AI 기반의 자살위험 예측 모델에 대한 연구

이지선, 윤수연*

국민대학교, *국민대학교

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

A Study on AI-Based Suicide Risk Detection from Counseling Conversations

Ji Sun Lee, Soo Yeon Yoon*
Kookmin Univ., *Kookmin Univ.

요 약

본 연구는 정신건강 상담 발화 내 자살 위험 징후를 조기에 탐지하기 위한 자동화 시스템을 구축하고, 그 성능을 비교·분석하였다. Whisper-tiny 및 Wav2Vec2.0 기반 음성 인식 모델을 활용해 발화 텍스트를 전처리한 후, GPT-4o-mini 기반 Scikit-LLM 분류기를 zero-shot 방식으로 적용하였다. 실험 결과, Whisper-tiny 기반 파이프라인은 전체 정확도 98.53%, 자살 클래스에 대한 F1-score 0.958 로 Wav2Vec2.0 보다 우수한 탐지 성능을 보였다. 특히 Whisper-tiny 는 높은 정밀도와 민감도를 유지하며 자살 발화를 보다 정확하게 분류하였다. 이는 Whisper 모델이 자살 관련 표현을 보다 정밀하게 전사함으로써 분류 정확도 향상에 기여했음을 시사한다. 본 연구는 자살 고위험군 조기 식별을 위한 실용적인 음성 기반 LLM 분류 파이프라인의 가능성을 제시한다.

Ⅰ. 서 론

한국은 OECD 국가 중 압도적 1 위의 자살률을 유지하고 있으며 이는 지속 증가하고 있다. 사회는 경쟁 심화와 사회적 고립 증가로 인해 불안, 우울, 스트레스 등 정신건강 문제가 심각해지고 있으며, 자살은 그 중 가장 극단적인 결과로 공공 보건의 중요한 이슈로 부상하고 있다. 한편, 국내 성인의 평생 정신질환 유병률은 25.4%에 달하지만, 실제 서비스 이용률은 9.6%에 불과해 접근성과 사회적 낙인이 큰 장벽으로 작용하고 있다(장규현, 2021).

자살은 환자의 주관적 응답에 의존한 진단 체계로 인해 정확한 선별이 어렵기 때문에 객관적이고 자동화된 예측 도구의 도입이 시급하다(신다운, 2022). 국내 정신건강 문제의 심각성과 진단 체계의 구조적 한계를 고려할 때, AI 기술을 활용한 자살위험 예측 연구는 그 필요성이 부각되고 있다.

Ⅱ. 관련 연구

2.1 STT(Speech to Text)

STT(Speech-to-Text)는 음성 데이터를 텍스트로 변환하는 핵심 전처리 기술이다. 기존에는 Google Speech API 나 kaldi 기반 모델들이 주로 사용되었으나, 최근에는 사전학습된 Transformer 기반의 Wav2Vec2.0 과 Whisper 계열 모델이 높은 정확도와 범용성을 바탕으로 주목받고 있다. 본 모델은 의료 사각지대에서도 적용 가능성이 높으며, 향후 다중 분류 및 설명 가능한 AI 로 확장될 수 있다.

2.2 Scikit-LLM

Scikit-LLM 은 Scikit-learn 과 호환되는 텍스트 분류 도구로, zero-shot 또는 few-shot 방식으로 고성능 분류기를 빠르게 적용할 수 있다는 장점이 있다. Scikit-LLM 은 텍스트 데이터를 바로 Pandas DataFrame 형태로 입력받고, .fit() 없이 .predict()만으로도 실행 가능하다는 점에서 간편한 LLM 인터페이스로 평가받고 있다. 또한 OpenAI API 와 연동이 가능해, 모델 내부 변경 없이 다양한 프롬프트 설정 실험이 가능하다.

Ⅲ. 실험

3.1 데이터셋 구성

AI-Hub 의 "복지 분야 콜센터 상담 데이터"를 활용하여 총 563,251 문장에 대한 발화를 STT 로 변환 후, 정제 및 병합 과정을 거쳐 suicide(자살위험) 또는 non-suicide(일반상담)로 태깅하여 이진 분류 데이터셋을 구성하였다.

3.1.1 STT(Speech To Text)데이터셋 전처리

STT 는 화자의 고위험 발화를 정형 텍스트로 변환해 LLM 기반 분류기의 입력으로 사용할 수 있도록 하는 정보 전달의 관문이자 전처리 핵심 단계로 기능한다.

모델은 Whisper-tiny 와 Wav2Vec2.0 를 동일한 조건에서 비교하여 사용하였다. 성능 평가지표로는 WER(Word Error Rate)와 CER(Character Error Rate)를 사용하였다.

<표 1-1. Wav2Vec2.0 로 STT 변환 후 데이터 예시>


```
{
  "file_id": "HOS110003122242A019",
  "speaker_id": "HOS0003122",
  "label": "nonsuicide",
  "pred_text": "거리 불가는한 생항이 죠",
  "gt_text": "거의 불가능한 상황이죠.",
  "Model": "Wav2Vec2.0",
  "WER": 1.33,
  "CER": 0.46
}
```


<표 1-2. Whisper-tiny 로 STT 변환 후 데이터 예시>


```
{
  "file_id": "MEN21000561752B007",
  "speaker_id": "MEN0005617",
  "label": "suicide",
  "pred_text": "이제는 솔직히 모르겠습니다.",
  "gt_text": "이젠 솔직히 모르겠습니다.",
  "Model": "Whisper-tiny",
  "WER": 0.33,
  "CER": 0.14
}
```


<표 2. WER, CER 비교>

Model	WER	CER
Wav2Vec2.0	1.0030	0.4719
Whisper-tiny	**0.6843**	**0.3584**

WER 과 CER 수치는 낮을수록 STT 모델의 성능이 우수함을 의미한다. Whisper-tiny 모델은 Wav2Vec2.0 에 비해 전반적인 오류율이 낮아, 품질이 더 뛰어남을 입증하였다.

3.2 Scikit-LLM 예측

정신건강 상담 발화에서 자살 위험 징후를 조기에 탐지하기 위해 GPT-4o-mini 기반 Scikit-LLM 분류기를 활용하였다. Zero-shot 으로 별도의 학습 없이 총 5,313 개의 발화 세션, 563,251 문장에 대해 suicide 와 non-suicide 여부를 예측하였다.

3.3 실험 결과 및 성능 평가

Whisper-tiny 기반 파이프라인은 정확도 98.53%, suicide F1-score 0.958 를 달성하며 자살 고위험 발화를 민감하게 탐지하였다.

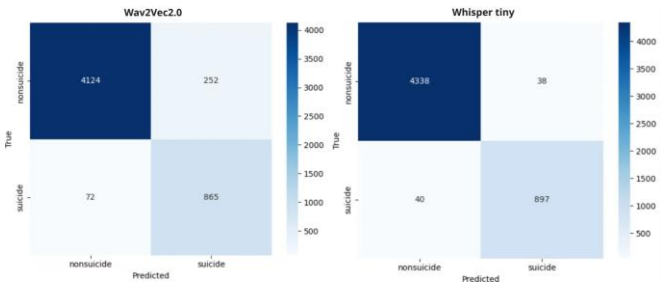
<표 3-1. Whisper-tiny 성능 평가 결과>

구분	Precision	Recall	F1-score	Support
non-suicide	0.991	0.991	0.991	4,376
suicide	0.959	0.957	0.958	937
Accuracy	-	-	0.985	5,313
Macro avg	0.975	0.974	0.975	5,313
Weighted avg	0.985	0.985	0.985	5,313

동일 환경에서 Wav2Vec2.0 을 사용한 경우 전체 정확도 93.90%를 보였고, 자살 클래스에 대해 F1-score 0.842, 민감도 0.923 을 기록하여 비교적 낮은 탐지 성능을 보였다.

<표 3-2. Wav2Vec2.0 성능 평가 결과>

구분	Precision	Recall	F1-score	Support
non-suicide	0.983	0.942	0.962	4,376
suicide	0.774	0.923	0.842	937
Accuracy	-	-	0.939	5,313
Macro avg	0.879	0.933	0.902	5,313
Weighted avg	0.946	0.939	0.941	5,313



<그림 1. Confusion Matrix 분석 결과>

Confusion Matrix 분석 결과, Whisper-tiny 는 자살 발화를 대부분 정확히 탐지하며 안정적인 분류 성능을 유지한 반면, Wav2Vec2.0 역시 자살 발화 937 건 중 865 건을 탐지하며 실용적인 수준의 성능을 확보하였으나, 비자살 발화에 대해 252 건의 오탐(False Positive)이 발생하여 정밀도에서 Whispe- tiny 에 비해 다소 낮은 수치를 보였다.

IV. 결론 및 시사점

본 연구는 음성 기반 자살 위험 탐지에서 STT 모델의 전사 품질이 LLM 분류 성능에 직접적인 영향을 미친다는 점을 실증하였다. Whisper-tiny 기반 파이프라인은 높은 전사 정확도로 인해 자살 발화 탐지에서 우수한 성능을 보였으며, 이는 공공상담센터, 군부대, 교육기관 등 사각지대에서 조기 개입 도구로 활용될 수 있는 가능성을 시사한다. 기술적 성능을 넘어 심리적 낙인 해소와 접근성 개선에도 기여할 수 있다는 점에서 사회적 가치가 크다. 향후 연구에서는 다중 클래스 분류, 설명 가능한 AI, 음성 품질 보정 기술을 포함한 고도화를 통해 보다 정교한 자살 예방 시스템으로의 확장이 필요하다. 이러한 접근은 자살 문제에 대한 통합적이고 실질적인 해결책을 제시하는 기반이 될 수 있다.

참 고 문 헌

[1] 통계청, “사망원인통계 결과,” 통계청 보도자료, 2023.

[2] 장규현, “상담자들의 심리상담 챗봇에 대한 인식,” *연세대학교 교육대학원 석사학위논문*, 2021.

[3] D. Shin, “음성과 텍스트를 이용하여 우울증 및 자살 위험을 평가하는 인공지능 기반 임상의사결정지원시스템에 관한 연구,” 박사학위논문, 서울대학교, 2022.

[4] S. Y. Min, “음성 분석을 이용한 인공지능 기반 자살 위험군 선별 및 모니터링,” 의학박사학위논문, 서울대학교, 2024.

[5] D. Lee, “Predicting Suicidality with Explainable Deep Learning Models,” Ph.D. Dissertation, 성균관대학교, 2024.