BERTopic을 활용한 불면증 소셜 데이터 토픽 모델링 및 불면증 경향 문헌 딥러닝 자동분류 모델 구축*

Topic Modeling Insomnia Social Media Corpus using BERTopic and Building Automatic Deep Learning Classification Model

고영수 (Young Soo Ko)**, 이수빈 (Soobin Lee)***
차민정 (Minjung Cha)****, 김성덕 (Seongdeok Kim)*****
이주희 (Juhee Lee)******, 한지영 (Ji Yeong Han)*******
송 민 (Min Song)********

초 록

불면증은 최근 5년 새 환자가 20% 이상 증가하고 있는 현대 사회의 만성적인 질병이다. 수면이 부족할 경우 나타나는 개인 및 사회적 문제가 심각하고 불면증의 유발 요인이 복합적으로 작용하고 있어서 진단 및 치료가 중요한 질환이다. 본 연구는 자유롭게 의견을 표출하는 소셜 미디어 'Reddit'의 불면증 커뮤니티인 'insomnia'를 대상으로 5,699개의 데이터를 수집하였고 이를 국제수면장애분류 ICSD-3 기준과 정신의학과 전문의의 자문을 받은 가이드라인을 바탕으로 불면증 경향 문헌과 비경향 문헌으로 태깅하여 불면증 말뭉치를 구축하였다. 구축된 불면증 말뭉치를 학습데이터로 하여 5개의 딥러닝 언어모델(BERT, RoBERTa, ALBERT, ELECTRA, XLNet)을 훈련시켰고 성능 평가 결과 RoBERTa가 정확도, 정밀도, 재현율, F1점수에서 가장 높은 성능을 보였다. 불면증 소셜 데이터를 심층적으로 분석하기 위해 기존에 많이 사용되었던 LDA의 약점을 보완하며 새롭게 등장한 BERTopic 방법을 사용하여 토픽 모델링을 진행하였다. 계층적 클러스터링 분석 결과 8개의 주제군('부정적 감정', '조언 및 도움과 감사', '불면증 관련 질병', '수면제', '운동 및 식습관', '신체적 특징', '활동적 특징', '환경적 특징')을 확인할 수 있었다. 이용자들은 불면증 커뮤니티에서 부정 감정을 표현하고 도움과 조언을 구하는 모습을 보였다. 또한, 불면증과 관련된 질병들을 언급하고 수면제 사용에 대한 담론을 나누며 운동 및 식습관에 관한 관심을 표현하고 있었다. 발견된 불면증 관련 특징으로는 호흡, 임신, 심장 등의 신체적 특징과 좀비, 수면 경련, 그로기상태 등의 활동적 특징, 햇빛, 담요, 온도, 낮잠 등의 환경적 특징이 확인되었다.

ABSTRACT

Insomnia is a chronic disease in modern society, with the number of new patients increasing by more than 20% in the last 5 years. Insomnia is a serious disease that requires diagnosis and treatment because the individual and social problems that occur when there is a lack of sleep are serious and the triggers of insomnia are complex. This study collected 5,699 data from 'insomnia', a community on 'Reddit', a social media that freely expresses opinions. Based on the International Classification of Sleep Disorders ICSD-3 standard and the guidelines with the help of experts, the insomnia corpus was constructed by tagging them as insomnia tendency documents and non-insomnia tendency documents. Five deep learning language models (BERT, RoBERTa, ALBERT, ELECTRA, XLNet) were trained using the constructed insomnia corpus as training data. As a result of performance evaluation, RoBERTa showed the highest performance with an accuracy of 81.33%. In order to in-depth analysis of insomnia social data, topic modeling was performed using the newly emerged BERTopic method by supplementing the weaknesses of LDA, which is widely used in the past. As a result of the analysis, 8 subject groups ('Negative emotions', 'Advice and help and gratitude', 'Insomnia-related diseases', 'Sleeping pills', 'Exercise and eating habits', 'Physical characteristics', 'Activity characteristics', 'Environmental characteristics') could be confirmed. Users expressed negative emotions and sought help and advice from the Reddit insomnia community. In addition, they mentioned diseases related to insomnia, shared discourse on the use of sleeping pills, and expressed interest in exercise and eating habits. As insomnia-related characteristics, we found physical characteristics such as breathing, pregnancy, and heart, active characteristics such as zombies, hypnic jerk, and groggy, and environmental characteristics such as sunlight, blankets, temperature, and naps.

키워드: 토픽 모델링, 불면증, 소셜 미디어, BERTopic, 딥러닝 모델
topic modeling, insomnia, social media, BERTopic, deep learning model

* 본 연구는 정부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A3A2075114).
** 연세대학교 문헌정보학과 석사과정(kosue@yonsei.ac.kr) (제1저자)
*** 연세대학교 문헌정보학과 박사과정(bini122@yonsei.ac.kr) (공동저자)
**** 연세대학교 소셜오믹스 연구센터(youhear1@hanmail.net) (공동저자)
***** 연세대학교 문헌정보학과 석사과정(ystetjdejr@yonsei.ac.kr) (공동저자)
****** 연세대학교 문헌정보학과 석사과정(juhee5795@yonsei.ac.kr) (공동저자)
******* 연세대학교 문헌정보학과 석사과정(jiyoung181@yonsei.ac.kr) (공동저자)
******** 연세대학교 문헌정보학과 교수(min.song@yonsei.ac.kr) (교신저자)

■ 논문접수일자 : 2022년 5월 13일 ■ 최초심사일자 : 2022년 5월 31일 ■ 게재확정일자 : 2022년 6월 8일
■ 정보관리학회지, 39(2), 111-129, 2022. http://dx.doi.org/10.3743/KOSIM.2022.39.2.111

1. 서 론

불면증은 현대인이 안고 살아가는 만성적인 질병이자 현대인의 특징이다. 국민건강보험공단에 따르면 매해 60만 명 이상이 불면증으로 의료기관을 찾고 있으며 최근 5년 새 진료 환자가 20% 이상 증가하는 등 매해 꾸준하게 증가세를 보이고 있다. 특히, 2020년부터 지속된 코로나19의 확산으로 국민 대다수가 의료기관 방문을 주저하여 환자 수가 감소했음에도 불면증 환자 수는 3.2% 증가의 이례적인 추세를 보였다(국민건강보험공단, 2020).

이처럼 현대 사회에서의 불면증은 날이 갈수록 심각해지고 있다. 불면증 환자는 잠을 잘 수 있는 적절한 환경에서도 잠을 시작하거나 유지하는 데 어려움을 겪으며 깨어 있는 동안 과민성이나 피로와 같은 증상을 필수적으로 동반한다. 불면증의 유병률은 약 6-20%이며, 이 중 50%는 만성적 불면증 상태이다(윤인영, 2013). 제3판 국제수면장애분류(International Classification of Sleep Disorders, version-3, ICSD-3)는 기존의 다양한 불면증 진단을 3가지로 축소하고 이를 포괄하는 불면증의 개념을 제시하고 있다. 불면증은 만성 불면증(chronic insomnia disorder), 단기 불면증(short-term insomnia disorder), 기타 불면증(other insomnia disorder) 3가지로 구분된다(Sateia, 2014). 불면증은 복합적 요인에 의해 발병되며 불면증을 일으키는 대표적인 요인은 생활습관 요인, 환경 요인, 신체 요인, 심리 요인이 있다(서울아산병원, 2014). 불면증은 단순히 성과 지향적인 현대 사회의 특징으로만 설명하기에는 개인 및 사회적인 심각한 문제를 일으킨다. 수면은 뇌를 포함한 우리 몸의 장기가 피로를 해소하기 위해 반드시 필요한 행위이기에 수면이 부족해지면 인지기능이 저하되고 염증이 증가하며 면역기능이 저하되는 등 개인 건강이 악화되고 수면 부족 현상이 계속되면 만성 질환의 위험이 증가할 수 있다(Buysse, 2013). 또한, 사회적으로는 관계가 악화되고 교통사고 유발 등 타인의 안전이 위협받고 업무 능력 저하를 통한 기업 경쟁력의 저하도 우려된다(안경진, 2014). 따라서, 불면증은 단순히 나타나는 현상이 아니라 치료해야 할 질병이며 우리 사회에 만연한 불면증에 대한 인식과 진단은 무엇보다도 중요하다.

대부분의 정신 질환 연구들은 실제 환자를 대상으로 임상 양상 및 치료 효과를 파악하는 방법으로 진행되는데 소셜 미디어가 발달하면서 자유로운 표현이 가능한 소셜 미디어 데이터를 바탕으로 아직 진단을 받지 않은 잠재 환자를 대상으로 하는 정신 질환 연구도 활발하게 이루어지고 있다(이수빈 외, 2021). 소셜 미디어 트위터에 나타난 불면증 연구에서 749개의 트윗을 특징을 분류하고 분석하여 5개의 큰 주제(감정과 정보공유, 불면증 언급, 마음의 표출, 수면의 양과 질, 불면증 관리, 환경 통제)를 찾았으나 데이터양이 부족하고 수동으로 분류했다는 한계가 있고(Jamison-Powell et al., 2012), 불면증 딥러닝 분류 모델 연구는 진행되지 않았다.

따라서, 본 연구는 익명성이 두드러지는 소셜 미디어 중 'Reddit'에서 불면증 관련 자료를 수집 및 분석하고자 하였다. Reddit은 다양한 주제를 다루는 익명의 소셜 네트워킹 플랫폼이며 세계에서 19번째로 많이 방문 되는 웹사이

트이자 미국에서는 7번째로 많이 방문 되는 웹사이트이다. 또한, 일회용 계정을 통해 작성자 본인의 신원을 밝히지 않고 자유롭게 토론할 수 있다는 차별화된 특징이 있다(van der Nagel & Frith, 2015).

본 연구에서는 'Reddit'의 불면증 커뮤니티인 'insomnia'에 있는 데이터를 수집하고 ICSD-3 기준과 전문가의 조언으로 구축된 가이드라인에 따라 불면증 경향 여부를 표기하여 불면증 경향 및 비경향 말뭉치를 구축하고자 하였다. 이 불면증 말뭉치를 훈련데이터로 하여 딥러닝 모델을 학습시켜 불면증 경향 여부를 자동으로 분류할 수 있는 모델을 제안하고자 한다. 또한, 불면증 소셜 미디어 데이터의 세부 주제들을 파악하기 위해 기존에 많이 쓰이는 LDA기법의 단점을 보완한 BERTopic을 활용하여 토픽 모델링을 수행하고 이를 통해 불면증 말뭉치에 공통으로 나타난 질환의 특성과 잠재 환자들의 생각을 파악하고자 한다. 본 연구는 불면증 증상으로 힘들어하는 현대 사회의 수많은 잠재 환자들에게 진단 및 치료의 도움을 줄 수 있다는 점에서 의의가 있다.

본 연구의 연구 질문은 다음과 같다.

- 연구 질문 1: BERTopic 토픽 모델링을 통해 살펴본 불면증 소셜 미디어 데이터의 세부 주제는 무엇인가?
- 연구 질문 2: 불면증 말뭉치를 학습시켜 생성한 불면증 경향 문헌 자동분류 딥러닝 모델의 성능은 어떠한가?

2. 이론적 배경

2.1 소셜 미디어를 활용한 정신건강 연구

소셜 미디어는 다양한 정신건강 문제를 연구하기 위해 광범위하게 활용되고 있다. 특히, 기계 학습 및 인공지능을 적용하여 텍스트의 중요한 패턴을 찾는 텍스트 마이닝 분야에서 소셜 미디어 데이터는 효과적으로 사용되고 있다. He et al.(2017)은 자연어 사용과 텍스트 마이닝 접근 방법을 사용해 외상 후 스트레스 장애(Post-Traumatic Stress Disorder; PTSD)를 선별하는 자동화된 평가 시스템을 제시하였고 웨이보(Weibo)에서 수집한 소셜 미디어 텍스트를 바탕으로 이용자가 자살 위험에 처해있거나 심리적 고통을 경험하고 있는지 자동으로 분류한 사례도 존재한다(Cheng et al., 2017).

최근 신종 코로나바이러스 감염증(코로나19)로 인해 발생하는 심리적 어려움을 파악하는데에도 소셜 미디어 데이터가 유용하게 사용되었다. Koh와 Liew(2020)는 코로나19의 장기화 상황에서 많은 사람이 사회적 단절로 느끼는 외로움 정도를 트위터(twitter) 바탕으로 조사하여 외로움의 주요 영역을 구분한 연구를 진행하였다. 또한, 코로나바이러스 대 유행으로 심각해지고 있는 자살문제와 관련하여 소셜 미디어인 네이버 지식iN에 나타난 자살 관련 문헌을 수집하고 자살 경향 문헌 자동분류 모델 구축 및 토픽모델링을 통한 자살 관련 세부 요인을 파악한 연구도 진행되었다(고영수, 이주희, 송민, 2021).

본 연구는 불면증 연구를 위해 소셜 미디어인 'Reddit'의 불면증 관련 커뮤니티 'insomnia'

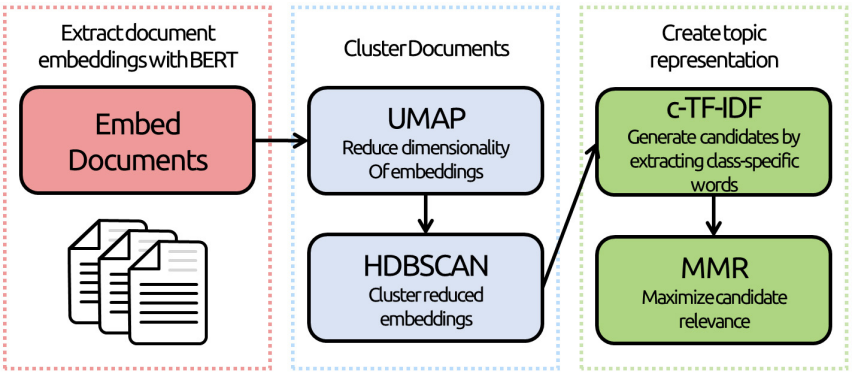
에서 데이터를 수집하고 불면증 경향 문헌인지 아닌지를 자동분류하는 딥러닝 모델을 구축하는 연구와 토픽 모델링을 통해 불면증 소셜 데이터의 세부 주제들을 확인하는 연구를 진행하고자 한다.

2.2 BERTopic 토픽 모델링

자연어처리 분야에서 대용량 텍스트의 주제를 확인하는 토픽 모델링은 중요한 연구 주제이다. 그동안 토픽 모델링으로 잠재 디리클레 할당(Latent Dirichlet Allocation) 및 확률적 잠재 의미 분석(Probabilistic Latent Semantic Analysis) 방법이 많이 사용됐으나 이 방법들은 최적의 결과를 얻기 위해 주제의 수, 불용어 목록, 형태소 분석이 필요하고 단어 간의 의미 관계가 무시된다는 약점을 가지고 있다. 이를 해결하기 위해 공동 문서 및 단어 의미론적 임베딩을 활용하는 Top2vec 방법이 제안되었고(Angelov, 2020), 사전 훈련된 단어 임베딩을 클러스터링하는 동시에 가중치 클러스터링 및 상위 단어 순위 재지정을 위한 문서 정보를 통합하는 연구도 진행되었다(Sia, Dlamia, & Mielke, 2020).

하지만, 이러한 토픽 모델링 기술은 클러스터의 중심에 가까운 단어를 가장 잘 대표하는 주제인 것으로 가정하는데, 실제로는 클러스터가 클러스터 중심 주위의 구 안에 항상 있는 것은 아니므로 모든 문서 클러스터와 해당 클러스터의 표현에 대한 가정이 성립되지 않아 해당 주제가 오해의 소지가 있을 수 있다(Grootendorst, 2022).

이에 클러스터링 기술과 TF-IDF의 클래스 기반 변형을 활용하여 일관된 주제 표현을 생성하는 BERTopic 기법이 제안되었다(Grootendorst, 2020). BERTopic은 텍스트의 임베딩 단계에서 BERT를 활용한 임베딩과 c-TF-IDF 단어 가중치를 활용한 다음, 각 도메인에 맞는 텍스트 클러스터링을 하여 텍스트에 잠재된 의미 있는 주제를 찾아내는 토픽 모델링 기법이다. 〈그림 1〉과 같이 사전 훈련된 SBERT를 사용하여 텍스트 데이터를 임베딩 표현으로 변환하고 UMAP을 사용하여 임베딩의 차원을 줄여 클러스터링 프로세스를 최적화한다. 또한, HDBSCAN으로 축소된 임베딩을 클러스터링하여 의미적으로 유



〈그림 1〉 BERTopic 알고리즘

사한 클러스터를 생성한다. 마지막으로 문서 클러스터에서 c-TF-IDF의 사용자 정의 클래스 기반 변형을 사용하여 주제 표현을 추출한다. BERTopic은 다른 토픽 모델링 기법과 비교했을 때 높은 주제 일관성(Topic Coherence)과 주제 다양성(Topic Diversity)을 보이는 것으로 확인되었다(Grootendorst, 2022).

이와 관련하여 인도네시아 전자 상거래 챗봇 데이터로 토픽 모델링을 진행하여 BERTopic 기법이 LDA보다 더 나은 주제 일관성, 다양성 평가 점수를 보인 연구가 수행되었고(Hendry et al., 2021), BERTopic을 LDA, NMF와 비교하여 NPMI(Normalized Pointwise Mutual Information) 방식으로 측정 평가한 연구에서도 BERTopic 기법이 더 나은 결과를 보여주었다(Abuzayed & Al-Khalifa, 2021). 본 연구에서는 높은 평가를 받았지만 새로운 기법으로 아직 관련 연구가 많이 진행되지 않은 BERTopic으로 토픽 모델링을 실행하여 불면증 데이터의 세부 주제를 탐색하고자 한다.

2.3 텍스트 분류를 위한 딥러닝 언어모델

최근 텍스트 마이닝 분야에서 정신 질환 관련 딥러닝 분류 모델 연구가 주목받고 있으며 BERT(Bidirectional Encoder Representations from Transformers), ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately) 등 다양한 언어모델을 사용한 연구가 진행되고 있다.

Nikhil, Sreekumar, Subha(2021)은 정신분열증 탐지를 위하여 환자 및 대조군을 대상으로 LSTM(Hochreiter & Schmidhuber, 1997) 모델을 사용하여 높은 정확도를 이끌어냈다. Martinez-Castaño et al.(2021)은 자해 징후 및 우울증 징후 탐지를 위해 BERT기반 분류 모델을 사용하였고 자해 징후 탐지에서 최대 91%의 정확도를 보였으며, Guo et al.(2021)은 BERT모델을 사용하여 지식그래프 기반의 정신건강 Q&A 시스템을 구축하였다. 본 연구는 불면증 경향 문헌 및 비경향 문헌 분류를 위해 최신 언어모델에 대한 성능실험을 수행하고자 다양한 자연어처리 영역에서 우수한 성능을 보인 5가지 언어모델(BERT, RoBERTa, ALBERT, ELECTRA, XLNet)을 선정하여 연구를 진행하고자 한다.

BERT는 Transformer(Vaswani et al., 2017)의 양방향 인코더를 심층적으로 쌓은 구조의 언어모델이다(Devlin et al., 2018). BERT는 사전학습의 목적 함수로 MLM(Masked Language Model)와 NSP(Next Sentence Prediction)를 사용한다. MLM은 입력된 문장의 토큰 중 일부를 [MASK] 토큰으로 변환하고 주변의 문맥 정보를 활용하여 모델을 통과한 이후 변환되기 전의 토큰을 예측하도록 하는 방법이며, NSP는 두 문장 주어진 상황에서 두 번째 문장이 첫 번째 문장의 다음에 해당하는지를 예측하는 방법이다.

RoBERTa(Robustly optimized BERT approach)는 기존 BERT모델에 충분히 사전학습시키는 최적화가 필요하다는 관점에서 제안되었다(Liu et al., 2019). 기존의 목적 함수였던 NSP를 제거하면서 MLM방식을 정적인 방식에서 동적인 방식으로 대체하였고, BERT의 훈련데이터보다 더 큰 용량의 다양한 데이터를 더 오랜 기간, 더 큰 배치 사이즈로 학습했

다는 점에서 차별점을 지닌다.

ALBERT(A Lite BERT)는 대용량의 말뭉치로 사전학습하는 거대 규모 언어모델의 문제점에 주목하였다(Lan et al., 2019). 저자는 언어모델의 크기가 커질수록 메모리 부족 현상이 나타나고 학습에 너무 과도한 시간이 소요되며, 파라미터의 증가가 성능을 저하하는 memory degradation 현상이 발생한다는 점에서 해결이 필요하다고 보았다. ALBERT는 기존 BERT의 토큰 임베딩 층 크기를 대폭 감소시키고 서로 다른 층들이 파라미터를 공유하도록 설계하였다. 또한, 언어 이해력 증진에 비효율적인 NSP를 제거하고 문장 순서를 예측하는 SOP(Sentence-Order Prediction) 방법을 새로운 목적 함수로 도입하였다.

ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)는 생성모델(Generator)과 판별모델(Discriminator)을 결합한 언어모델이다(Clark et al., 2020). 생성모델은 BERT와 동일하게 MLM 방식으로 학습이 진행되나, 판별모델의 학습에는 RTD(Replaced Token Detection)라는 새로운 방식을 적용하였다. 생성모델에서는 일부 토큰을 [MASK] 처리한 뒤 토큰의 원형을 예측하고, 판별모델은 생성된 문장의 각 토큰이 원본 문장의 토큰과 같은지 혹은 변경되었는지를 예측하는 원리이다.

XLNet(Generalized Autoregressive Pretraining for Language)는 OpenAI-GPT와 같은 Auto-Regressive(AR) 방식과 BERT와 같은 Auto-Encoder(AE) 방식의 장점을 유지하고 단점을 보완하기 위해 설계되었다(Yang et al., 2019). 이 모델은 사전훈련 단계에서 PLM(Permutation Language Modeling)이라는 새로운 목적 함수를 제안한다. PLM는 모든 입력 시퀀스 순서를 조합한 AR 방식의 학습을 통해 양방향 정보를 모두 고려하고 토큰 간의 의존성(dependency) 학습을 가능하게 했다. 한편, PLM 방식에서는 토큰의 위치 정보를 명확하게 식별할 수 없다는 문제가 발생하는데, 이를 식별하는 방법인 Target Position-Aware Representation과 Two-Stream Self-Attention 구조를 제안하여 해결하였다.

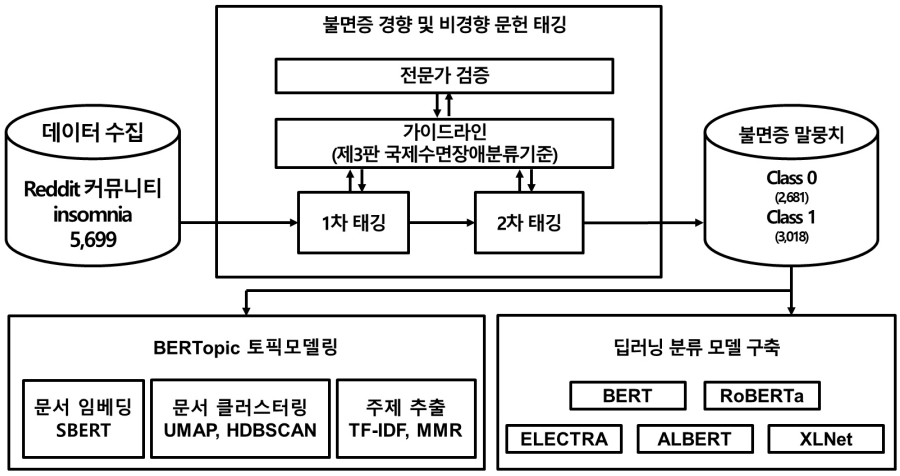
본 연구는 앞서 언급된 5가지 모델로 불면증 관련 문헌과 비관련 문헌을 분류하는 자동분류 딥러닝 모델을 구축하고 각 모델의 성능을 비교하여 최적의 모델을 찾고자 한다.

3. 연구 설계

본 연구는 불면증 관련 소셜 데이터를 수집하여 불면증 말뭉치를 구축하고 토픽 모델링 기법을 활용해 이를 심층적으로 분석하였다. 또한, 딥러닝 모델을 활용하여 자동분류 모델을 만드는 연구를 진행하였고 연구 모형은 〈그림 2〉와 같다.

3.1 불면증 말뭉치 구축

연구를 위해 소셜 미디어 Reddit에서 2020년 1월부터 2021년 5월까지 5,699개의 데이터를 수집하였고 불면증 관련 문헌과 비관련 문헌으로 분류하는 태깅을 진행하였다. 태깅을 위해 제3판 국제수면장애분류기준(ICSD-3)을 기반으로 정신의학과 전문의 1명의 검증을 받아



〈그림 2〉 불면증 소셜 데이터 토픽 모델링 분석 및 딥러닝 모델 구축 연구 모형

〈표 1〉 ICSD-3 국제수면장애분류 내 불면증 판단 기준

A. 환자는 다음 중 하나 이상을 보고하거나 환자의 부모 또는 간병인이 다음 중 하나 이상을 관찰함.
A-1. 수면개시의 어려움
A-2. 수면을 유지하기 어려움
A-3. 원하는 것보다 일찍 깨어남
A-4. 적절한 시간에 잠자리에 들지 않으려고 함
A-5. 부모 또는 간병인 개입없이 자기가 어려움
B. 야간 불면증과 관련된 다음 사항 중 하나 이상을, 환자가 자기보고하거나 환자 부모 또는 간병인이 관찰함.
B-1. 피로/권태
B-2. 주의, 집중 또는 기억장애
B-3. 사회, 가족, 직업 또는 학업 수행 장애
B-4. 기분장애/과민성
B-5. 주간졸림
B-6. 행동문제(예: 과잉행동, 충동성, 공격성)
B-7. 동기/에너지/기획력 감소
B-8. 오류/사고 발생 경향
B-9. 수면에 대한 우려 또는 불만족

작성된 가이드라인을 사용하여 5개의 기준 A, B, C, D, E 중 주요 기준인 A와 B를 모두 만족시킬 경우 불면증 관련 문헌으로 분류하였다. 관련된 기준은 〈표 1〉과 같다. 문헌 분류는 기존에 말뭉치 구축의 경험이 있는 연구원 5명이 직접 진행하였다. 총 2차례의 교차 검증과정을 거쳐 진행되었고 불일치 하는 문헌이 있을 때는 전문의의 검증을 받고 연구원들 간의 합의를 통해 결정하였다. 최종적으로 2,681개의 불면증 비경향 문헌과 3,018개의 불면증 경향 문

헌으로 구성된 불면증 말뭉치를 구축하였다.

3.2 Bertopic 토픽모델링

불면증 말뭉치의 구조와 세부 주제를 살펴보기 위해 기존에 많이 사용되는 LDA 방식의 단점을 보완한 BERTopic을 활용하고자 하였다. 이를 위해 Github에 올려져 있는 BERTopic[1] 패키지를 다운받고 파이썬 프로그램을 통해 분석을 진행하였다. 먼저, 만들어진 불면증 말뭉치를 문장 단위로 분할한 후 SBERT를 통해 임베딩하는 과정을 거쳤다. 다음으로 UMAP을 통해 임베딩 차원이 축소되고 HDBSCAN을 통해 클러스터링이 진행되었다. 마지막으로, c-TF-IDF를 활용하여 적합한 주제가 추출되었다. 불용어 제거가 필요하지는 않지만 더 정교한 분석을 위해서 불용어를 제거하고자 vectorizer_model의 CountVectorizer를 사용하였고 영어에서 많이 사용되는 불용어들을 제거하였다. 이를 통해 59개의 주제가 생성되었고 이를 계층적 클러스터링을 진행하여 총 8개의 대주제로 구분하였다.

3.3 불면증 분류 딥러닝 모델 비교 실험

본 연구에서는 최적화된 딥러닝 모델을 사용하기 위해 5가지의 언어모델을 비교 실험하였다. 먼저, 언어모델들이 대용량의 말뭉치를 통해 사전학습 된 버전을 Hugging Face library[2])를 통해 다운받았고 언어모델 버전은 GPU 메모리 한계를 고려하여 〈표 2〉를 사용하였다. 사전 훈련된 언어모델을 로딩한 후, 사전에 구축된 불면증 말뭉치를 훈련데이터로 사용하여 파인튜닝을 진행하였다. 파인튜닝 단계에서는 기존 언어모델의 구조를 크게 변경하지 않고 가장 상단에 linear layer와 softmax를 추가하였다. 그리고 입력 시퀀스의 [CLS] 토큰에 대응되는 마지막 hidden state h는 입력 시퀀스에 대한 표상으로 간주하며, 벡터 h는 linear layer 및 softmax를 거쳐 클래스에 대한 확률값을 계산하였다. 비용 함수(loss function)로는 cross-entropy를 사용하였다.

〈표 2〉 불면증 딥러닝 분류 언어 모델 버전

Model	Version
BERT	bert-base
RoBERTa	roberta-base
ELECTRA	electra-small-discriminator
ALBERT	albert-base-v2
XLNet	xlnet-base

파인튜닝시 설정한 주요 하이퍼 파라미터값의 범위는 다음과 같다. 훈련 반복 횟수(epoch)는 최대 10회까지로 하고, 검증 데이터셋에 가장 높은 성능을 보인 시점의 모델을 평가 데이터셋에 적용하였다. 배치 사이즈(batch size)는 메모리 가용 범위 내에서 8, 16, 32로 하였고 옵티마이저(optimizer)는 Adam(Kingma & Ba, 2015)을 사용하였다. 학습률(learning rate)은 5e-5, 3e-5, 2e-5로 설정하고, 선형 스케줄러를 통해 점진적으로 조절하였다. 이때 warm-up ratio는 0.1으로 설정하였다. 다른 상세한 하이퍼 파라미터는 개별 언어모델들의 저자가 사용한 값을 그대로 사용하였다.

1) https://maartengr.github.io/BERTopic/
2) https://huggingface.co/docs/hub/libraries

4. 결 과

4.1 Bertopic 토픽모델링 결과

4.1.1 문장 개수 상위 10개 주제

BERTopic 토픽 모델링 결과 29,886개의 문장에서 59개의 주제가 생성되었고 문장 개수가 높은 상위 10개의 주제와 주요 단어는 〈표 3〉과 같다. 전체 문장의 약 90%에 해당하는 주제 0과 주제 -1은 단순히 불면증을 언급하는 내용으로 특별한 의미를 담고 있지 않다. 그 외 약 10%에 해당하는 문장에서 주요 의미들을 찾을 수 있음을 알 수 있다. 의미를 담은 주제 중 가장 많은 문장을 가지고 있는 주제 1은 꿈과 관련이 있다. 'dream', 'nightmares', 'vivid' 등의 단어가 등장하는데 악몽과 자각몽(vivid dream)이 주요 내용임을 확인할 수 있다. 다음으로 많이 나오는 주제 2는 'meditation', 'tried', 'exercise' 등으로 보아 명상과 관련된 주제임을 알 수 있다. 다음으로 많이 나오는 주제 3은 'seroquel', 'prescribed', 'dose' 등의 단어가 등장하는데 수면제 'seroquel' 관련임을 알 수 있다. 수면제 관련 주제는 5, 6에도 등장하여 자주 등장하는 주제임을 확인할 수 있다. 그 밖에 심계항진, 심장박동 관련 주제(4)와 감사와 격려 관련 주제(7), 코로나 자가격리와 관련된 주제(8), 수면위생 관련 주제(9)가 상위 빈도를 차지하였다.

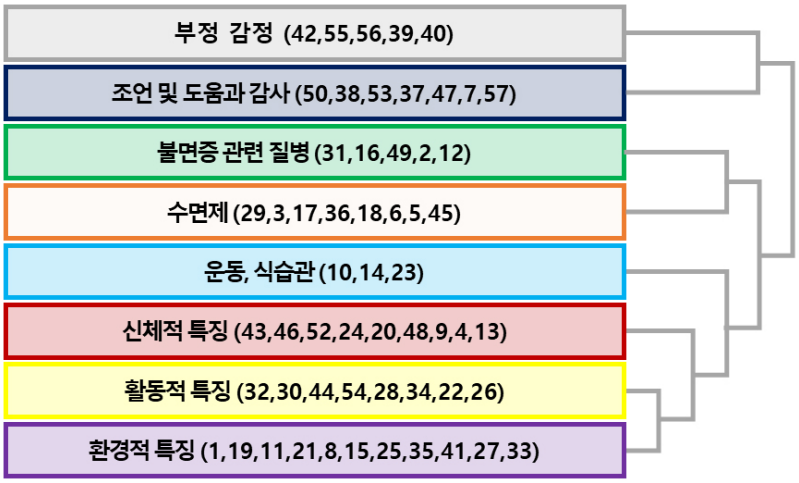
4.1.2 계층적 클러스터링

59개의 주제를 계층적 클러스터링(Hierarchical Clustering)으로 확인한 결과는 〈그림 3〉과 같다. 계층화된 그래프를 바탕으로 의미가 비슷한 주제를 그룹화하여 총 8개의 대주제로 묶었으며, 의미를 지니지 않은 주제 51, -1, 0은 그룹에서 제외하였다. 총 8개의 주제는 각각 '부정 감정', '조언 및 도움과 감사', '불면증 관련 질병', '수면제', '운동 및 식습관', '신체적 특징', '활동적 특징', '환경적 특징'으로 확인할 수 있다.

첫번째 주제는 '부정적 감정'에 대한 주제이며 대표 단어와 대표 문장은 〈표 4〉와 같다. 단어와 문장을 살펴보면 기분이 싫고 이상하거나 비속어 같은 단어들이 주로 나오고 있다. 커뮤

〈표 3〉 BERTopic 토픽모델링 문장 개수 상위 10개 주제

번호	주제	문장 개수	비율	단어1	단어2	단어3	단어4
0	0	15169	50.8%	insomnia	sleep	I'm	I've
1	-1	11735	39.3%	sleep	I'm	asleep	like
2	1	309	1.0%	dreams	dream	nightmares	vivid
3	2	255	0.9%	meditation	tried	exercise	hypnosis
4	3	178	0.6%	seroquel	prescribed	dose	mg
5	4	131	0.4%	heart	rate	palpitations	chest
6	5	128	0.4%	trazodone	trazadone	prescribed	mg
7	6	109	0.4%	mg	5mg	dose	50mg
8	7	92	0.3%	thanks	thank	cheers	thx
9	8	91	0.3%	quarantine	schedule	quarantined	started
10	9	86	0.3%	hygiene	sleep	good	tried



〈그림 3〉 BERTopic 계층적 클러스터링 결과(괄호 내 숫자는 주제 번호)

〈표 4〉 '부정적 감정' 대표 단어와 대표 문장

대주제	주제	단어1	TF -IDF	단어2	TF -IDF	단어3	TF -IDF
부정 감정	42	strange	0.76	weird	0.61	weirdo	0.20
	55	hey	1.10	guys	0.74	folkskind	0.29
	56	yetand	2.10	moment	0.96	thats	0.62
	39	sucks	2.25	fucking	0.22	blows	0.20
	40	hate	2.19	fucking	0.17	ugh	0.16
	주제별 대표 문장						
	42	"Its the strangest thing."					
	55	"Hey folks,Kind of an odd one here."					
	56	"Never in my life."					
	39	"Absolutely sucks"					
	40	"I hate this!"					

니티 이용자들이 글을 통해 불면증과 관련된 부정적 감정을 가감 없이 표출하고 있으며 비속어를 사용할 정도의 부정적인 상황으로 유추할 수 있다.

두번째 주제는 조언, 도움을 구하거나 그에 대한 감사와 관련된 주제이다. 대표 단어와 대표 문장은 〈표 5〉와 같다. 주로 제안이나, 아이디어 및 도움, 조언을 구하는 내용과 공감 및 감사 관련 내용이 나오고 있다. 이를 통해 이용자들은 reddit의 insomnia 커뮤니티를 통해 불면증과 관련된 도움이나 조언을 구하고 있음을 알 수 있다.

세번째 주제는 불면증과 관련된 질병이며 〈표 6〉을 통해 주요 내용을 확인할 수 있다. 대표 문장과 단어를 통해 간질 발작(epileptic seizure), 양극성 장애(bipolar disorder), 주의력

〈표 5〉 '조언, 도움에 대한 감사' 대표 단어와 대표 문장

대주제	주제	단어1	TF-IDF	단어2	TF-IDF	단어3	TF-IDF
조언, 도움과 감사	50	ideas	2.07	suggestions	1.66	idk	0.24
	38	help	1.31	btw	0.19	thank	0.14
	53	advice	1.59	advices	0.81	redditors	0.38
	37	advance	1.85	thanks	0.80	thank	0.58
	47	friends	1.80	thanks	1.63		0.00
	7	thanks	1.63	thank	1.24	cheers	0.14
	57	hey	2.06	thanks	0.81	symphathize	0.70
	주제별 대표 문장						
	50	"Any suggestions?"					
	38	"Someone please please help me."					
	53	"Any advice?"					
	37	"Thanks in advance"					
	47	"Thanks for hearing me out, I appreciate any responses."					
	7	"Thank you!"					
	57	"Hey all, I symphathize with all of you"					

〈표 6〉 '불면증 관련 질병' 대표 단어와 대표 문장

대주제	주제	단어1	TF-IDF	단어2	TF-IDF	단어3	TF-IDF
불면증 관련 질병	31	seizure	0.26	seizures	0.13	focal	0.09
	16	bipolar	0.27	manic	0.14	mania	0.10
	49	adhd	0.68	aspergers	0.16	autism	0.16
	2	meditation	0.12	tried	0.04	exercise	0.03
	12	hallucinations	0.26	hallucinating	0.09	auditory	0.08
	주제별 대표 문장						
	31	"A week alter walking the dogs I get an aura for an epileptic seizure"					
	16	"One of the practitioners asked me if I had ever heard of bipolar disorder"					
	49	"With ADHD, I have access to some, but I feel modafinil would be more applicable"					
	2	"But you have to make sure you are susceptible to hypnosis"					
	12	"They were mostly visual but had a few later that were auditory. What do I do now? I can't cope with insomnia and hallucinations!"					

결핍 과잉행동장애(ADHD), 아스퍼거 증후군(aspergers), 자폐증(autism), 환각(hallucination), 환청(auditory) 등의 내용이 포함된 것을 알 수 있다. 특이한 점으로 명상(meditation)이 이 주제로 함께 포함되었는데 이는 명상과 환각이 비슷한 의미를 지녔기 때문으로 추측된다.

네번째 주제는 불면증과 관련된 질병이다. 〈표 7〉을 통해 seroquel, benzos, amitriptyline, trazodone, temazepam 등의 수면제가 주로 등장하고 비타민(vitamin) 같은 보충제도 함께

〈표 7〉 '수면제' 대표 단어와 대표 문장

대주제	주제	단어1	TF -IDF	단어2	TF -IDF	단어3	TF -IDF
수면제	29	migraines	0.35	migraine	0.29	september	0.06
	3	seroquel	0.24	prescribed	0.02	dose	0.02
	17	benzos	0.32	benzo	0.14	addicted	0.03
	36	amitriptyline	0.46	pain	0.06	prescribed	0.06
	18	supplements	0.19	vitamin	0.15	supplement	0.10
	6	mg	0.18	5mg	0.06	dose	0.05
	5	trazodone	0.23	trazadone	0.13	prescribed	0.04
	45	temazepam	0.35	gave	0.05	15mg	0.04
	주제별 대표 문장						
	29	"I now get migraines and I return to school in 2 weeks and I'm afraid"					
	3	"Wish I was less worried about my weight so I could try seroquel."					
	17	"I have yet to try a Benzo, but have obvious reservations.PleAse help"					
	36	"Last night I began the Amitriptyline and got about 3 hours of sleep"					
	18	"You doctor shop and try supplements and they might work for a little while"					
	6	"He increased it to 2 mg"					
	5	"Been on trazodone for years, unfortunately"					
	45	"Right now I take temazepam, but it's starting to feel like my body is just adjusting to it and it's less effective than it used to be"					

등장하는 것을 알 수 있다. 특이점으로 편두통(migraine)이 등장하는데 이 증상이 나타날 때 수면제를 함께 복용하고 있어서 함께 묶인 것으로 추측된다.

다섯번째 주제는 운동, 식습관 관련 내용이다. 〈표 8〉을 통해 운동을 통해 불면증을 개선하려는 의지와 불면증이 식습관에 영향을 미치는 내용들이 포함되어 있음을 알 수 있다.

여섯번째 주제는 신체적 특징 관련 내용이다. 〈표 9〉를 살펴보면 주제 43과 46은 임신과

〈표 8〉 '운동 및 식습관' 대표 단어와 대표 문장

대주제	주제	단어1	TF -IDF	단어2	TF -IDF	단어3	TF -IDF
운동 및 식습관	10	exercise	0.18	miles	0.06	weights	0.05
	14	eat	0.11	hungry	0.09	eating	0.09
	23	weight	0.18	pounds	0.12	lost	0.11
	주제별 대표 문장						
	10	"I do daily exercise classes."					
	14	"I'm getting fatter and fatter because my sleep is terrible so I crave junk food but then when I try to eat healthy my sleep is even worse at night etc."					
	23	"I had no idea what I was doing at the time but I started with eating 1200 calories and exercising like crazy, the weight dropped like crazy and I thought this was normal."					

〈표 9〉 '신체적 특징' 대표 단어와 대표 문장

대주제	주제	단어1	TF-IDF	단어2	TF-IDF	단어3	TF-IDF
신체적 특징	43	unisom	0.48	otc	0.07	pregnancy	0.06
	46	hormonal	0.18	menopause	0.15	hormones	0.14
	52	snores	0.45	snoring	0.17	roommate	0.11
	24	breathing	0.24	breath	0.24	shortness	0.13
	20	apnea	0.27	study	0.05	snore	0.04
	48	nose	0.39	nasal	0.17	plastic	0.13
	9	hygiene	0.21	sleep	0.03	good	0.03
	4	heart	0.25	rate	0.12	palpitations	0.08
	13	music	0.13	listen	0.09	podcasts	0.09
	주제별 대표 문장						
	43	"I managed it by taking Unisom, which is safe to take during pregnancy"					
	46	"I kept telling doctors that I thought the problem might be hormonal but kept getting shooed off."					
	52	"He also snored, which is something I can never sleep through"					
	24	"you notice how unusually shallow your breaths are"					
	20	"I've had a sleep apnea test and I've had bloodwork done"					
	48	"just makes my nose stuffy!"					
	9	"I've also done sleep hygiene and everything according to the rules"					
	4	"There are several things that I have noticed speed up my heart rate"					
	13	"I listened to sleep music for an hour, still up."					

관련된 수면제(OTC)와 갱년기 호르몬 문제를 언급하고 있다. 주제 52, 24, 20, 48은 호흡계통 관련으로 코골이, 수면 무호흡 관련 내용이 포함되어 있다. 주제 9는 수면 위생과 관련된 내용인데 위생이라는 단어가 신체적 특징과 연결되기 때문에 함께 묶인 것으로 보인다. 주제 4는 심장박동이나 심계항진 관련 단어로 보아 심장과 관련된 것으로 보인다. 주제 13의 음악 관련 주제가 이 그룹에 속하게 되었는데 'listen'이라는 단어가 신체적 특징으로 연결된 것으로 추측된다.

일곱 번째 주제는 활동적 특징과 관련이 있다. 〈표 10〉에 보면 주제 32, 34를 통해 좀비와 그로기 상태를 나타내는 내용이 보이고 주제 54의 잠깐 잠을 뜻하는 wink 와 주제 28의 수면 경련(hypnic jerk)이 나타난다. 다만, 주제 30의 lockdown은 도시의 폐쇄를 뜻하고 있는데 이는 lockdown 자체가 활동적인 의미를 담고 있어서 이 그룹으로 묶인 것으로 추측된다. 마찬가지로 주제 44의 단어 'remeron'은 항우울제로 쓰이는 약품 이름인데 단어 'stopped'와 묶이면서 활동적 특징으로 그룹화된 것으로 보인다. 주제 22의 SFI(Sporadic Fatal Insomnia)와 FFI(Fatal Familial Insomnia)는 유전적 불면증을 뜻하는 단어인데 scared와 묶이면서 이 그룹에 포함된 것으로 보인다.

여덟 번째 주제는 환경적 특징으로 보인다. 〈표 11〉에 나타난 단어와 문장을 살펴보면 주

〈표 10〉 '활동적 특징' 대표 단어와 대표 문장

대주제	주제	단어1	TF -IDF	단어2	TF -IDF	단어3	TF -IDF
활동적 특징	32	zombie	0.43	mode	0.05	like	0.05
	30	lockdown	0.36	lock	0.08	uk	0.05
	44	remeron	0.53	stopped	0.07	sleepint	0.05
	54	wink	0.69	didnt	0.20	night	0.09
	28	hypnic	0.42	jerks	0.33	jerk	0.31
	34	groggy	0.52	little	0.06	grogginess	0.06
	22	sfi	0.38	ffi	0.07	scared	0.05
	26	vent	0.53	venting	0.22	needed	0.14
	주제별 대표 문장						
	32	"I'm afraid I will take them not sleep and be a zombie"					
	30	"My states super locked down and I don't have a PC anyway but I can't do this sleepless night shit anymore."					
	44	"Usually I just take Remeron, but lately that hasn't been working"					
	54	"I couldn't sleep a wink and it was definitely related to anxiety"					
	28	"I know, hypnic jerk"					
	34	"4 hours at most and I have a good day more then 4 and I'm groggy and can't function"					
	22	"I say this cuz I hear stories about sfi in other parts of the world"					
	26	"Just a little venting"					

〈표 11〉 '환경적 특징' 대표 단어와 대표 문장

대주제	주제	단어1	TF -IDF	단어2	TF -IDF	단어3	TF -IDF
환경적 특징	1	dreams	0.12	dream	0.10	nightmares	0.10
	19	alarm	0.33	alarms	0.14	set	0.11
	11	nap	0.24	naps	0.24	napping	0.05
	21	pee	0.29	bladder	0.10	urinate	0.08
	8	quarantine	0.29	schedule	0.05	quarantined	0.04
	15	covid	0.16	pandemic	0.13	virus	0.09
	25	sun	0.40	comes	0.09	rise	0.08
	35	blanket	0.35	weighted	0.29	blankets	0.17
	41	room	0.23	degrees	0.14	temp	0.13
	27	driving	0.20	wheel	0.11	drive	0.11
	33	toss	0.39	turn	0.28	tossing	0.26
	주제별 대표 문장						
	1	"These are extremely vivid and completely lucid"					
	19	"I also don't get out of bed right away when the alarm rings"					
	11	"Naps don't seem to work and I'm fed up"					
	21	"Also my bladder wakes me up a lot, so that's fun too"					
	8	"So the past few weeks I've been trying to fix my sleep schedule because it went a little haywire when quarantine began"					
	15	"Like so many others, this whole pandemic thing has me spiraling"					
	25	"I really, really want to be awake when the sun rises"					
	35	"'Add a blanket for weight"					
	41	"The temperature in my room is 65F (said to be the perfect temperature for sleeping)"					
	27	"'I fell asleep 10+ times while driving"					
	33	"I see people here saying they toss and turn but I don't have that problem"					

제 1, 19, 11, 21, 27, 33은 잠과 관련되어 꿈(dream), 알람(alarm), 낮잠(nap), 소변(pee), 뒤척임(toss), 졸음운전(drive)의 내용이 함께 나타나는 것으로 확인된다. 주제 8과 15는 코로나와 관련되어 자가격리(quarantine), 팬데믹(pandemic), 바이러스(virus) 등의 단어가 나타난다. 주제 25, 35, 41은 햇빛(sun), 담요(blanket), 방의 온도(room degree) 등 수면과 관련된 주변 환경적 요소들이 나타나고 있다.

4.2 딥러닝 분류 모델

불면증 관련 문헌과 비관련 문헌 분류를 통해 구축된 말뭉치를 바탕으로 딥러닝 자동분류 모델을 구축 및 비교한 성능 결과는 〈표 12〉와 같다. RoBERTa와 XLNet는 정확도 0.813으로 여타 언어모델보다 우수한 성능을 보이는 것으로 확인되었다. 이중 RoBERTa는 정밀도, 재현율, F1 점수에서 모두 가장 높은 성능을 보였다. RoBERTa가 BERT와 XLNet보다 더 뛰어나다는 논문(Liu et al., 2019)에 따르면, 이는 RoBERTa 모델이 기본적으로 대용량(160GB)의 데이터와 BOOKCORPUS, English WIKIPEDIA, OpenWebText 등의 다양한 데이터로 학습된 결과인 것으로 확인된다. 또한, BERT에서 NSP를 제거함으로써 NSP loss를 제거하였고 Dynamic Masking을 적용하여 성능이 더 높아진 것으로 확인된다. 반면, ALBERT는 정확도, 정밀도, 재현율, F1 점수에서 모두 가장 낮은 성능을 달성하였다. 이는, BERT보다 모델의 크기를 줄였고 다양한 데이터가 사전훈련되지 않았기 때문으로 보인다. 이를 통해 구축된 RoBERTa 모델은 이후 일별, 월별 불면증 지수를 개발하는 연구 및 불면증 관련 문헌을 분류하는 과업과 불면증 치료 연구 등에서 사용 가능할 것으로 판단된다.

5. 결 론

본 연구는 현대 사회에 만연해 있는 질병인 불면증의 심층적인 주제를 파악하고 딥러닝 자동분류 모델 구축을 통한 잠재 환자 진단 및 치료에 도움을 주기 위해 진행되었다. 이를 위해 자유롭게 의견을 남길 수 있는 소셜 미디어의 데이터를 수집하고 토픽 모델링의 새로운 기술인 BERTopic을 사용하여 심층 주제를 파악하였다. 또한, 수집된 데이터를 ICSD-3 기준과 전문가의 도움을 바탕으로 불면증 경향 문헌과 비경향 문헌으로 분류하여 학습데이터를 만들었고 5개의 딥러닝 언어모델을 통해 학습시킨 후 각각의 모델 성능을 비교 분석하였다.

〈표 12〉 딥러닝 자동분류 모델 성능 비교

Model	정확도	정밀도	재현율	F1 점수
BERT	0.810	0.784	0.801	0.792
RoBERTa	**0.813**	**0.809**	**0.811**	**0.810**
ELECTRA	0.806	0.768	0.792	0.780
ALBERT	0.795	0.771	0.766	0.768
XLNet	**0.813**	0.799	0.805	0.802

먼저, 토픽 모델링의 결과 8개의 대주제를 확인하였다. 이용자들은 불면증 커뮤니티를 통해 부정적 감정을 표출하고 도움과 조언을 구하고 있었으며 불면증과 관련된 질병을 같이 언급하고 있었다. 또한, 수면제 사용 및 처방에 대한 조언과 운동 및 식습관 변화 등의 내용을 언급하고 있다. 불면증과 관련된 특징들도 함께 발견되는데 호흡, 임신, 심장 등의 신체적 특징과 좀비, 수면 경련, 그로기상태 등의 활동적 특징, 햇빛, 담요, 온도, 낮잠 등의 수면 직간접적 영향을 주는 환경적 특징이 확인되었다.

전반적으로 BERTopic을 통해 분석된 주제들은 의미 있게 묶이고 각 주제들의 세부 내용(치료제의 종류, 질병의 종류 등)이 자세하게 표현되고 있음을 알 수 있다. 하지만, 몇 가지 제한 사항도 발견되었다. 음악(music)은 보통 환경적 특징으로 생각되지만, 단어 'listen'과 묶여 신체적 특징으로 구분되었고 수면제인 unisom, remeron은 수면제로 분류되지 않고 각각 임신과 연결되어 신체적 특징, stopped과 연결되어 환경적 특징으로 구분되는 것을 확인할 수 있다. 이는 단어의 의미를 사용하여 구분하는 BERTopic의 특징이자 한계로 판단된다.

딥러닝 자동분류 모델의 구축 및 성능 비교 결과는 RoBERTa가 가장 높은 성능을 보였다. 대용량의 다양한 데이터를 사전학습시킨 모델의 성능이 더 유용한 것으로 판단되며 구축된 RoBERTa 모델은 향후 불면증 트렌드를 파악하는 불면증 지수 개발 연구 및 불면증을 진단 및 치료하는 챗봇과 애플리케이션 개발에도 유용하게 활용될 수 있다.

본 연구는 아직 연구가 많이 진행되지 않은 불면증 소셜 미디어 데이터를 세부적으로 분석하기 위해 BERTopic이라는 새로운 토픽 모델링 기법을 사용했고 불면증 경향 딥러닝 자동분류 모델을 구축했다는 점에서 연구적 의의가 있으며 나아가 불면증의 세부적인 특징을 살펴봄으로써 질환을 깊이 이해할 수 있었다는 점에서 연구의 시사점이 있다.

참 고 문 헌

고영수, 이주희, 송민 (2021). 딥러닝 및 토픽모델링 기법을 활용한 소셜 미디어의 자살 경향 문헌 판별 및 분석. 한국비블리아학회지, 32(3), 247-264, https://doi.org/10.14699/kbiblia.2021.32.3.247

국민건강보험공단 (2020). 2020년 건강보험통계연보.

서울아산병원 (2014). 질환백과 불면증.
출처: https://www.amc.seoul.kr/asan/healthinfo/disease/diseaseDetail.do?contentId=31586

안경진 (2014). 수면장애, 국가 건강까지 위협한다. 메디칼업저버,
출처: http://www.monews.co.kr/news/articleView.html?idxno=76359

윤인영 (2013). 수면질환의 종류. Hanyang Medical Reviews, 33(4), 197-202.

https://doi.org/10.7599/hmr.2013.33.4.197

이수빈, 김성덕, 이주희, 고영수, 송민 (2021). 딥러닝 자동 분류 모델을 위한 공황장애 소셜미디어 코퍼스 구축 및 분석. 정보관리학회지, 38(2), 153-172. https://doi.org/10.3743/KOSIM.2021.38.2.153

Abuzayed, A. & Al-Khalifa, H. (2021). BERT for arabic topic modeling: an experimental study on BERTopic technique. Procedia Computer Science, 189, 191-194. http://doi.org/10.1016/j.procs.2021.05.096

Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint, arXiv:2008.09470. https://doi.org/10.48550/arXiv.2008.09470

Buysse D. J. (2013). Insomnia. The Journal of the American Medical Association, 309(7), 706-716. https://doi.org/10.1001/jama.2013.193

Cheng, Q., Li, T. M., Kwok, C. L., Zhu, T., & Yip, P. S. (2017). Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. Journal of Medical Internet Research, 19(7), e243. https://doi.org/10.2196/jmir.7276

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint, arXiv:2003.10555. http://doi.org/10.48550/arXiv.2003.10555

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Grootendorst, M. (2020). Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. Zenodo. https://doi.org/10.5281/zenodo.4381785

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint, arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

Guo, C., Lin, S., Huang, Z., & Yao, Y. (2021). Mental health question and answering system based on bert model and knowledge graph technology. Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences, 472-476. https://doi.org/10.1145/3500931.3501011

He, Q., Veldkamp, B. P., Glas, C. A., & de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. Assessment, 24(2), 157-172. https://doi.org/10.1177/1073191115602551

Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., & Taufik, N. (2021). Topic modeling for customer service chats. In 2021 International Conference on Advanced

Computer Science and Information Systems, 1-6. https://doi.org/10.1109/ICACSIS53237.2021.9631322

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jamison-Powell, S., Linehan, C., Daley, L., Garbett, A., & Lawson, S. (2012). "I can't get no sleep" discussing insomnia on twitter. Proceedings of the Sigchi Conference on Human Factors in Computing Systems, 1501-1510. https://doi.org/10.1145/2207676.2208612

Kingma, D. P. & Ba, J. (2015). Adam: A method for stochastic optimization. ICLR. 2015. arXiv preprint, arXiv:1412.6980, 9. https://doi.org/10.48550/arXiv.1412.6980

Koh, J. X. & Liew, T. M. (2020). How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds. Journal of Psychiatric Research, 145, 317-324. https://doi.org/10.1016/j.jpsychires.2020.11.015

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint, arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Martínez-Castaño, R., Htait, A., Azzopardi, L., & Moshfeghi, Y. (2021). BERT-Based transformers for early detection of mental health illnesses. In International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 189-200. https://doi.org/10.1007/978-3-030-85251-1_15

Nikhil Chandran, A., Sreekumar, K., & Subha, D. P. (2021). EEG-based automated detection of schizophrenia using long short-term memory (LSTM) network. In Advances in Machine Learning and Computational Intelligence, 26, Springer, Singapore, 229-236. https://doi.org/10.1007/978-981-15-5243-4_19

Sateia M. J. (2014). International classification of sleep disorders-third editiond. Chest, 146(5), 1387-1394. https://doi.org/10.1378/chest.14-0970

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!. arXiv preprint, arXiv:2004.14914. https://doi.org/10.48550/arXiv.2004.14914

van der Nagel, E. & Frith, J. (2015). Anonymity, pseudonymity, and the agency of online identity: Examining the social practices of r/Gonewild. First Monday, 20(3).

https://doi.org/10.5210/fm.v20i3.5615

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, T., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32. https://doi.org/10.48550/arXiv.1906.08237

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Ahn Kyung-jin (2014). Sleep disorder threatens national health. Medical observer. Available: http://www.monews.co.kr/news/articleView.html?idxno=76359

Asan Medical Center (2014). disease encyclopedia insomnia. Asan Medical Center, Available: https://www.amc.seoul.kr/asan/healthinfo/disease/diseaseDetail.do?contentId=31586

Ko, Young-Soo, Lee, Ju-Hee, & Song, Min (2021). Examining suicide tendency social media texts by deep learning and topic modeling techniques. Journal of the Korean Biblia Society for library and Information Science, 32(3), 247-264. https://doi.org/10.14699/kbiblia.2021.32.3.247

Lee, Soobin, Kim, Seongdeok, Lee, Juhee, Ko, Youngsoo, & Song, Min (2021). Building and analyzing panic disorder social media corpus for automatic deep learning classification model. Journal of the Korean Society for Information Management, 38(2), 153-172. https://doi.org/10.3743/KOSIM.2021.38.2.153

National Health Insurance Service (2020). 2020 National Health Insurance Statistical Yearbook.

Yoon, In-Young (2013). Introduction to sleep disorders. Hanyang Medical Reviews, 33, 197-202. https://doi.org/10.7599/hmr.2013.33.4.197