복지상담 음성데이터 기반 우울증 조기진단을 위한 감성 분류 모델 성능 비교에 관한 연구

이제민, 윤수연*
국민대학교, *국민대학교

jm20008@kookmin.ac.kr, *1104py@kookmin.ac.kr

A Comparative Study on Sentiment Classification Models for Early Depression Detection Using Welfare Counseling Voice Data

Lee Jemin, Yoon Soo Yeon*
Kookmin Univ., *Kookmin Univ.

요 약

본 연구는 비정형 상담 음성 데이터를 기반으로 화자 단위 감정 특성이 반영된 전사 텍스트를 생성하고, 이를 입력으로 활용한 GPT 기반 감정 분류 모델의 성능을 정량적으로 비교하였다. Whisper-tiny로 전사된 텍스트를 활용하여, Scikit-LLM의 ZeroShotGPTClassifier를 통해 '우울'과 '비우울' 간 zero-shot 이진 분류를 수행하였다. 분류 모델은 GPT-3.5 turbo와 GPT-4o를 적용하였으며, 정확도, 정밀도, 재현율, F1-score 및 혼동행렬을 통해 성능을 평가하였다. 실험 결과, GPT-4o는 GPT-3.5 turbo와 비교했을 때 정밀도 측면에서 0.8169로 더 높은 수치를 기록하며 오탐 방지에는 효과적이었지만, 재현율이 0.1928로 낮아 실제 우울 발화를 충분히 식별하는 데에는 한계가 있었다. 이러한 결과는 GPT-4o가 감정 분류에 있어 보다 감정 고위험군 탐지를 위한 보완적 접근이 필요함을 보여준다. 본 연구는 상담 음성 기반 감정 분류의 자동화 가능성을 확인하였으며, 자연어처리 기반 감정 인식 기술이 심리상담 및 정신건강 분야에서 조기 선별 도구로 활용될 가능성을 제시한다.

Ⅰ. 서 론

우울증은 전 세계적으로 가장 널리 퍼진 정신질환 중 하나로, 삶의 질을 현저히 떨어뜨릴 뿐 아니라 자살이라는 극단적 결과로 이어질 수 있는 주요 요인이다. 보건복지부에 따르면, 국내 전체 인구의 약 18.1%가 일생 동안 한 번 이상 우울 증상을 경험하며, 연간 약 5% 이상의 유병률을 나타내는 것으로 보고되었다. 이에 따라 상담 현장에서의 우울 상태 조기 진단은 정신건강 고위험군을 선별하기 위한 핵심 과제로 부각되고 있다.

상담은 우울 상태를 감지할 수 있는 중요한 접점임에도 불구하고, 대부분의 상담 현장은 내담자의 주관적 진술과 상담자의 경험에 의존하고 있어 정량적 진단과 선제적 개입에는 한계가 있다. 특히 비정형 대화 데이터에서 우울 표현은 은유나 간접적인 언어로 나타나는 경우가 많아, 이를 자동으로 해석하고 예측할 수 있는 기술이 요구된다.

최근 STT 기술과 대규모 언어 모델(LLM)의 진보는 자연어 기반 정서 분석의 새로운 가능성을 열고 있다. 본 연구는 복지상담 음성데이터를 전사하고, 이를 기반으로 우울 표현을 탐지하는 조기 예측 모델을 제안함으로써, 정서 기반 상담 지원의 자동화 가능성을 실증하고자 한다.

Ⅱ. 관련 연구

2.1 STT(Speech-to-Text) 모델

STT 기술은 음성 데이터를 자연어 처리 입력으로 활용하기 위해 텍스트로 변환하는 핵심 전처리 과정이다. Whisper는 OpenAI에서 개발한 STT 모델로, 다양한 억양, 발화 속도, 잡음 환경에서도 안정적인 전사 성능을 보인다. 특히 상담 음성과 같이 비정형적이고 감정이 섞인 발화에 대해 Whisper는 상대적으로 높은 정확도를 기록하고 있다. 본 연구에서는 Whisper-tiny 모델을 활용하여 상담 음성을 텍스트로 변환한다.

2.2 Scikit-LLM

Scikit-LLM은 대규모 언어 모델을 활용한 텍스트 분류기이며, scikit-learn과 호환 가능한 인터페이스를 제공한다. 사전학습된 LLM을 기반으로 zero-shot 분류가 가능하다는 점에서 별도의 학습 없이도 감정 분류가 가능하다는 장점이 있다. 본 연구에서는 GPT-4o 모델 기반 Scikit-LLM을 활용하여 전사된 상담 텍스트로부터 우울 표현을 자동 감지하는 분류기를 구현한다.

Ⅲ. 실험 설계

3.1 실험환경 및 데이터셋

본 연구는 AI Hub의 "정신건강 상담" 데이터 중 '우울증' 및 '비우울' 클래스에 해당하는 음성 파일을 각각 2,500개씩 샘플링하여 총 5,000개를 실험에 활용하였다. 데이터가 방대한 관계로 Google Colab 환경에서 실험을 수행하였으며, Whisper-tiny 모델을 통해 STT 전처리를 진행하였다.

3.1.1 STT(Speech To Text)데이터셋 전처리

Whisper-tiny 모델은 각 화자 ID(speaker_id)를 기준으로 음성 파일(.wav)과 라벨링 파일(.json)을 매칭하여, 정답 문장(gt_text)과 모델이 생

성한 전사 결과(pred_text)를 구성하였다. 생성된 전사 결과는 CSV 형식으로 병합 및 저장되었다.

감정이 포함된 비정형 상담 발화에 대해서도 Whisper-tiny는 일정 수준 이상의 전사 품질을 보였으며, 전사 정확도는 Word Error Rate(WER)와 Character Error Rate(CER)를 기준으로 평가되었다. 이 지표는 수치가 낮을수록 전사 문장에서 누락, 삽입, 대체와 같은 오류가 적다는 것을 의미한다. [표 1]은 해당 모델의 전사 성능을 요약한 결과이다.

[표 1] Whisper-tiny 전사 성능 평가 결과

WER	CER
0.5847	0.2542

3.2 LLM 기반 우울 감정 분류

전사된 텍스트는 동일한 speaker_id를 갖는 복수의 발화 문장을 시간 순으로 정렬한 뒤 하나의 텍스트로 통합하여, 화자 단위의 감정 특성이 보다 뚜렷하게 반영되도록 구성하였다. 해당 텍스트를 입력값으로 하여 Scikit-LLM 라이브러리를 기반으로 감정 분류 실험을 수행하였고, 분류기는 OpenAI GPT API를 이용한 ZeroShotGPTClassifier로 구현하였다. 실험에는 GPT-3.5 turbo와 GPT-4o 모델을 적용하였으며, "우울"과 "비우울" 두 클래스를 대상으로 zero-shot 방식의 이진 분류를 수행하였다. 분류 성능 평가지표로는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score를 사용하였고, 혼동행렬을 통해 각 모델의 예측 특성을 시각적으로 비교하였다.

3.3 실험 결과 및 성능 평가 분석

본 연구에서는 Whisper-tiny를 활용하여 전사된 비정형 상담 음성 데이터를 Scikit-LLM 기반 감정 분류 모델의 입력으로 활용하였다. 감정 분류에는 OpenAI GPT API를 이용한 ZeroShotGPTClassifier로 구현하였으며, GPT-3.5 turbo와 GPT-4o 모델 실험을 진행하였다. 두 모델 모두 “우울”과 “비우울”의 이진 분류를 수행하였고, 평가 지표로는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score를 사용하였다.

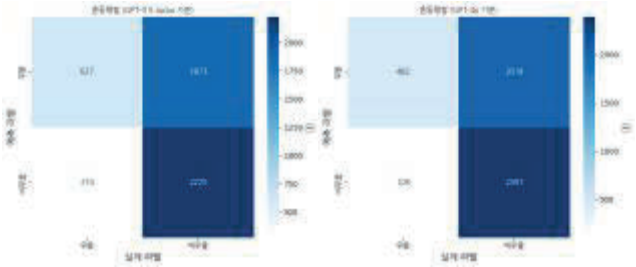
[표 2] GPT 기반 모델별 우울 감정 분류 성능 비교

Model	Accuracy	Precision	Recall	F1-score
GPT-3.5 turbo	0.5705	0.6959	0.2508	0.3687
GPT-4o	0.5747	0.8169	0.1928	0.3120

실험 결과 [표 2] 에 따르면, GPT-4o는 GPT-3.5 turbo 대비 전반적인 정확도 비교에서는 0.5747 → 0.5705을 나타내었고, 정밀도에서는 0.8169 → 0.6959 비교 수치를 기록하며 더 우수한 성능을 보였다. 특히 정밀도에서 0.12 이상의 향상이 확인되었으며, 이는 모델이 우울 발화를 예측할 때 상대적으로 적은 오탐률을 보여주고 있음을 의미한다.

그러나 GPT-4o의 재현율은 0.1928로 GPT-3.5 turbo 수치0.2508보다 낮았으며, 결과적으로 F1-score 는0.3120으로 낮은 수치를 기록하였다. 이는 GPT-4o가 우울 감정 분류에 있어 실제 우울 발화를 놓치는 경우가 많음을 보여주고 있다.

[그림 1]의 혼동행렬 비교 분석을 통해 이러한 경향은 더욱 명확히 나타난다. GPT-3.5 turbo는 우울 클래스의 정확한 식별은 627건이었으나, 비우울을 우울로 잘못 분류한 경우도 274건에 달하였다. 반면 GPT-4o는 우울 발화를 482건만 정확히 분류했으며, 비우울을 우울로 오분류한 경우는 108건으로 GPT-3.5 turbo 대비 크게 감소하였다. 이는 GPT-4o가 오탐 방지에는 효과적이지만, 실제 우울 발화의 인식률이 낮아 Recall과 F1-score에 낮은 성능을 보이는 것으로 해석된다.



[그림 1] GPT-3.5 turbo 및 GPT-4o 기반 혼동행렬 (Confusion Matrix) 비교 결과

IV. 결론 및 시사점

본 연구는 GPT 기반 Zero-shot 감정 분류 모델의 성능을 상담 음성 전사 텍스트에 적용하여 분석한 결과, 다음과 같은 결론을 도출할 수 있었다.

첫째, GPT-4o는 GPT-3.5 turbo 대비 높은 정밀도를 보이며 오탐 방지에 강점을 지닌 모델로 이는 실제 상담 현장에서 비우울 화자를 우울로 잘못 분류하여 발생할 수 있는 과잉 개입 가능성을 낮추는 데 유리할 수 있다.

둘째, 반면 GPT-4o의 낮은 재현율은 실제 우울 화자를 놓치는 위험을 내포하고 있으며, 이는 우울 고위험군의 조기 선별이 중요한 상담 및 정신건강 도메인에서 중요한 제한점으로 작용할 수 있다.

셋째, Whisper-tiny 기반 전사 데이터는 비교적 높은 품질의 음성 텍스트를 제공함으로써, 후속 감정 분류 실험에서 신뢰할 수 있는 입력값으로 활용되었으며, 다양한 음성 기반 정서 분석 연구의 출발점으로 활용될 수 있음을 확인하였다.

향후 연구에서는 단일 모델 기반 zero-shot 접근 외에도, 다수 모델 앙상블 혹은 Prompt Engineering 기법을 접목한 방법론을 적용하여 재현율과 F1-score를 동반 향상시킬 수 있는 균형적 모델 설계가 필요하다. 또한 상담 문맥을 반영한 사전학습 또는 도메인 튜닝 기법을 적용할 경우, 보다 정밀한 감정 인식 성능 확보가 가능할 것으로 기대된다.

참 고 문 헌

[1] 보건복지부, “2023년 정신건강실태조사,” 보건복지부, 2023.

[2] S. W. Son, “메타버스에서 텍스트 및 음성 데이터를 활용한 우울증 분류 시스템 개발,” 석사학위논문, 한신대학교 일반대학원 IT영상데이터융합협동과정, 2025.

[3] S. J. Yeo, “국내 AI 기반 심리상담 연구동향: 주제범위 문헌고찰,” 석사학위논문, 숙명여자대학교 교육대학원 상담교육전공, 2024.

[4] D. W. Shin, “자살 고위험군 조기진단을 위한 AI 기반 알고리즘 연구,” 의학박사학위논문, 서울대학교 대학원 의학과 정신과학교실, 2022.