 데이터에서 자동으로 추정하도록 하였으며, iterations=400으로 설정하였다.

각 문서는 토픽 확률 분포 중 가장 높은 확률의 토픽에 할당되었다. 단, 토픽 분류의 명확성을 높이기 위해 본 연구에서 설정한 임계값인 최대 토픽 확률이 0.3 미만인 경우 해당 문서는 명확한 토픽에 속하지 않는 것으로 간주하여 노이즈로 분류하였다.

4. BERTopic Modeling

BERTopic 분석은 한국어에 최적화된 사전학습 모델인 'jhgan/ko-sroberta-multitask'를 임베딩 모델로 사용하였다. 이 모델은 한국어 문장 임베딩에 특화되어 있으며, 의미적 유사성을 효과적으로 포착할 수 있다[25].

차원 축소를 위한 UMAP 파라미터는 n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine'으로 설정하였다. 클러스터링을 위한 HDBSCAN 파라미터는 min_cluster_size=100, min_samples=10으로 설정하여 최소 100개 이상의 문서를 포함하는 클러스터만 유효한 토픽으로 인정하였다.

BERTopic에서 토픽 -1에 할당된 문서는 어떤 클러스터에도 속하지 않는 노이즈로 처리하였다. 이는 HDBSCAN의 특성으로, 밀도가 낮은 영역의 문서들을 명시적으로 노이즈로 분류한다.

5. Evaluation Metrics

두 방법론의 성능을 비교하기 위해 다음의 지표를 사용하였다.

- 토픽 수: 각 방법론이 도출한 유효 토픽의 개수
- 노이즈 비율: 전체 문서 중 노이즈로 분류된 문서의 비율
- 분포 엔트로피: 토픽 분포의 균등성을 측정하는 지표로, 값이 클수록 균등한 분포
- 분포 균등성: 실제 엔트로피를 최대 엔트로피로 나눈 정규화된 값
- 토픽 일관성(C_V): 토픽 내 상위 키워드 간의 의미적 일관성
- 주요 토픽 집중도: 상위 3개 토픽에 할당된 문서의 비율

또한, 두 방법론의 토픽 할당 일치도를 분석하기 위해 교차표(Cross-tabulation) 분석과 카이제곱 검정을 수행하였다.

IV. Results

1. Topic Extraction Results

BERTopic과 LDA의 토픽 도출 결과를 Table 2에 제시하였다. BERTopic은 9개의 유효 토픽을 도출하였으며, LDA는 9개의 토픽 중 3개(토픽 1, 5, 8)가 노이즈 성격의 키워드를 포함하여 실질적으로 6개의 유효 토픽을 도출하였다.

Table 2. Comparison of Topic Modeling Performance

Metrics	BERTopic	LDA
Total Documents	8,002	8,002
Valid Topics	9	6
Noise Documents	1,825	74
Noise Ratio (%)	22.8	0.9
Distribution Entropy	2.701	2.077
Distribution Uniformity	0.852	0.804
Coherence Score (C_V)	-	0.5287
Top-3 Topic Concentration (%)	66.8	85.1

BERTopic은 22.8%의 노이즈 비율을 보인 반면, LDA는 0.9%의 낮은 노이즈 비율을 나타냈다. 이는 BERTopic이 HDBSCAN 클러스터링을 통해 어떤 토픽에도 명확히 할당되지 않는 문서를 명시적으로 분리하는 반면, LDA는 확률 분포에 기반하여 모든 문서에 토픽을 할당하기 때문이다. 단, LDA의 노이즈 비율 0.9%는 최대 토픽 확률이 0.3 미만인 문서만을 노이즈로 분류한 값이다. LDA의 9개 토픽 중 3개(토픽 1, 5, 8)가 해석이 어려운 노이즈 성격의 키워드를 포함하고 있으며, 이들 토픽에 할당된 문서까지 고려하면 LDA의 실질적 노이즈 비율은 더 높아질 수 있다. 따라서 두 방법론의 노이즈 비율은 정의 방식의 차이로 인해 직접 비교에 주의가 필요하다.

분포 균등성은 BERTopic(0.852)이 LDA(0.804)보다 높게 나타나, BERTopic이 문서를 더 균등하게 토픽에 할당함을 확인하였다. 반면, 주요 토픽 집중도는 LDA(85.1%)가 BERTopic(66.8%)보다 높아, LDA에서 상위 토픽에 문서가 더 집중되는 경향을 보였다.

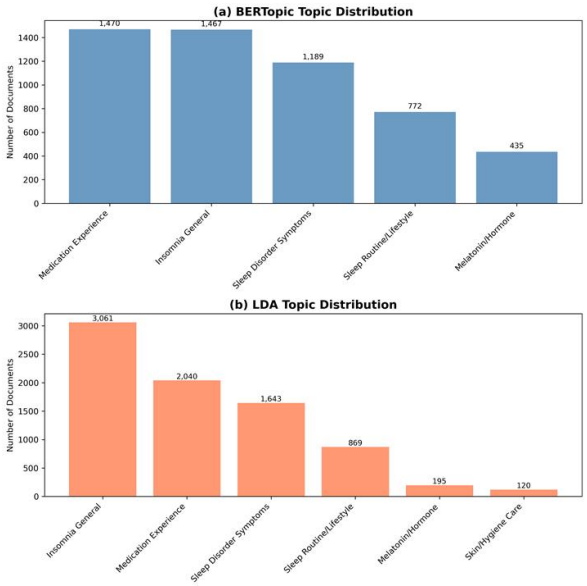

Fig. 2. Topic Distribution Comparison (BERTopic vs LDA)

2. Comparison of Representative Keywords

Table 3은 BERTopic과 LDA에서 도출된 주요 토픽과 대표 키워드를 비교한 것이다. 두 방법론 모두 수면 건강 관련 주요 주제를 포착하였으나, 키워드 구성에서 차이를 보였다.

Table 3. Representative Keywords by Topic (Top-5)

Topic	BERTopic Keywords	LDA Keywords
Sleep Routine/Lifestyle	루틴, 청소, 아침, 운동, 저녁	루틴, 침실, 청소, 수면, 운동
Insomnia General	불면증, 수면, 밤, 새벽, 극복	수면, 불면증, 적정, 관리, 극복
Medication Experience	스틸녹스, 수면제, 처방, 졸피뎀, 복용	스틸녹스, 수면제, 처방, 졸피뎀, 효과
Melatonin/Hormone	멜라토닌, 분비, 호르몬, 숙면, 유도	매트리스, 멜라토닌, 분비, 호르몬, 유도
Sleep Disorder Symptoms	증상, 치료, 장애, 피로, 통증	증상, 수면, 불면증, 치료
Skin/Hygiene Care	-	피부, 관리, 방법, 위생, 결핍

Note: Some BERTopic topics do not directly correspond to LDA topics

BERTopic은 '의존', '부작용', '진단'과 같이 더 구체적이고 맥락적인 키워드를 추출하는 경향을 보였다. 이는 BERT 임베딩이 단어의 문맥적 의미를 포착하기 때문으로 해석된다. 반면 LDA는 '먹다', '자다'와 같은 일반적인 동사가 포함되어 토픽의 명확성이 다소 낮았다.

3. Analysis of Topic Assignment Agreement

두 방법론 간 토픽 할당의 일치도를 분석하기 위해 교차표 분석을 수행하였다. Fig. 3은 BERTopic 토픽과 LDA 토픽 간의 문서 할당 비율을 히트맵으로 시각화한 것이다.

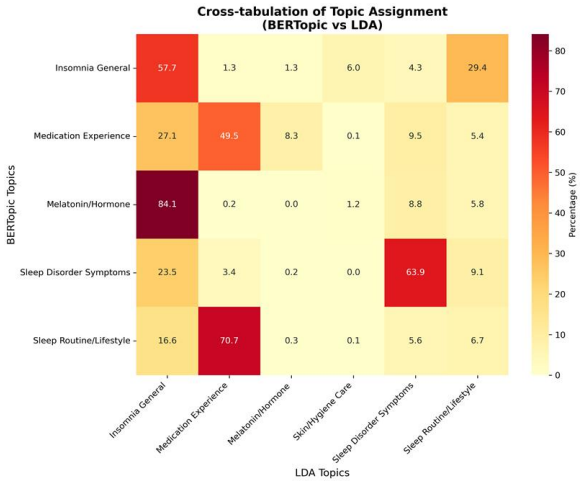

Fig. 3. Cross-tabulation Heatmap of Topic Assignment

분석 결과, 특정 토픽에서 높은 일치율을 보였다. BERTopic의 '약물 경험' 토픽으로 분류된 문서 중 49.5%가 LDA에서도 동일 토픽에 할당되었다. 반면, '멜라토닌/호르몬' 토픽의 경우 84.1%가 LDA의 '불면증 일반' 토픽에 할당되어 가장 높은 집중도를 보였다. 이는 수면제 관련 담론이 두 방법론 모두에서 명확하게 구분되는 주제임을 시사한다.

반면, '수면/불면증 일반(Insomnia General)' 토픽의 경우 일치율이 상대적으로 낮았다. BERTopic에서 '불면증 일반' 토픽으로 분류된 문서 중 LDA에서 동일 토픽에 할당된 비율은 57.7%였으며, 나머지는 '수면루틴/생활습관'(29.4%)이나 '수면장애 증상/치료'(6.0%) 등으로 분산되었다. 이는 불면증 관련 담론이 다양한 하위 주제를 포함하고 있어, 두 방법론이 이를 다르게 세분화했기 때문으로 해석된다.

카이제곱 검정 결과, 두 방법론의 토픽 할당 간에 통계적으로 유의한 연관성이 확인되었다(χ^2=1815.85, p<0.001).

4. Model Performance by Number of Topics

LDA 모델의 최적 토픽 수를 결정하기 위해 토픽 수를 3부터 15까지 변화시키며 일관성 점수(C_V)를 측정하였다. Fig. 4는 토픽 수에 따른 일관성 점수 변화를 나타낸 것이다.

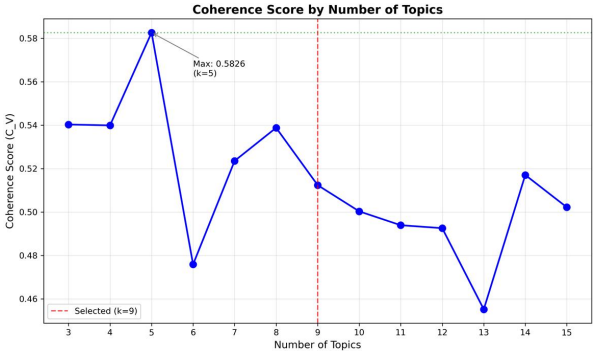

Fig. 4. Coherence Score by Number of Topics

분석 결과, 토픽 수 5에서 일관성 점수가 0.5826으로 최대값을 보였다. 그러나 다음과 같은 이유로 토픽 수 9(일관성 점수 0.5287)를 최종 선정하였다. 첫째, BERTopic과의 공정한 비교를 위함이다. BERTopic은 HDBSCAN의 min_cluster_size=100 설정에 따라 9개의 유효 토픽을 자동 도출하였으므로, LDA도 동일하게 9개로 설정하여 방법론 간 직접 비교가 가능하도록 하였다. 둘째, 토픽 수 5의 경우 수면 건강 담론의 세부 주제(약물 경험, 멜라토닌, 수면장애 증상 등)가 하나의 토픽으로 통합되어 해석의 세분화가 어려웠다. 셋째, 토픽 수 9의 일관성 점수(0.5287)도 일반적으로 양호한 수준(0.5 이상)에 해당한다[24].

BERTopic의 경우 HDBSCAN의 min_cluster_size 파라미터에 따라 토픽 수가 자동으로 결정된다. 본 연구에서는 min_cluster_size=100으로 설정하여 9개의 유효 토픽이 도출되었다. BERTopic은 토픽 수를 사전에 지정하지 않고 데이터의 밀도 구조에 따라 자동으로 결정한다는 점에서 LDA와 차이가 있다.

5. Summary of Methodological Characteristics

본 연구 결과를 바탕으로 BERTopic과 LDA의 방법론적 특성을 Table 4에 요약하였다.

Table 4. Topic Distribution Comparison

Topic Name	BERTopic n (%)	LDA n (%)
Sleep Routine/Lifestyle	892 (11.1)	869 (10.9)
Skin/Hygiene Care	-	120 (1.5)
Insomnia General	1,847 (23.1)	3,061 (38.3)
Medication Experience	1,523 (19.0)	2,040 (25.5)
Melatonin/Hormone	634 (7.9)	195 (2.4)
Sleep Disorder Symptoms	1,281 (16.0)	1,643 (20.6)
Noise (Topic -1)	1,825 (22.8)	74 (0.9)
Total	8,002 (100.0)	8,002 (100.0)

첫째, 접근 방식에서 BERTopic은 사전학습된 언어 모델의 임베딩을 활용하여 문서의 의미적 유사성을 포착하는 반면, LDA는 단어의 동시출현 확률에 기반한 생성 모델이다. 이로 인해 BERTopic은 동의어나 문맥적 의미를 더 효과적으로 처리할 수 있다.

둘째, 토픽 수 결정에서 BERTopic은 클러스터링 알고리즘에 의해 자동으로 토픽 수가 결정되어 탐색적 분석에 유리한 반면, LDA는 연구자가 사전에 토픽 수를 지정해야 하므로 일관성 점수 등을 통한 최적화 과정이 필요하다.

셋째, 재현성 측면에서 LDA는 random seed를 고정하면 동일한 결과를 얻을 수 있으나, BERTopic은 UMAP과 HDBSCAN의 확률적 특성으로 인해 실행 시마다 결과가 다소 달라질 수 있다. 이는 연구의 재현성이 중요한 학술 연구에서 고려해야 할 사항이다.

V. Discussion

선행연구에서는 BERTopic이 LDA보다 우수한 성능을 보이는 것으로 보고되었다. Egger와 Yu[17]는 호텔 리뷰 데이터에서, Abuzayed와 Al-Khalifa[19]는 아랍어 트위터 데이터에서 BERTopic이 더 해석 가능하고 일관된 토픽을 생성함을 확인하였다. 그러나 본 연구에서는 LDA가 낮은 노이즈 비율(0.9%)과 높은 주요 토픽 집중도(85.1%)를 보여 선행연구와 다른 양상을 나타냈다.

이러한 차이는 다음과 같은 요인에 기인한다. 첫째, 데이터 특성의 차이이다. 선행연구들은 주로 영어 텍스트나 짧은 트윗, 리뷰 데이터를 사용한 반면, 본 연구의 네이버 블로그 데이터는 문서당 평균 24.9개의 토큰을 포함하는 상대적으로 긴 텍스트이다. 선행연구에서 LDA는 짧은 텍스트에서 성능이 저하되는 것으로 보고되었으며[6], 이는 역으로 본 연구의 블로그 데이터와 같이 문서 길이가 긴 경우 LDA가 유리할 수 있음을 시사한다. 둘째, 도메인 특수성이다. 수면 건강 담론은 '불면증', '수면제', '멜라토닌' 등 명확한 핵심 키워드를 중심으로 형성되어 있어, 단어 빈도 기반의 LDA가 토픽 구조를 효과적으로 포착할 수 있었다. 셋째, 노이즈 정의 방식의 차이이다. BERTopic은 HDBSCAN을 통해 노이즈를 명시적으로 분리하는 반면, LDA는 모든 문서에 토픽을 할당하므로 두 방법론의 노이즈 비율을 동일 기준으로 비교하기 어렵다.

1. Comparison of Methodological Characteristics

본 연구 결과를 바탕으로 BERTopic과 LDA의 특성을 비교하면 다음과 같다.

첫째, 노이즈 처리 방식에서 근본적인 차이가 있다. BERTopic은 HDBSCAN의 밀도 기반 클러스터링을 통해 어떤 토픽에도 명확히 할당되지 않는 문서를 노이즈(토픽 -1)로 명시적으로 분리한다. 본 연구에서 BERTopic의 노이즈 비율은 22.8%였다. 이는 소셜미디어 텍스트의 특성상 수면 건강과 직접적 관련이 없거나 내용이 모호한 게시물이 상당수 존재함을 반영한다. 반면, LDA는 확률적 생성 모델로서 모든 문서에 토픽 확률 분포를 할당하며, 본 연구에서는 최대 토픽 확률이 0.3 미만인 경우만 노이즈로 분류하였다. 따라서 두 방법론의 노이즈 비율 차이(BERTopic 22.8% vs LDA 0.9%)는 모델 성능의 차이라기보다 노이즈 정의 방식의 근본적 차이에 기인한다. LDA의 낮은 노이즈 비율은 '노이즈가 없다'는 의미가 아니라, 노이즈 성격의 문서들이 기존 토픽에 낮은 확률로 분산 할당되었음을 의미한다. 실제로 LDA의 9개 토픽 중 3개(토픽 1, 5, 8)가 해석이 어려운 노이즈 성격의 키워드를 포함하고 있어, 이를 고려하면 두 방법론의 실질적 노이즈 처리 수준은 유사할 수 있다.

둘째, 토픽 분포의 균등성에서 차이가 있다. BERTopic은 분포 균등성 0.852로 문서가 토픽에 비교적 균등하게 분포되었다. 반면, LDA는 균등성 0.804로 상위 3개 토픽에 전체 문서의 85.1%가 집중되었다. 이는 LDA가 빈도가 높은 단어 조합을 중심으로 토픽을 형성하는 경향이 있어, 일부 우세 토픽에 문서가 편중될 수 있음을 시사한다.

셋째, 키워드 해석성에서 차이가 있다. BERTopic은 BERT 임베딩을 기반으로 문맥적 의미가 유사한 문서들을 군집화하므로, '의존', '부작용'과 같이 특정 맥락에서 사용되는 키워드를 효과적으로 추출하였다. LDA는 단어 동시 출현 빈도에 기반하므로 '먹다', '자다'와 같은 일반적 동사가 포함되어 토픽 해석에 추가적인 노력이 필요하였다.

2. Practical Implications

본 연구 결과를 바탕으로 토픽 모델링 방법론 선택에 대한 실무적 지침을 제시하면 다음과 같다.

첫째, 탐색적 분석이나 노이즈가 많은 소셜미디어 데이터 분석에는 BERTopic을 권장한다. BERTopic은 노이즈 문서를 명시적으로 분리하여 분석자가 유효한 데이터에 집중할 수 있게 하고, 의미론적으로 일관된 토픽을 생성하여 해석이 용이하다.

둘째, 대규모 데이터셋의 효율적 처리나 기존 연구와의 비교가 필요한 경우 LDA를 권장한다. LDA는 계산 효율성이 높고, 토픽 일관성(C_V) 등 표준화된 평가 지표가 잘 정립되어 있어 정량적 비교가 용이하다.

셋째, 명확한 주제(예: 특정 약물명)를 추출하는 경우 두 방법론 모두 유사한 결과를 제공한다. 본 연구에서 '멜라토닌/호르몬' 토픽은 LDA의 '불면증 일반' 토픽과 84.1%의 높은 집중도를 보였다.

넷째, 한국어 건강 텍스트 분석에서는 적절한 형태소 분석과 불용어 처리가 두 방법론 모두에서 중요하다. 특히 BERTopic은 한국어 사전학습 모델의 선택이 성능에 영향을 미치므로, 도메인에 적합한 모델 선정이 필요하다.

3. Limitations and Future Research

본 연구의 한계점으로는 첫째, 단일 플랫폼(네이버 블로그)의 데이터만을 분석하여 일반화에 제한이 있다는 점이다. 트위터, 인스타그램 등 다른 소셜미디어 플랫폼에서는 텍스트 길이와 특성이 다를 수 있어, 방법론별 성능 차이가 달라질 수 있다.

둘째, BERTopic의 토픽 일관성(C_V)을 LDA와 동일한 방식으로 직접 비교하지 못했다. BERTopic은 임베딩 기반 접근법으로 전통적인 C_V 계산 방식과 다른 평가 방법이 필요할 수 있다.

셋째, 토픽의 실제 유용성을 검증하기 위한 전문가 평가가 수행되지 않았다. 향후 연구에서는 수면 건강 전문가를 통한 토픽 품질 평가가 필요하다.

향후 연구에서는 다양한 플랫폼의 데이터를 포함하고, 다국어 데이터에서의 방법론 비교, 그리고 동적 토픽 모델링 기법과의 비교 연구가 수행될 필요가 있다.

VI. Conclusion

본 연구는 한국어 수면 건강 관련 소셜미디어 텍스트를 대상으로 BERTopic과 LDA의 토픽 모델링 성능을 비교 분석하였다. 연구 결과, 두 방법론은 각각의 장단점을 가지고 있음을 확인하였다.

BERTopic은 노이즈 문서를 명시적으로 분리하고, 의미론적으로 일관된 토픽을 생성하며, 균등한 토픽 분포를 보여 탐색적 분석에 적합하다. 반면, LDA는 모든 문서에 토픽을 할당하고, 계산 효율성이 높으며, 일관성 점수와 같은 정량적 평가가 용이하여 대규모 데이터 처리와 비교 연구에 적합하다.

실무적 관점에서, 연구 목적에 따른 방법론 선택 지침을

제시하면 다음과 같다. 첫째, 새로운 데이터셋의 탐색적 분석이나 노이즈가 많은 소셜미디어 데이터 분석에는 BERTopic을 권장한다. 둘째, 대규모 데이터셋의 효율적 처리나 기존 연구와의 비교가 필요한 경우 LDA를 권장한다. 셋째, 명확한 주제(예: 약물 경험)를 추출하는 경우 두 방법론 모두 유사한 결과를 제공한다.

본 연구의 한계점으로는 단일 플랫폼(네이버 블로그)의 데이터만을 분석하여 일반화에 제한이 있다는 점, 그리고 BERTopic의 토픽 일관성(C_V)을 직접 비교하지 못했다는 점이 있다. 향후 연구에서는 다양한 플랫폼의 데이터를 포함하고, 동일한 평가 지표로 두 방법론을 비교하는 연구가 필요하다.

REFERENCES

[1] D. Riemann, C. Baglioni, C. Bassetti, B. Bjorvatn, L. Dolenc Groselj, J. G. Ellis, et al., "European guideline for the diagnosis and treatment of insomnia," Journal of Sleep Research, Vol. 26, No. 6, pp. 675-700, December 2017. DOI: 10.1111/jsr.12594

[2] Health Insurance Review & Assessment Service, "Analysis of Mental Health Disease Treatment Status 2023," HIRA Press Release, 2024.

[3] H. Kim, K. S. Yoo, M. Han, and Y. J. Cho, "Trusting Social Media as a Source of Health Information: Online Surveys Comparing the United States, Korea, and Hong Kong," Journal of Medical Internet Research, Vol. 18, No. 3, e25, March 2016. DOI: 10.2196/jmir.4193

[4] D. M. Blei, "Probabilistic Topic Models," Communications of the ACM, Vol. 55, No. 4, pp. 77-84, April 2012. DOI: 10.1145/2133806.2133826

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol. 3, pp. 993-1022, January 2003.

[6] J. H. Lau, D. Newman, and T. Baldwin, "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality," Proceedings of the 14th Conference of the European Chapter of the ACL, pp. 530-539, Gothenburg, Sweden, April 2014.

[7] M. Grootendorst, "BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure," arXiv preprint arXiv:2203.05794, March 2022.

[8] M. Grootendorst, "BERTopic: Leveraging BERT and c-TF-IDF to Create Easily Interpretable Topics," https://github.com/Maarten Gr/BERTopic, 2022.

[9] T. L. Griffiths, and M. Steyvers, "Finding Scientific Topics," Proceedings of the National Academy of Sciences, Vol. 101, pp. 5228-5235, April 2004. DOI: 10.1073/pnas.0307752101

[10] K. Koh, S. Lee, S. Park, and J. Lee, "Media Reports on COVID-19 Vaccinations: A Study of Topic Modeling in South Korea," Vaccines, Vol. 10, No. 12, 2166, December 2022. DOI: 10.3390/vaccines10122166

[11] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the History of Ideas Using Topic Models," Proceedings of EMNLP, pp. 363-371, Honolulu, USA, October 2008.

[12] X. Han, E. Zhou, and D. Liu, "Electronic Media Use and Sleep Quality: Updated Systematic Review and Meta-Analysis," Journal of Medical Internet Research, Vol. 26, No. 1, e48356, April 2024. DOI: 10.2196/48356

[13] S. Kim, J. Lee, and H. Park, "Understanding Patient Concerns Through Topic Modeling of Online Health Communities," Journal of Medical Internet Research, Vol. 23, No. 8, e28479, August 2021. DOI: 10.2196/28479

[14] T. Park, "COVID-19 Research Trends in Social Work: LDA Topic Modeling Analysis in South Korea," Journal of Social Service Research, Vol. 50, No. 4, pp. 484-499, 2024. DOI: 10.1080/01488376.2024.2354528

[15] N. Reimers, and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proceedings of EMNLP-IJCNLP, pp. 3982-3992, Hong Kong, China, November 2019. DOI: 10.18653/v1/D19-1410

[16] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical Density Based Clustering," Journal of Open Source Software, Vol. 2, No. 11, p. 205, March 2017. DOI: 10.21105/joss.00205

[17] R. Egger, and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," Frontiers in Sociology, Vol. 7, 886498, May 2022. DOI: 10.3389/fsoc.2022.886498

[18] S. Sia, A. Dalmia, and S. J. Mielke, "Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics Too!," Proceedings of EMNLP, pp. 1728-1736, November 2020. DOI: 10.18653/v1/2020.emnlp-main.135

[19] A. Abuzayed, and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," Procedia Computer Science, Vol. 189, pp. 191-194, 2021. DOI: 10.1016/j.procs.2021.05.096

[20] J. Chun, N. R. Han, J. D. Hwang, and J. D. Choi, "Building Universal Dependency Treebanks in Korean," Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), pp. 2194-2202, Miyazaki, Japan, May 2018.

[21] Korea Internet & Security Agency, "2023 Survey on Internet Usage," KISA, pp. 45-52, December 2023.

[22] E. L. Park, and S. Cho, "KoNLPy: Korean Natural Language Processing in Python," Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, pp.

133-136, Chuncheon, Korea, October 2014.

[23] R. Rehurek, and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50, Valletta, Malta, May 2010.

[24] M. Roder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 399-408, Shanghai, China, February 2015. DOI: 10.1145/2684822.2685324

[25] J. Ham, Y. J. Choe, K. Park, I. Choi, and H. Soh, "KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding," in Findings of the Association for Computational Linguistics: EMNLP 2020, Online, November 2020, pp. 422-430. DOI: 10.18653/v1/2020.findings-emnlp.39

Authors

JongHwi Song received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Yonsei University, Korea, in 2012, 2015 and 2023, respectively. Dr. Song joined the Institute of AI Convergence Science at Yonsei University, Wonju, Korea, in 2025. He is currently a Research Professor at the Institute of AI Convergence Science, Yonsei University. He is interested in text mining, large language models (LLMs), and on-device AI.