

한국어 기반 정신 질환 예측 모델의 개발 및 적용: 일기 분석을 통한 장기 모니터링

김정우^{1*}, 이도현^{2*}, 조혜민^{2*}, 김미수²¹전남대학교 에너지자원공학과²전남대학교 인공지능학부

* 공동 1저자

jeongwu510@naver.com, leedo573@naver.com, cho26739830@gmail.com,

misoo.kim@jnu.ac.kr

Development and Application of a Korean Language-Based Mental Health Disorder Prediction Model: Long-term Monitoring through Diary Analysis

Jeong-Wu Kim¹, Do-Hyun Lee², Hye-Min Jo², Misoo-Soo Kim²¹Dept. of Energy and Resources Engineering, Chonnam University²Dept. of Artificial Intelligence, Chonnam University

요 약

본 연구는 한국어 텍스트를 기반으로 정신 질환 예측 AI 모델을 개발하고, 이를 일기 분석 플랫폼에 적용하여 장기적인 정신 건강 모니터링 시스템을 제안한다. 현재 한국은 정신 건강 문제의 심각성에도 불구하고, 정신 건강 서비스 이용률이 매우 낮다. 이러한 문제를 해결하기 위해 우울증, 불안, 조울증 등의 정신 질환 발생 가능성을 예측하는 모델을 구축한다. 이 모델은 사용자가 일기를 통해 감정 변화를 기록하고, 정신 건강 문제를 조기에 진단하도록 돕는다. 연구 결과는 정신 질환의 조기 발견 및 예방에 기여하고, 한국어 기반 정신 건강 관리 서비스에 중요한 기초 자료를 제공할 것이다.

1. 서론

현대 사회의 급격한 변화로 인해 많은 사람들이 정신 건강 문제를 겪고 있다. 2024년 보건부 조사에 따르면, 응답자의 73.6%가 정신 건강 문제를 경험했으며, 한국은 OECD 국가 중 우울감을 가장 많이 경험하는 국가로 기록되고 있다[1]. 그러나 정신 건강 서비스 이용률은 4.3%에 불과하며, 많은 이들이 자신의 문제를 정신 질환으로 인식하지 않거나 스스로 해결할 수 있다고 믿는 경향이 있다[2].

최근 인공지능 기술의 발전으로 정신 질환 예측에 대한 관심이 높아지고 있으며[3,4,5], 특히 한국어 기반의 예측 모델 개발이 요구되고 있다. 본 연구는 기존의 영어 데이터를 중심으로 한 연구들과 달리, 한국어 텍스트 데이터를 활용한 모델을 구축하는 것을 목표로 한다.

또한, 정신 건강 상태는 시간에 따라 변화하기 때문에 장기적인 모니터링과 예측이 매우 중요하다[5]. 이에 따라 본 모델은 사용자가 작성한 일기를 분석하여 정신 상태를 진단하는 플랫폼으로 확장되

어, 감정 변화를 기록하고 조기 진단을 통해 정신 질환을 예방할 수 있도록 설계한다.

2. 관련 연구

Murarka et al. (2021)은 우울증, 불안, 조울증 등을 포함한 다섯 가지 주요 정신 질환을 다루는 데이터셋을 구축하였다. 이 데이터셋을 기반으로 LSTM, BERT, RoBERTa 세 가지 모델을 비교 분석한 결과, RoBERTa가 F1-score 0.89로 가장 우수한 성능을 기록했다. LSTM은 0.76, BERT는 0.87을 기록하며, RoBERTa가 두 모델보다 성능이 뛰어났다. 특히 ADHD와 PTSD의 탐지에서 강점을 보였으나, 우울증과 불안 분류에서는 여전히 개선이 필요한 부분이 발견되었다.

Ameer et al. (2022) 또한 Murarka et al. (2021)이 구축한 데이터셋을 활용하여 전통적인 머신러닝, 딥러닝, 그리고 전이 학습 모델의 성능을 평가하였다. RoBERTa 모델은 F1-score 0.83을 기록하며, 다른 전이 학습 모델(BERT: 0.82, XLNet: 0.80)보다

높은 성능을 보였다. 딥러닝 모델 중에서는 Bi-LSTM이 0.79의 F1-score로 가장 우수한 성능을 나타냈으며, 머신러닝 모델 중에서는 Logistic Regression이 F1-score 0.80을 기록하였다.

Iyortsuun et al. (2023)은 정신 질환 진단에서 머신러닝과 딥러닝 기술을 검토한 연구로, 조현병, 우울증, 불안장애 등 여섯 가지 정신 질환을 다룬 논문들을 분석하였다. 연구진은 PRISMA 방법론을 사용해 33개의 관련 논문을 분석하고, 각 논문에서 제안된 기술을 정신 질환별로 분류해 성능을 평가하였다.

3. 연구 방법

3.1. 사전 학습

본 연구에서는 Murarka et al. (2021)이 구축한 영어 데이터 셋을 활용하여 모델을 사전 학습시킨다. 이 데이터는 2015년부터 2020년까지 수집된 게시물로, 각 게시물은 해당 서브레딧의 주제에 따라 정신 질환과 관련된 라벨이 자동으로 부여되었다.

데이터는 Reddit API를 통해 13개 서브레딧에서 우울증, 불안, 조울증, ADHD, PTSD 관련 17,159개의 게시물을 수집했다. 비교군 데이터는 음악, 여행, 정치 등의 서브레딧에서 수집된 None 라벨의 게시물들로 구성되었으며, 이를 통해 정신 질환과 비정신 질환 게시물 간 차이를 학습할 수 있도록 했다. 각 게시물은 최소 10개의 추천 수와 30개 이상의 토큰을 포함하여 데이터 품질을 보장했다.

3.2. 파인 튜닝

한국어 데이터 셋을 구축하기 위해, 비교적 최근의 Reddit 데이터(2021년 이후)를 추가로 수집한다. 이를 위해 우울증, 불안, 조울증, ADHD, PTSD와 관련된 서브레딧에서 게시물들을 수집한 후, 이를 한국어로 번역한다.

번역된 한국어 데이터 셋은 파인튜닝 작업을 통해 모델에 적용된다. 이를 통해 모델은 영어로 학습된 지식을 바탕으로 한국어 텍스트에 적응하며, 한국어 데이터에 맞춰 성능을 최적화한다.

4. 예상 결과 및 향후 과제

본 연구는 한국어 텍스트를 기반으로 정신 질환 예측 모델을 개발하여 우울증, 불안, 조울증 등 다섯 가지 정신 질환을 예측하는 것을 목표로 한다. 이 모델은 일기 플랫폼에 적용되어 사용자의 기록을 분

석하고, 장기적인 정신 건강 상태 변화를 추적한다. 향후 연구에서는 다양한 질환으로 예측 범위를 확장하고, 인구 집단별 맞춤형 모델을 개발할 것이다. 또한, 임상 데이터로 모델의 효과를 검증하고, 의료진과 협력하여 실제 진단에 유용한 도구로 발전시킬 것이다.

사사

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 소프트웨어중심대학사업 (2021-0-01409)과 인공지능융합혁신인재양성사업 (IITP-2023-RS-2023-00256629) 연구 결과로 수행되었음

참고문헌

- [1] 보건복지부. "2024년 국민 정신건강 실태조사 결과 발표." 보건복지부, 2024. https://www.mohw.go.kr/board.es?mid=a10503010100&bid=0027&act=view&list_no=1482175&tag=&nPage=1
- [2] 국립정신건강센터. "정신건강 실태조사 결과." 정신건강포털, 2024. <https://www.mentalhealth.go.kr/portal/bbs/bbsDetail.do?bbsId=BBSSTATS&nttId=105>
- [3] Murarka, Ankit, Balaji Radhakrishnan, and Sushma Ravichandran. "Classification of mental illnesses on social media using RoBERTa." Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, Mexico City, 2021, pp. 45-52.
- [4] Ameer, Iqra, et al. "Mental illness classification on social media texts using deep learning and transfer learning." arXiv preprint, arXiv:2207.01012, 2022.
- [5] Iyortsuun, Ngumimi Karen, et al. "A review of machine learning and deep learning approaches on mental health diagnosis." Healthcare, vol. 11, no. 3, MDPI, 2023, pp. 285. <https://doi.org/10.3399/healthcare11030285>