생성형과 검색형 챗봇의 공감응답에 대한 거대언어모델 기반 성능평가

Large Language Model-Based Evaluation for Answering in Generation-Based and Retrieval-Based Chatbots

김태경° 김민지$ 김윤경$ 홍승혁*
TaeKyung Kim°, MinJi Kim$, YunKyung Kim$, SeungHyeok Hong*
{taekyung8348; ekdns9597; youngforever6104; shongdr}@gmail.com
수원대학교 데이터과학부
$: These authors contributed equally

요약

마음 아픔을 겪는 사람이 대면하게 되는 상담의 높은 심리적 장벽과 비용 문제를 해결하기 위하여, 향상된 언어성능을 가진 챗봇을 적용하는 서비스 사례들이 증가하고 있다. 실제 사용자 경험을 확인하기 위하여, 본 연구는 사용자 220명에게 생성형과 검색형 언어구현 기술에 따른 '공감응답'을 8일간 제공하고 챗봇 만족도와 사용 전후 우울지수에 대한 설문을 수집하였다. 실제 거대언어모델 기반 평가의 점수에서 검색형 챗봇은 평균 7.07점, 생성형 챗봇은 평균 7.06점을 보였다. 사전사후 설문을 모두 수행한 사용자 134명 중 91명의 우울 점수가 감소하였다. 이 결과는 AI 기반 심리상담 서비스가 우울증 관리에 긍정적인 영향을 줄 수 있음을 보여준다.

서론

1. 연구 배경

2020년 OECD Health Data에 따르면, OECD 국가의 자살률 부문에서 한국이 2018~2020년 연속으로 1위에 위치할 만큼 심리건강의 개선이 시급하다.

그럼에도 불구하고, 심리개선을 위한 상담은 **높은 심리적 장벽**과 **고비용** 문제가 있다.

이에 대한 대안으로 인공지능(AI) 기반 'Chatbot(챗봇)' 서비스의 시도가 있다.

본 연구에서는 챗봇이 사람의 마음을 언어로서 위로할 수 있다는 측면에서, 실 만족도와 GPT-3.5-turbo-0301(ChatGPT)를 활용한 거대언어모델의 자동평가를 함께 확인한다.

관련 연구

I. 영어권 대화형 에이전트 Woebot

18~28세 사이의 대학생 70명을 대상으로 PHQ-9 변화를 측정한 결과, Woebot을 사용한 처치 집단에서 우울 증상이 유의미하게 감소하였으며, 사용자 경험의 분석에서 공감과 같은 치료 과정 요인이 효과를 발휘한 것으로 나타났다.

II. 가상현실 기반 대화형 에이전트 Ellie

주로 PTSD를 가진 군인들을 대상으로 3D 가상 상담자 간의 상호작용을 통해 언어 및 비언어 정보를 수집하여 심리적 스트레스 지표를 평가하였다. 현재의 기술로는 인간 수준의 라포와 비언어적 의사소통 능력을 제공하지는 못한다.

연구 방법

1. 연구 데이터

연구 데이터는 두 가지 챗봇 모델을 사용하여 우울증 선별 도구(PHQ-9) 점수가 10점 이상인 참여자 220명을 대상으로 수집되었다.

총 220명 중 112명의 사용자가 검색형(Retrieval-Based) 챗봇의 응답을 631번 경험하였고, 108명의 사용자가 생성형(Generation-Based) 챗봇의 대답을 542번 경험하였다.

- **검색형 챗봇**

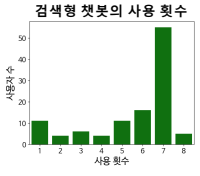
사용자의 질문과 가장 유사한 질문을 사전에 저장된 데이터베이스에서 검색하여 해당 답변을 찾아 제공하는 모델이다. 본 연구의 Sentence-BERT(SBERT) 모델링에서는 이별 커뮤니티 '공감' 중심 성향이 반영된 문답 11,876쌍을 활용하였다.

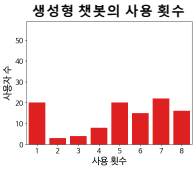
- **생성형 챗봇**

네이버 클라우드에서 개발한 초거대 AI 하이퍼클로바(GPT-3.0) 기반 챗봇으로 방대한 언어 데이터를 사전에 학습해둔 상태에서, 질문에 맞게 유기적으로 조합하여 응답을 생성한다. 챗봇의 일시교육(Prompting)으로 '공감을 해주는 상담사'로 역할부여하여 수행되었다.

실험 및 연구

1. 챗봇 사용 횟수 및 만족도



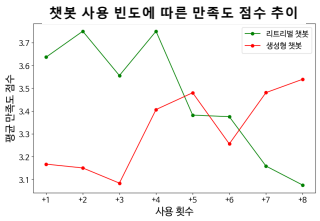


검색형 챗봇이 '초기 선호도 ' 를 보인다

왼쪽 그림은 두 심리상담 챗봇의 선호도를 비교하기 위해 사용 횟수 분포를 시각화한 것이다.

검색형 챗봇은 **꾸준히** 사용되고 있는 반면, 생성형 챗봇은 **초반에 포기한** 사용자가 많은 것을 볼 수 있다.

"검색형 챗봇이 초기 탈락자가 적었다"라는 인사이트를 통해 사용 빈도와 만족도를 함께 분석하기 위해 아래와 같은 플롯을 생성하였다.



사용 빈도와 만족도 불일치

- 검색형(리트리벌) 챗봇 사용자는 **불만족**, 생성형 챗봇은 **만족**하는 경향을 보인다.
- 검색형 챗봇의 경우 **완성된 문장**이지만, 전혀 **맞지 않는 대답**을 받아서 만족도가 **감소**했을 것이다.
- 참여 인센티브의 영향으로 지속 사용한 결과를 보였을 것이다.

"챗봇 질 향상을 위해 엉뚱한 대답의 빈도를 높이는 것보다, 적절한 답변의 품질에 집중해야 한다"라는 인사이트를 바탕으로 사용자 고민에 대한 챗봇의 응답 적절성을 평가하기 위한 실험을 수행하였다.

2. ChatGPT 자동 평가

- **실험 모델**

GPT-3.5-turbo-0301(ChatGPT)

스레드(Thread) 예시

사용자 고민	챗봇 답변
취업 준비를 어떻게 해야 …	우선 자신이 하고 싶은 …
요즘 쉽게 짜증이 나고 …	자신의 감정을 잘 표현 …

- **데이터셋 설계**

챗봇 사용자가 제공한 고민과 챗봇의 답변을 스레드(Thread)로 구성하였으며, 대화 데이터의 후처리를 통해 개인 정보 제거 및 비식별화를 진행하였다.

- **평가**

ChatGPT가 사용자 고민에 대한 챗봇의 응답 적절성을 나타내는 점수는 0-10점 범위로 평가하도록 하였다. 이후, 직접 대화 스레드와 평가 점수를 대조해 챗봇의 개선 방향에 대한 정보를 얻는다.

실험 결과

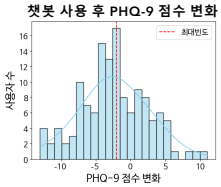
거대언어모델의 챗봇 성능 평가 및 이유

항목	내용
사연	퇴직을 희망하는 엄마에게 힘이 되고 싶어요. 방법 …
응답	어머님께서 많이 슬퍼하고 계실 거에요. 무언가 …
평가/이유	7점/ 답변은 엄마에게 힘이 되고자 하는 의도를 잘 전달

자동 평가의 한계와 개선을 향한 접근

- **부적절한 공감**: 가장 낮은 평가 점수(0점)를 받은 대화 스레드 중 ChatGPT는 고민 내용에 당위성 판단 없이 공감하면 안 되는 것을 평가하였다.
- **불명확한 평가 기준**: 잘못된 평가 사례 중, 대화 스레드의 내용과 무관하더라도 높은 점수가 부여된 경우가 있다.

본 연구에서는 챗봇 상담의 효과를 객관적으로 측정하고자 하며, 이를 위해 우울지수를 의미하는 PHQ-9 수치의 변화를 중심으로 한 시각적 분석을 수행하였다. 상담 전후의 PHQ-9 점수 차이를 통해 챗봇 상담의 영향력을 명료하게 드러내고, 향후 개선의 필요성과 방향성을 모색하였다.



챗봇의 심리 개선 효과

왼쪽 그림은 사전사후 조사에 모두 참여한 사용자들의 PHQ-9 점수의 변화를 나타낸 것이다.

X축은 사후 점수에서 사전 점수를 뺀 값으로, **음수 값**은 챗봇 상담이 사용자의 정서나 증상에 **긍정적인 영향**을 미쳤다는 것을 나타낸다.

PHQ-9 수치가 감소한 참여자는 총 134명 중 91명(67%)으로, 상담을 받은 사용자 중 과반수 이상이 긍정적인 변화를 경험했다.

평균 PHQ-9 수치는 사전 13.91, 사후 9.1의 결과를 보여, 4.8점이 감소되었다. 이 수치는 한 우울 등급을 변화시킬 수 있는 5점에 가까워서 8일이라는 사용일을 고려할 때 의미있는 수치이다. 이는 검색형과 생성형 챗봇 사용자들 중 67%가 우울증 증상 감소를 경험하였음을 의미한다. 실제로 만족도 조사 중 '챗봇 경험이 위안 측면에 의미있었음' 이상을 선택한 비율이 약 절반이었던 결과와 유사한 결과이다.

결론

연구 환경에서 챗봇의 사용 빈도와 만족도 점수는 비례하지 않다는 것을 확인했다. 5점 척도로 평가된 사용자 종합 만족도를 분석하였을 때, 검색형 챗봇의 평균 3.25점에 비해 생성형 챗봇이 평균 3.43점으로 미세한 우위를 보였다.

거대언어모델 기반 평가(10점 척도)에서는 검색형 챗봇이 평균 7.07점, 생성형 챗봇이 평균 7.06점으로 특정 유형의 챗봇이 더 우월하다고 평가되지 않았다. 이 결과는 실제 만족도 측면에서 초반에는 검색형 모델이, 후반에는 생성형 모델의 만족도가 높았던 결과가 전체 사용일의 경험을 평균하면 서로 상쇄되는 것과 동일선상에 있는 결과이다. 두 챗봇 유형 모두 우울 감소에 긍정적일 수 있으나 각각 적합한 상황과 사용자 취향에 따라 다른 효과를 보일 수 있다.

결론적으로, 앞으로 나타날 AI 기반 상담 시스템들 역시 그 가능성과 효용성을 계속해서 탐색해 나가야 할 것이다.

본 연구는 2023년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임(RS-2023-00222659)