대규모 언어모델 기반 우울장애 진단에서의 성별 편향 완화

양시현 [1], 강미경 [1], S. Shyam Sundar[1,2], 한진영 [1*]
[1] 성균관대학교, [2] 펜실베니아 주립대학교

thdal147@g.skku.edu, gy77@g.skku.edu, sss12@psu.edu, jinyounghan@skku.edu*

Debiasing Gender Bias in Large Language Models for Depression Detection

Sihyeon Yang[1], Migyeong Kang[1], S. Shyam Sundar[1,2], Jinyoung Han[1*]
[1]Sungkyunkwan Univ., [2]Pennsylvania State Univ.

요 약

대규모 언어모델(LLM)은 활용 범위가 빠르게 확산되고 있으며, 우울장애 조기 탐지 등 정신건강 분야에 적용하려는 시도가 이루어지고 있다. 그러나 학습 데이터에 내재된 사회적 편향이 모델 판단에 반영되어 결과의 공정성 및 신뢰성을 저해할 수 있다는 우려에도 불구하고, 정신건강 영역에서의 성별 편향 문제는 아직 충분히 검증되지 않았다. 이에 따라, 본 연구는 성별 정보가 LLM 의 우울장애 진단에서 편향을 유발하는지 검증하고, 현존하는 편향 완화 기법들을 통해 이를 통제할 수 있는지 확인하고자 하였다. 구체적으로, 동일한 증상을 서술한 텍스트에 대해 성별 단서 포함 유무에 따른 우울장애 진단 결과 차이를 분석한 결과, 약 29.1% 샘플에서 진단 불일치가 발생함을 확인하였다. 이후 편향 완화 전략을 적용한 후 동일한 조건에서 진단을 수행한 결과, 전체 샘플의 90.1%에서 진단이 일치함을 확인하였다. 이러한 결과는 성별 단서와 같은 비임상적 요인이 실제로 LLM 을 통한 우울장애 진단에 영향을 미칠 수 있음을 입증하고, 공정하고 신뢰할 수 있는 결과를 얻기 위해서는 LLM 활용 시 편향 완화 전략을 함께 고려해야 함을 시사한다.

Ⅰ. 서론

대규모 언어모델(LLM, Large Language Model)의 추론 능력이 향상됨에 따라, 다양한 분야에서 활용 범위가 빠르게 확산되고 있다. 특히 정신건강 분야에서도 대규모 언어모델의 임상적 배경지식을 활용하여 우울장애와 같은 정신장애 조기 탐지 및 정서 상태를 평가하려는 시도가 증가하고 있다.

대규모 언어모델의 학습 과정에서 성별·인종·연령 등 사회적 편향을 반영할 수 있다는 우려가 보고되며, 이에 따라 다양한 분야에서 대규모 언어모델의 편향 관련 연구가 활발히 진행되고 있다[1]. 그러나, 대규모 언어모델의 정신건강 영역에서의 성별 편향 문제는 충분히 다루어지지 않았다. 특히, 정신장애 진단 영역에서 편향이 발생할 경우, 진단의 공정성과 신뢰성을 저해하여 특정 집단에 대한 과소 진단 또는 과잉 진단으로 이어질 수 있다.

따라서 본 연구는 성별 정보가 대규모 언어모델의 우울장애 진단에서 편향을 유발하는지 검증하고, 편향 완화 기법을 통해 이를 통제할 수 있는지 확인하고자 한다. 구체적으로, 동일한 증상을 서술한 텍스트에 대해 성별 정보 포함 여부에 따라 우울장애 진단 결과 차이를 비교하였다. 실험 결과, GPT-4.1 은 동일한 증상임에도 성별 단서 유무에 따라 서로 다른 Patient Health Questionnaire-9(PHQ-9) 점수를 예측하는 경향을 보였으며, 약 29.1% 샘플에서 진단 불일치가 발생하였다. 이는 성별 단서와 같은 비임상적 요인이 실제로 LLM 을 통한 우울장애 진단에 영향을 미칠 수 있음을 시사한다.

이를 해결하기 위해 프롬프트 기반 편향 완화 접근법 중 Chain-of-Thought(CoT)와 Self-Refinement(SR) 기법을 적용하였다. CoT 는 LLM 연구에서 활용되는 대표적 추론 기법으로, 단계적 추론을 통해 응답을 생성하여 편향적 사고를 줄일 수 있다[2]. SR 은 모델이 스스로 초기 응답을 검토하는 방법으로, 최근 편향 완화 연구에서 효과적인 방법으로 주목받고 있다[3]. 실험 결과, Role-based SR 은 기존의 진단 일치도를 90.1%로 크게 향상시켜 사용자 수준에서의 성별 편향 완화 가능성을 입증하였다. 이는 정신장애 진단 영역에서 LLM 활용 시, 공정하고 신뢰할 수 있는 결과를 얻기 위해 편향 완화 전략을 함께 고려해야 함을 시사한다.

Ⅱ. 성별 편향 측정

2.1 측정용 데이터셋 확보

본 연구의 목적은 동일한 증상 묘사에 대해 LLM 이 성별 정보 유무에 따라 다른 우울장애 진단 결과를 내리는지 검증하는 데 있다. 이를 위해 온라인 게시글 기반 우울장애 심각도 예측을 위한 데이터셋 DepSeverity 를 활용하였다[4].

데이터셋 원문에는 작성자의 성별을 유추할 수 있는 표현(예: “I’m pregnant”)이 포함되어 있으며, 이는 모델이 증상 자체보다 성별 정보에 의존하여 결과를 다르게 예측하는 핵심 요인이 될 수 있다. 따라서 본 연구에서는 GPT-4.1 을 이용해 성별 단서를 제거하거나 중립적 표현으로 대체하였고, 이 과정을 통해 성별 단서가 포함된 데이터와 제거된 데이터를 구축하였다. 최종적으로 1,000 개의 쌍을 구성하여 성별 단서 유무가 진단 결과에 미치는 영향을 비교하였다.

2.2 편향 측정 지표

성별 단서 유무에 따른 모델 예측의 일관성을 Agreement(일치도)로 정의한다. 일치도는 동일한 텍스트에 대해 성별 포함 여부에 따른 예측 결과가 얼마나 유사한지를 측정하는 지표이다. 점수가 완전히 동일하거나

±1 점 이내인 경우를 '일치'로 정의하며 전체 데이터 중 일치 비율을 계산한다.

$$Agreement = \frac{1}{N}\sum_{i=1}^{N} 1\left[y_{no\ Gendered}^{(i)} = y_{Gendered}^{(i)}\right] \quad (1)$$

수식 1. Agreement 지표 정의

일치도가 높을수록 성별 정보 유무에 관계없이 모델이 일관된 판별을 하고 있음을 의미하고, 낮을수록 성별 정보에 의존하여 판별 결과가 크게 달라짐을 의미한다.

2.3 편향 측정 및 결과

본 연구에서는 OpenAI 의 GPT-4.1 을 사용하여 성별 단서 유무가 진단 결과에 미치는 영향을 분석하였다. GPT-4.1 은 내부 구조와 학습 데이터가 공개되지 않았으나 현재 가장 널리 활용되는 상용 모델이므로 실험 모델로 선정하였다. 재현성 확보를 위해 temperature 는 0.3 으로 고정하였으며, 각 텍스트에 대해 PHQ-9 점수(0-27 점)를 예측하였다.

실험 결과, 동일한 증상임에도 성별 정보의 포함 여부에 따라 예측 점수가 달라지는 편향이 존재함을 확인하였다. 기본 프롬프트로 각 게시글의 우울장애 점수를 예측했을 때 일치도는 70.9%로 나타났다. 이는 모델이 증상 자체보다 성별 단서에 의존해 서로 다른 수준의 우울장애 심각도로 판단했음을 시사한다. 따라서 진단의 신뢰성 확보를 위해 편향 완화 기법의 적용이 필요하다.

Ⅲ. 성별 편향 완화

3.1 편향 완화 기법

편향 완화 기법은 모델 내부 수정 방법과 프롬프트 기반 방법으로 구분된다. 모델 내부 수정은 파라미터 재학습이나 파인튜닝이 필요하기 때문에 일반 사용자 수준에서는 활용이 어렵다. 본 연구에서는 추가 학습 없이도 적용 가능한 프롬프트 기반 기법에 초점을 맞추었으며, 구체적으로는 CoT 와 SR 방법을 사용한다.

CoT 는 모델이 단계적 추론 과정을 거쳐 답변하도록 유도하는 방법이다[2]. 이를 통해 텍스트에서 증상을 추출하여 PHQ-9 항목에 매핑하고, 최종 점수를 출력하였다. 이러한 절차를 활용해 모델이 성별 단서에 의해 편향된 응답을 내릴 위험을 줄이고자 하였다.

SR 은 모델이 생성한 초기 응답을 다시 입력하여 스스로 편향 여부를 검토하도록 유도하는 방법으로, 기존 편향 완화 연구에서 효과적인 방법으로 보고된 바 있다[3]. 이는 적용 방식에 따라 Role-based 와 Instruction-based 접근으로 구분할 수 있다. Role-based 는 모델에 특정 역할을 부여해 자기 점검 과정을 수행하도록 하는 방법이며, Instruction-based 는 별도의 지시문을 통해 초기 응답을 재평가하고 수정하도록 유도하는 방법이다. 본 연구에서는 두 가지 기법을 모두 적용해 성별 편향 완화 효과를 검증하였다.

3.2 편향 완화 및 결과

기존의 일치도를 향상시키기 위해 프롬프트 수준의 편향 완화 전략을 적용하였으며, 결과는 표 1 과 같다.

프롬프트 (Type of Prompt)	일치도 (Agreement)	편향 완화율 (Δ)
Base prompt	0.709	-
CoT prompt	0.834	+12.5%
SR (Instruction-based)	0.649	-8.5%
SR (Role-based)	**0.901**	**+19.2%**

표 1. 편향 완화 결과

실험 결과, Role-based SR 방법이 가장 큰 효과를 나타내어 90.1%의 일치도를 기록하였으며, 이는 기존 일치도 대비 19.2% 향상된 결과다. CoT 도 83.4%로 기존 일치도 대비 12.5% 향상되어 단계적 추론이 편향 완화에 기여함을 확인하였다. 반면, Instruction-based SR 방법은 64.9%로 하락하는 결과를 보였다.

이는 Role-based SR 과 CoT 모두 성별 편향을 완화하는 데 효과적임을 보여준다. 특히 Role-based SR 방법은 모델에게 "편향이 없는 평가자" 역할을 부여하여 가장 큰 완화율을 보였으며, CoT 도 단계적 추론을 통해 증상 단서에 집중하도록 유도하여 성능을 향상시켰다. Instruction-based SR 은 일치도가 하락하였는데, 세밀한 지시를 추가하는 방식이 반드시 성능 개선으로 이어지지 않을 수 있음을 시사한다.

Ⅳ. 결론

본 연구는 LLM 이 성별 정보 포함 여부에 따라 우울장애 진단에서 편향을 보이는지 확인하고, 이를 프롬프트 기법으로 완화할 수 있는지 검증하였다. 실험 결과, 성별 정보 유무에 따라 서로 다른 우울장애 점수를 예측하는 경향이 나타났다. 또한, Role-based SR 방법을 적용하여 일치도를 90.1%로 향상시켰다. 이는 적절한 프롬프트 설계가 우울장애 진단 과정에서 성별 편향을 완화하는 대응책이 될 수 있음을 시사한다. 다만, 제한된 프롬프트 기법과 GPT-4.1 에 국한된다는 한계가 있으며, 향후에는 다양한 multi-step 기법과 오픈소스 모델을 대상으로 모델별 편향 특성을 분석할 필요가 있다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부 및 한국산업단지공단의 산업집적지 경쟁력강화사업(VCSK2502)의 연구결과로 수행되었으며, 과학기술정보통신부 및 정보통신기획평가원의 디지털분야해외석학유치지원 연구결과로 수행되었음(RS-2024-00459638).

참 고 문 헌

[1] Kotek, H., Dockum, R., & Sun, D. (2023, November). Gender bias and stereotypes in large language models. In Proceedings of the ACM collective intelligence conference (pp.12-24).

[2] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., … & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

[3] Furniturewala, S., Jandial, S., Java, A., Banerjee, P., Shahid, S., Bhatia, S., & Jaidka, K. (2024, November). "Thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 213-227).

[4] Naseem, U., Dunn, A. G., Kim, J., & Khushi, M. (2022, April). Early identification of depression severity levels on reddit using ordinal classification. In Proceedings of the ACM web conference 2022 (pp. 2563-2572).