

감정분석 기반의 상담용 생성형 AI모델

김석준[†], 이병문^{††*}

Generative AI Model for Counseling based on Emotion Analysis

Seok Jun Kim[†], Byung Mun Lee^{††*}

ABSTRACT

The demand for psychological counseling among students is increasing, but the number of counselors is limited and it is difficult to satisfy the demand. AI-based counseling attempts to solve this shortage, but it is difficult to counsel with directive-based models. There are also models based on deep learning, but they do not consider the connection with the counselor and have the limitation that the entire counseling must be viewed again. In this paper, we propose a generative AI model for counseling based on emotion analysis. This model conducts psychological counseling and emotion analysis, and summarizes the conversation after the end of the counseling, so it is easy to understand the entire counseling. To achieve this goal, the model consists of a Solar-based counseling AI, a summary AI, and an ELECTRA-based analysis AI model. Then, using the collected data, we use various fine-tuning techniques to specialize the models for counseling. In the experiment to validate the model, the analysis AI model showed an F1 Score of about 0.69 on 6,640 sentiment data, and the counseling AI, summary AI model was validated by G-Eval Framework. Thus, we demonstrated the validity of the proposed model.

Key words: Generative AI, Counseling, Artificial Intelligence, Emotion Analysis

1. 서 론

오늘날, 학교폭력, 학업 스트레스, 대인관계 등으로 학생들의 심리상담 수요는 꾸준히 증가하고 있다. 그래서 이러한 수요를 만족시키기 위해 전문 상담사가 청소년 상담센터에서 학생들의 심리상담을 진행하고 있다. 하지만 상담사의 수가 한정되어 있을 뿐만 아니라 지역별로 상담하는 담당 인원수가 적어 모든 학생들을 포함한 상담받고자 하는 사람들의 상담수요를 만족시키기 어렵다. 그러다 보니, 경우에 따라서는 수요가 많아서 상담사들의 피로도 역시 증가하고 있는 실정이다.

그래서 이러한 문제점을 해결할 수 있는 상시 접근가능한 AI상담에 대한 연구가 이루어지고 있다 [1]. AI상담은 스마트폰 앱과 같이 온라인으로 접근할 수 있어 사용자의 수요를 쉽게 만족시킬 수 있다. 하지만 명령어나 키워드 기반의 단순한 AI상담은 복잡하거나 키워드에 없는 내용을 답변할 수 없고 일상적인 대화방식으로 진행할 수 없다는 한계가 있다.

그래서 최근에는 심리상담에 NLP(Natural Language Processing)기술을 이용한 AI를 도입하려는 시도가 있다. NLP기반의 AI는 문장의 맥락에 따라 답변이 달라지므로 맥락이 중요한 상담에서 사용하기 적합하다. 그래서 사용자의 문장을 입력받아 우

* Corresponding Author : Byung Mun Lee, Address: (13120) 1342 Seongnamdaero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Korea, TEL : +82-31-750-4756, FAX : +82-31-750-4756, E-mail : bmlee@gachon.ac.kr
Receipt date : Mar. 4, 2024, Revision date : Apr. 25, 2024
Approval date : May 24, 2024

[†] Dept. of IT Convergence Engineering, Graduate School, Gachon University
(E-mail : bnb1324@naver.com)

^{††} Dept. of Computer Engineering, College of IT Convergence, Gachon University

* This work was supported by the Gachon University research fund of 2024.(GCU-202404280001).

울, 스트레스, 식욕부진, 자살충동을 예측할 수 있는 심리상담 보조모델로 사용할 수도 있고, 상담내역을 입력받아 실시간 자살위험 지수를 예측하는 분석사례도 있다[2,3]. 하지만 이러한 분석사례는 AI가 직접 상담을 진행하는 것이 아니기 때문에 심리상담 보조에 있어 한계가 있다.

그래서 AI만을 사용하여 심리상담을 진행한 “Woebot”의 사례도 있다[4]. 이 사례에서는 2주 동안 상담 전문가가 개입하지 않고, “Woebot”을 통해서만 상담을 진행하여 우울증과 불안증세가 감소한 경우를 보여주고 있다. 그러나 상담자의 약 30%는 봇이 자연스럽게 대화하지 못했다고 느꼈다는 결과가 보고되었기 때문에 생성형 AI방식으로의 접근방법이 필요하다. 그리고 한국어어를 이용한 상담에서는 “Woebot”이 데이터 세트나 모델이 공개되지 않아, 그 상담효과를 확인할 수 없다.

AI 심리상담을 위한 한국어 기반의 챗봇 시스템에 대한 사례도 있다[5]. 이 사례에서는 AI로만 상담을 진행하며 실시간으로 사용자가 입력한 문장에 대한 긍정과 부정평가를 진행한다. 만약, 문장이 부정적으로 평가되었다면 일정한 순서를 따라 사용자의 불안 점수와 불안원인 검사를 진행한다. 하지만 사용자의 문장마다 부정적이라고 평가된다면 20개 이상의 문항에 대한 질문을 진행하는 것은 상담의 흐름을 방해할 가능성이 있다. 또한, 이렇게 생성된 데이터가 인간 상담사에게 전달될 경우에는 너무 방대해질 수 있다는 한계점도 있다.

따라서 본 논문에서는 감정분석 기반의 상담용 생성형 AI모델을 제안한다. 이 모델을 활용한다면, 한국어로 사람과 대화할 때 심리상담을 진행할 수 있을 뿐만 아니라, 그와 동시에 사람이 작성한 대화문장에서 감정분석도 가능하도록 한다. 같은 단어와 문장이더라도 감정상태에 따라서 다른 의미로 해석될 수 있기 때문에 문맥과 감정은 상담자의 심리상태를 해석하는데 매우 중요하며, 상담자에게 공감있는 답변을 만들어 낼 수 있어 보다 효과적인 상담서비스를 구현해 낼 수 있다.

이와 같이 그리고 본 논문에서 제안하는 모델이 보다 효과적으로 목적을 달성할 수 있도록 3가지 기능(분석, 상담, 요약)의 AI모델로 통합하여 구성할 필요가 있다. 분석 AI모델은 ELECTRA(Efficiently Learning an Encoder that Classifies Token Re-

placements Accurately) 모델을 사용하며, 문장을 입력받아 사람의 심리를 분석하도록 한다. 상담 AI모델은 Solar 모델을 사용하여, 사람의 문장에 대한 답변을 진행하면서 심리상담을 제공하도록 한다. 요약 AI모델은 전체상담 내용을 요약하여 상담가에게 상담소견을 모델에 피드백하여 상담효과를 높일 수 있도록 한다.

논문의 2장인 관련연구에서는 기존 심리상담 사례와 AI모델에 대해 탐구한다. 3장에서는 제안하는 모델인 감정분석 기반의 상담용 생성형 AI모델을 살펴본다. 4장에서는 제안한 AI모델의 성능을 확인하는 실험을 진행한다. 5장은 실험 결과와 함께 결론을 정리한다.

2. 관련연구

2.1 심리상담

심리상담은 상담의 대상인 내담자와 상담사 사이의 대화를 통해 심리적인 문제를 해결하는 과정이다. 그런데 심리상담을 받기 위해서는 내담자가 상담사가 있는 장소로 직접 방문하거나, 상담사의 일정에 여유가 있어야 하므로 학생과 같이 접근성이 떨어지는 경우에는 심리상담의 진행이 쉽지 않을 수 있다. 그래서 스마트폰 어플리케이션이나 웹 기반으로 시간과 위치에 상관없이 사용할 수 있는 디지털 치료제에 대한 수요가 있다[6,7]. 이러한 디지털 치료제 중에서도 AI 기반의 심리상담 사례를 연구하거나, 치료효과를 실제로 검증하려는 시도가 이루어지고 있다[4,8,9].

다음 Table 1은 AI챗봇으로 심리상담을 진행한 후, 심리분야에서 사용되고 있는 우울진단도구인 PHQ-9(Patient Health Questionnaire-9) 설문을 통해 효과를 확인하는 실험 사례들이다. Table 1의 Case 1을 보면, “Woebot”을 통해 참가자의 우울과 불안이 변화하는지 확인하는 연구로, 약 2주 동안 전체 70명의 모집단중에서 36명은 비교군으로, 34명은 “Woebot”을 사용한 실험군으로 참여한 후, PHQ-9를 설문하였더니 평균 2점 이상의 감소로 우울정도가 낮아짐을 알수있다[4].

Case 2는 “Shim”이라는 스마트폰 앱 형태의 챗봇을 사용한 실험이다. 이경우도 2주 동안 28명의 참가자를 통해 실험한 결과 유의미한 스트레스의 감소한

Table 1. Case studies about AI counseling.

Case	Metrics	Method	Duration	Participant
1	Depression, Anxiety	Woebot	2 weeks	70
2	Stress, Well-being score, Satisfaction with life scale	Shim	2 weeks	28
3	Depression, Anxiety, Ties with others	Rule based chatbot	2 weeks	30

사례이다[8]. 하지만 Case 1과 Case 2의 AI는 한국어 기반의 AI가 아니기 때문에 한국어를 사용하여 실험한다면 AI의 성능이 감소할 수 있으며, 그에 따라 스트레스와 같은 부정적인 요소의 경감 효과가 달라질 수 있다.

Case 3은 한국어 규칙 기반의 AI챗봇을 바탕으로 실험을 진행한 사례로, 2주 동안의 실험에서 60세 이상 노인 30명을 대상으로 우울, 불안, 삶의 질, 사회적 지지라는 4가지 요소의 변화 여부를 평가한 사례이다. 그리고 각 항목별로 평균 5점 이상의 점수가 개선되어, 유의미한 정신건강 개선이 있었음을 확인한 사례이다[9].

이와같이 AI기반의 심리상담은 유의미한 심리적 개선효과가 있다. 그러나 참가자의 설문조사를 통해 그 효과를 평가하기 때문에 자연스러운 대화과정에서 심리상태나 감정상태를 파악하기 어렵다. 다행히도 최근 생성형 AI기술 발전으로 개발된 Solar와 같은 AI모델을 심리상담에 도입할 수 있다면 기존의 상담모델의 한계를 개선할 수 있다.

2.2 Solar 모델

Solar 모델은 32개의 계층을 갖는 Llama2 모델을 기본구조로 Mistral 7B 모델의 사전학습된 가중치를 사용하여 만들어진 디코더 기반의 생성형 AI모델이다[10]. 특히 DUS(Depth Up-Scaling) 방식을 사용하면 모델을 효율적으로 확장할 수 있다. 예를 들면, 기본모델인 32개의 계층을 8개와 24개로 나누고 이를 8개만 중첩시킨후 복제하면 결국 24개와 24개가 만들어지게 되고 결국 32개를 48개의 계층으로 확장할 수 있는 방식이다. 하지만 계층이 확장되면서 성능이 감소할 수 있기 때문에, 추가적인 사전학습이 필요하다. 이 사전학습이 끝난 Solar 모델은 비슷한 파라미터 수를 가지는 생성형 AI모델과 비교하여 더 우수한 성능을 보인다.

높은 성능을 보이는 생성형 AI모델의 선택은 상담 대상자가 AI기반의 상담을 더 자연스럽게 느낄 수 있다는 점에서 중요하다. 그래서 비슷한 크기의 다른 모델보다 성능이 좋은 Solar 모델을 선택하면 맥락을 이해하고 문장을 생성하는 능력이 우수하므로 복잡한 대화능력이 요구되는 업무에 적합하다. 그러나 기존의 Solar 모델은 영어 데이터 세트로 사전학습을 진행해서 한국어 입력에 대한 성능이 부족할 수 있기 때문에 한국어 데이터를 사용하여 사전학습을 진행한 “Open-Solar-Ko-10.7B” 모델을 사용할 필요가 있다[11].

Open-Solar-Ko-10.7B 모델은 Solar 모델과 동일한 구조를 가진 모델을 영어 데이터 세트 대신 공개된 한국어 데이터 세트를 통해 사전학습을 시켜 한국어에 대한 성능을 높인 모델이다. Solar 모델과 구조가 동일하면서, 사전학습에 사용한 데이터 세트만 다르기 때문에 원본 모델과 유사한 성능을 기대할 수 있어 한국어 상담을 진행할 때 적합하다. 그러나 그럼에도 불구하고 대화과정에서 감정을 분석하여 어감을 추정해야 하는 경우에는 좋은 성능을 기대할 수 없다. 따라서 이를 보조할 ELECTRA와 같은 인코더 기반의 모델이 필요하며, 이를 사용한다면 감정 분석에 보다 효과적일 수 있다.

2.3 ELECTRA 모델

ELECTRA 모델은 인코더 기반의 AI모델로 입력된 문장을 이해하는데 최적화되어 주어진 문장에 대해 감정판별이나 상태분류에 높은 성능을 보이는 모델이다[12]. 이 모델은 인코더 층을 사용한다는 점에서 BERT(Bidirectional Encoder Representations from Transformers) 모델과는 비슷하지만, 학습 방식에 있어서 토큰을 마스킹하는 방식 대신 RTD(Replaced Token Detection)를 사용하므로 BERT 모델보다 학습효율이 높다.

RTD는 모델을 생성모델과 판별모델로 구분하여 학습하는 방식이다. 생성모델은 입력받은 문장중 일부를 조작된 단어로 대체하는 모델로, 판별모델보다 크기가 작다. 판별모델은 생성모델이 생성한 문장의 모든 단어에 대하여 각 단어가 원본인지, 조작된 단어인지를 판단한다. 이러한 방식으로 모든 단어에 대하여 학습할 수 있어 15% 미만의 단어만 학습될 수 있는 BERT에 비해 학습능력이 우수하다.

그런데 기존 ELECTRA 모델은 대소문자를 구분하지 않고 영어 데이터 세트를 이용해서 사전학습하기 때문에 한국어 문장을 분석할 때 그 성능이 감소된다. 따라서 이를 보완한 KoELECTRA와 같은 개선된 모델이 필요하다[13]. KoELECTRA와 같은 인코더 기반의 모델은 목적에 맞는 분류기 계층을 추가하여 파인튜닝을 진행해야 사용할 수 있다. 예를 들어, 주어진 문장에 대하여 이진분류를 해야한다면 분류기 계층은 시그모이드 함수가 포함된 분류기를 연결해야 한다. 문장에 대한 감정분석은 여러 감정 중 하나를 고르는 다중 분류이기 때문에 소프트맥스 함수가 포함된 분류기를 연결하여 파인튜닝을 진행해야 한다. 따라서 상담은 상담자와 상담자간의 상호대화를 통해서 이루어지기 때문에 상담내용의 문맥 파악과 감정분석이 동시에 필요하다. 이 두 가지가 상호협력을 하게 될 경우 좋은 상담효과를 기대할 수 있어, 이에 기반한 모델을 제안하고자 한다.

3. 감정분석 기반의 상담용 생성형 AI모델

3.1 개요

본 논문에서는 감정분석 기반의 상담용 생성형 AI 모델을 세가지 종류의 AI모델로 통합하여 제안한다. 첫 번째는 상담 AI모델로, 사람이 입력한 문장의 맥락을 파악하고 상담을 진행하는 역할을 한다. 두 번째는 분석 AI모델로, 사람의 문장으로부터 감정을 분석하고 6개 감정(분노, 슬픔, 불안, 상처, 당황, 기쁨)중에서 하나로 분류한다. 세 번째 모델인 요약 AI 모델은 상담이 끝나면, 전체 상담을 요약한다.

이러한 세가지 모델을 통해 Fig. 1과 같이 내담자는 AI모델에게 자신의 고민이나 질문을 물어보고 등 상담받을 수 있는데, 그 전체과정에서는 사람이 개입하지 않고 AI모델과 상담하는 것이다. 대화과정에서 수집한 데이터는 대상의 현재 감정 상태를 분석하거나, 또는 상담이 끝나고 요약본을 통해 전체 맥락을 인간 상담사에게 제공할 수 있다. 그래서 상담사가 개인에게 최적화된 상담 의견을 상담 AI모델에게 입력하는 것으로 개인별 최적화된 상담을 제공할 수도 있다.

이러한 상담과정을 제공하려면 Fig. 2에서처럼 상담용 AI모델을 생성하여야 한다. 그 생성과정을 살펴보면 먼저 원본데이터를 수집한다. 원본데이터는 KoAlpaca 데이터 세트와 감성 대화 말뭉치 데이터 세트를 사용한다. 각 데이터는 서로 다른 AI모델을 학습시키기 위해 사용하며, 적절한 전처리 과정을 거쳐 모델을 파인튜닝하는데 사용한다.

전처리 과정이 끝난 후 사전학습된 모델들(Open-Solar-Ko, KoELECTRA)을 파인튜닝하여야 한다. 파인튜닝에 사용할 원본 모델들은 Huggingface를 통해 확보할 수 있다[11,13]. Solar 모델은 명령어 데이

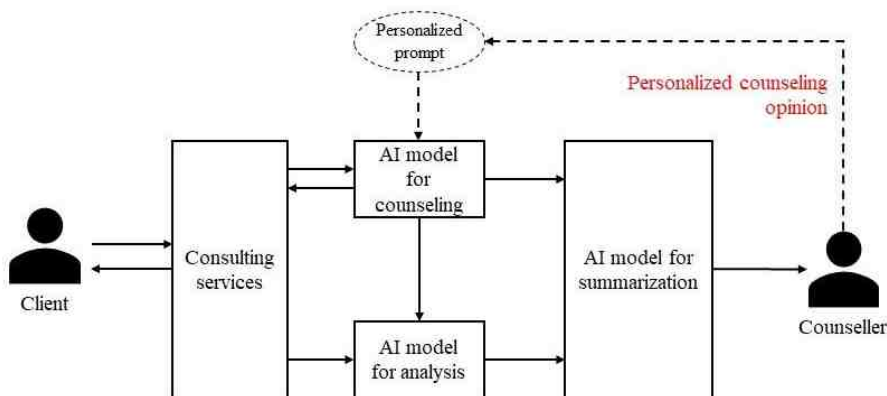


Fig. 1. Counseling process with generative AI model for counseling based on emotion analysis.

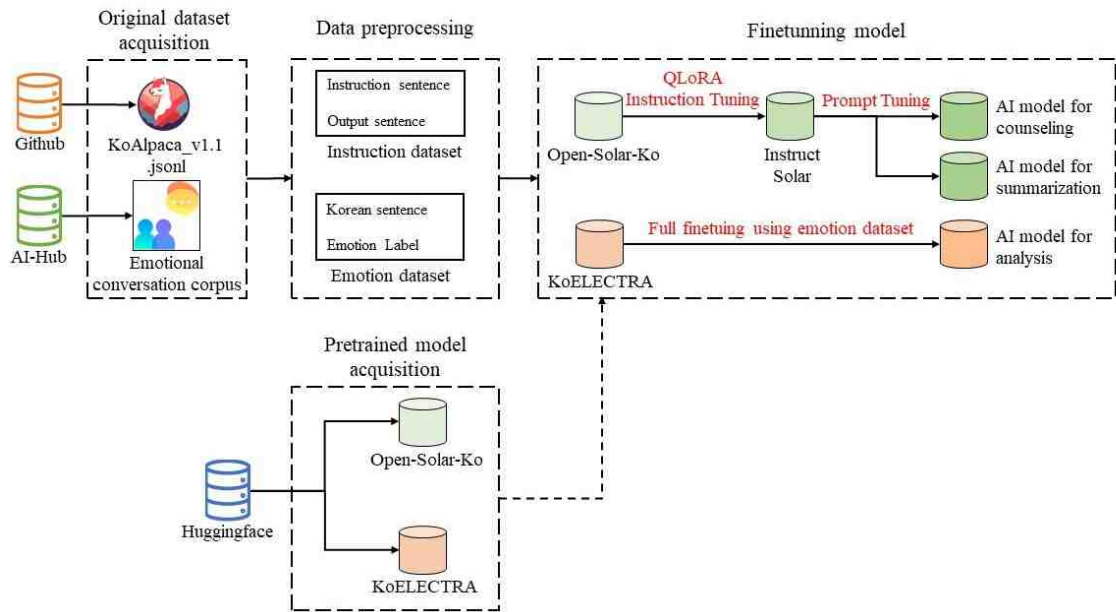


Fig. 2. The overview of the generative AI model for counseling based on emotion analysis.

터 세트를 통해 튜닝하는 것으로 주어진 명령을 인지하고 수행하는 모델로 만든다. 이때, 시스템 리소스를 고려하여 일부 가중치만 수정하는 QLoRA (Quantized Low Rank Adapters) 방식으로 명령어 튜닝을 진행한다[14,15]. 명령어 튜닝이 끝난 Solar 모델은 입력한 명령을 따라 적절한 출력값을 생성하는 모델이 된다. 그래서 학습 없이 모델이 이해할 수 있는 명령어를 제시하는 프롬프트 튜닝을 서로 다르게 적용하는 것으로 요약과 상담 모델을 만들 수 있다.

프롬프트 튜닝은 모델에 입력하는 프롬프트를 수정하여 원하는 결과를 출력하는 방식이므로 내담자와 AI모델간의 상담요약과 내용을 보고 상담소견을 AI모델로 피드백하여 상담효과를 배가시키도록 한다. 이것은 상담을 개인별로 차별화를 할 수 있다는 장점이 있다. ELECTRA 모델은 전체 가중치를 수정하는 풀 파인튜닝을 진행하는 것으로 감정 분석 AI 모델을 만든다. 이러한 파인튜닝은 파인튜닝의 방식도 중요하지만, 데이터도 성능에 영향을 많이 끼치기

때문에 원본데이터의 확보도 중요하다.

3.2 데이터 수집

AI모델마다 구조와 목적이 다르기 때문에 서로 다른 데이터를 수집해야 할 필요가 있다. Table 2와 같이, 상담모델에서 사용할 수 있도록 모델을 튜닝하기 위해 오픈소스 데이터 세트를 사용한다. 첫 번째는 KoELECTRA를 튜닝하기 위한 KoAlpaca 데이터 세트이고, 두 번째는 Open-Solar-Ko를 튜닝할 수 있는 “감성 대화 말뭉치” 데이터 세트다[16,17].

“KoAlpaca” 데이터 세트는 명령과 명령에 대한 응답으로 구성하며, Github에 업로드된 파일을 다운로드하여 확보할 수 있다[16]. 이 데이터 세트는 네이버 지식인에 등록된 질문과 답변을 ChatGPT로 이를 재구성하여 JSON Lines의 형태로 저장한 것이다. 하나의 데이터는 “instruction”(질문), “output”(답변), “url”(출처링크)의 세 가지로 구성된다.

“감성 대화 말뭉치” 데이터 세트는 AI-Hub에서

Table 2. The original dataset for AI model.

Purpose	Model	Dataset	Amount	Source
Instruct-Following	Open-Solar-Ko	KoAlpaca Dataset (v1.1)	21,155	Github[16]
Emotion analysis	KoELECTRA	Emotional conversation corpus	270,000	AI Hub[17]

제공하는 데이터 세트로, 사용 요청을 하는 것으로 확보할 수 있다. 이 데이터 세트에는 사람과 AI 간 대화한 내용이 저장되어 있다[17]. 한 개의 데이터에는 사람과 AI가 각각 한번씩 번갈아 작성한 문장이 최소 4개부터 최대 8개 까지 포함될 수 있다. 이때, 각 데이터 별 사람의 첫 번째 문장에 해당하는 감정도 같이 라벨링되어 있다. 이러한 감정은 모두 60개가 있으며, 6개의 대분류와 각 대분류별로 10개의 소분류로 구분한다.

다음 Fig. 3과 Fig. 4는 각각 KoAlpaca 데이터 세트와 감성 대화 말뭉치 데이터 세트의 예시다. Fig 3을 보면, instruction에서 양파와 고구마에 대한 질문을 볼 수 있으며, 질문에 따른 답변인 output에서 그에 맞는 답변을 하는 것을 확인할 수 있다.

Fig. 4와 같이, 감성 대화 말뭉치 데이터 세트는 대화를 하는 인간과 AI의 정보인 “profile”과 대화 내

용인 “talk”로 구성된다. 이 중 profile 내에 있는 emotion.type이 첫 번째 문장에 맞는 감정을 의미한다. Fig. 4의 type에서 “E31”은 “불안”(대분류 명)과 “두려운”(소분류 명)에 해당한다.

그런데 수집한 원본데이터에는 모델을 튜닝하는데 필요하지 않은 데이터도 포함되어 있어 불필요한 데이터를 제거하도록 전처리 과정을 거쳐야 한다.

3.3 데이터 전처리

데이터 전처리과정에서는 수집한 원본데이터에서 튜닝에 필요없는 데이터를 제거한다. 먼저, KoAlpaca 데이터 세트에서는 데이터의 출처를 나타내는 “url” 값을 삭제한다. 이 값은 수집한 데이터의 출처를 나타내는 하이퍼링크로 학습에서 사용하지 않는다. 따라서, 가공된 데이터의 형태는 Fig. 5에서처럼 “instruction”과 “output”의 두가지로만 이루어

```
{
  "instruction": "양파는 어떤 식물 부위인가요? 그리고 고구마는 뿌리인가요?",
  "output": "양파는 잎이 아닌 식물의 줄기 부분입니다. 고구마는 식물의 뿌리 부분입니다. \n\n식물의 부위의 구분에 대해 궁금해하는 분이라면 분명 이 질문에 대한 답을 찾고 있을 것입니다. 양파는 잎이 아닌 줄기 부분입니다. 고구마는 다른 질문과 답변에서 언급된 것과 같이 뿌리 부분입니다. 따라서, 양파는 식물의 줄기 부분이 되고, 고구마는 식물의 뿌리 부분입니다.\n\n 덧붙이는 답변: 고구마 줄기도 볶아먹을 수 있나요? \n\n고구마 줄기도 식용으로 볶아먹을 수 있습니다. 하지만 줄기 뿐만 아니라, 잎, 씨, 뿌리까지 모든 부위가 식용으로 활용되기도 합니다. 다만, 한국에서는 일반적으로 뿌리 부분인 고구마를 주로 먹습니다.",
  "url": "https://kin.naver.com/qna/detail.naver?d1id=11&dirId=1116&docId=55320268"
}
```

Fig. 3. Example of the KoAlpaca dataset.

```
{
  "profile": {
    "persona-id": "Pro_03802",
    "persona": {"persona-id": "A02_G01_C01", "human": ["A02", "G01"], "computer": ["C01"]},
    "emotion": {"emotion-id": "S06_D02_E31", "type": "E31", "situation": ["S06", "D02"]}
  },
  "talk": {
    "id": {"profile-id": "Pro_03802", "talk-id": "Pro_03802_00007"},
    "content": {
      "HS01": "이번 프로젝트에서 발표를 하는데 내가 실수하는 바람에 우리 팀이 감점을 받았어. 너무 미안해.",
      "SS01": "실수하시다니 정말 미안한 마음이 크겠어요.",
      "HS02": "내 능력이 부족한 거 같은데 그만 다녀야 될거같아.",
      "SS02": "능력을 올리려면 어떤 방법이 있을까요?",
      "HS03": "퇴근 후 여가에 회사 일을 더 열심히 해서 피해가 가지 않도록 해야겠어.",
      "SS03": "꼭 좋은 결과 있길 바라요."
    }
  }
}
```

Fig. 4. Example of the emotional conversation corpus.


```

{
  "instruction": "양파는 어떤 식물 부위인가요? 그리고 고구마는 뿌리인가요?",
  "output": "양파는 잎이 아닌 식물의 줄기 부분입니다. 고구마는 식물의 뿌리 부분입니다. \n\n식물의 부위의  
구분에 대해 궁금해하는 분이시면 분명 이 질문에 대한 답을 찾고 있을 것입니다. 양파는 잎이 아닌 줄기 부  
분입니다. 고구마는 다른 질문과 답변에서 언급된 것과 같이 뿌리 부분입니다. 따라서, 양파는 식물의 줄기  
부분이 되고, 고구마는 식물의 뿌리 부분입니다.\n\n덧붙이는 답변: 고구마 줄기도 볶아먹을 수 있나요?  
\n\n고구마 줄기도 식용으로 볶아먹을 수 있습니다. 하지만 줄기 뿐만 아니라, 잎, 씨, 뿌리까지 모든 부위가  
식용으로 활용되기도 합니다. 다만, 한국에서는 일반적으로 뿌리 부분인 고구마를 주로 먹습니다.",
}

```

Fig. 5. Example of preprocessed KoAlpaca dataset.

저 있으며 이를 KoELECTRA 모델의 튜닝 과정에서 사용한다.

감성 대화 말뭉치 데이터 세트는 감정을 분석할 문장과 문장에 해당하는 감정의 두가지 데이터가 필요하다. 따라서 이 구조에 맞게 데이터를 가공하여야 하는데, 먼저 문장에 해당하는 감정을 추출하기 위해서 데이터에서 “profile.emotion.type”의 값을 추출한다. 이 값은 E01, E22와 같이 E와 숫자 두 개가 있는 형태로 하나의 감정을 나타낸다.

Table 3은 이러한 데이터 세트의 감정 분류기준으로 가로축은 일의 자리 숫자를 결정하고, 세로축은 십의 자리 숫자를 결정한다. 예를 들어, 불안(“anxiety”)은 E30에, 자신감(“confident”)는 E69에 해당한다. 그런데, 소분류에 따라 분류하면 감정이 모두 60개가 되어 분류정확도가 떨어질 수 있다. 따라서, 대분류를 기준으로 분노(E10), 슬픔(E20), 불안(E30), 상처(E40), 당황(E50), 기쁨(E60)의 6가지 감정으로 병합하고 각각 0, 1, 2, 3, 4, 5의 범주형 값으로 치환한다.

두 번째로 Fig. 4에서 보면 “talk.content” 내의 문장을 추출하는 것으로 감정에 해당하는 문장을 수집한다. 이때, “HS01”이라는 사람의 첫 번째 문장만 추출하여 사용한다. “감성 대화 말뭉치” 데이터 세트에서는 사람의 첫 번째 문장에 따라 감정을 태깅하기 때문이다.

다음 Fig. 6은 전처리가 끝난 감성 대화 말뭉치 데이터 세트의 예시로 Fig. 6을 보면, 감정과 문장으로 구성한다. 이 데이터에서는 “emotion”의 값이 2이므로, 불안 감정에 해당한다. 이렇게 두 데이터의 전처리 과정이 끝나면 모델의 튜닝 과정에서 이 데이터 세트들을 통해 AI모델의 파인튜닝을 진행한다.

3.4 모델 파인튜닝

모델의 파인튜닝 과정에서는 사전학습된 Open-Solar-Ko 모델과 KoELECTRA 모델을 튜닝하며, NVIDIA의 RTX A5000 24GB GPU가 탑재된 서버를 통해 모든 튜닝을 진행한다. Solar모델은 두가지

Table 3. Emotion classification criterion in emotional conversation corpus.

E	1X	2X	3X	4X	5X	6X
X0	anger	sadness	anxiety	wound	panic	happiness
X1	grunting	disappointed	scared	jealous	embarrassed	grateful
X2	frustrated	heartbroken	stressed	betrayed	conscious	trusted
X3	irritating	regrettable	weak-minded	isolated	lonely	comfortable
X4	defensive	gloomy	confused	shocked	inferior	satisfying
X5	malicious	paralysed	perplexed	disadvantaged	guilty	excited
X6	fretting	pessimistic	skeptical	victimized	shameful	easygoing
X7	nauseate	tearful	worrisome	unfair	disgusting	relieved
X8	angry	discouraged	careful	distressed	pathetic	thrilled
X9	annoying	disillusioned	nervous	abandoned	confused	confident

```
{
  "emotion" : 2,
  "sentence" : "이번 프로젝트에서 발표를 하는데 내가 실수하는 바람에 우리 팀이 감점을 받았어. 너무
  미안해."
}
```

Fig. 6. Example of preprocessed emotional conversation corpus dataset.

방식을 사용하여 튜닝하는데, 첫 번째로 QLoRA 방식을 사용한 명령어 튜닝을 한다. 그리고 이 모델을 프롬프트 튜닝을 통해 모델별로 개별 프롬프트를 구성하는 것으로 상담AI모델과 요약AI모델을 만들 수 있다.

먼저, 사전학습이 끝난 Open-Solar-Ko 모델을 명령어 데이터 세트를 사용하여 QLoRA 방식으로 튜닝하여 명령어 튜닝을 진행한다. Solar 모델은 10.3억 개의 파라미터를 가지고 있기 때문에 24GB VRAM의 단일 그래픽카드로는 파인튜닝을 진행하기 어렵다. 그래서 QLoRA의 4비트 양자화와 LoRA 어댑터를 통해 모델의 튜닝에 요구되는 메모리를 줄여 튜닝을 진행해야 한다.

다음 Table 4는 Solar 모델의 QLoRA 방식의 튜닝에 사용한 하이퍼파라미터를 나탄낸 것이다. 본 논문에서는 튜닝에서 사용하는 파라미터로 양자화 방식과 LoRA 방식을 적용한다. 이때, 양자화 관련 설정으로는 양자화 타입과 정확도 계산에 사용하는 타입의 두가지가 있는데, 각각 NF4(Normalized Float 4 bit)와 float16로 설정한다. LoRA 설정에서는 크게 Rank와 Alpha가 중요하다. 이러한 Rank와 Alpha 값에 따라 모델의 학습시킬 파라미터 양이 달라지기 때문이다. 명령어 튜닝에서는 명령에 따른 답변을 한다는 스타일을 학습시켜야 하므로 Rank와 Alpha 값을 낮게 설정할 필요가 있다. 따라서 각각 8과 32를 사용한다.

명령어 튜닝이 끝난 모델은 주어진 명령을 따르는 Instruction-following 모델이 되어, 학습하지 않은

방식의 명령에도 유의미한 출력값을 만들 수 있다. 그래서 명령어 튜닝이 끝난 모델을 다시 프롬프트 튜닝을 통해 주어진 프롬프트에 따라 학생의 상담이나 전체 상담의 요약물 진행할 수 있는 모델로 만들 수 있다.

프롬프트 튜닝은 모델이 이해할 수 있는 명확한 문장을 통해 원하는 형태의 출력을 얻는 방식으로 “다음 내용에 대하여 답변하세요.”라는 프롬프트를 입력한다면 모델은 학생의 질문에 답변하는 상담 AI 모델이 될 수 있다. 반대로, 프롬프트를 “다음 내용을 요약하세요.”라고 입력한다면 요약 AI모델이 될 수도 있다.

프롬프트 튜닝은 프롬프트 작성만으로 결과에 변화를 줄 수 있기 때문에 파인튜닝에 비하면 데이터나 시간이 적게 사용된다. 뿐만 아니라 이러한 인간 상담사의 상담 소견을 명령어에 같이 입력하여 개인별로 특화된 상담모델을 만든다. 다만, 프롬프트에 따라 답변이 달라질 수 있으므로 최적의 프롬프트를 작성하는 것이 중요하다.

Fig. 7은 상담 AI모델에 사용하는 프롬프트로, 상담 대상이 입력한 문장에 대응할 수 있도록 작성한다. 이때, “### 내용 :”에는 상담 대상이 문장을 입력한다고 가정한다. 프롬프트를 보면 AI모델이 생성할 수 있는 답변에 모두 8개의 제약조건을 제시하고, 이에 따라 답변을 작성하도록 제시한다. 예를들어, 5번 조건인 “답변에는 하이퍼링크와 같은 출처를 포함하면 안됩니다.”는 모델이 환각에 의해 정확하지 않은 하이퍼링크를 작성하는 문제를 방지할 수 있다.

Table 4. Hyperparameters for Solar model fine tuning.

Quantization		LoRA				
Type	Precision type	Rank	Alpha	Dropout	Batch Size	Micro Batch Size
NF4	float16	8	32	0.05	128	8
Others						
Optimizer		Learning Rate		Epochs		
adamw_torch		3e-4		3		

다음 내용에 대한 답변을 생성해 주세요.
다음 요구 사항을 준수해야 합니다.

요구 사항은 다음과 같습니다:

1. 당신의 역할은 AI 상담사입니다. 상대방의 대화에 최대한 공감해 주세요.
2. 상대방은 학생입니다.
3. 답변은 한국어로 작성해야 합니다.
4. 답변은 최대 8문장을 초과하면 안됩니다.
5. 답변에는 하이퍼링크와 같은 출처를 포함하면 안됩니다.
6. 답변 외에는 어떠한 추가 설명을 작성하면 안됩니다.
7. 만약 상대방이 질문했다면, 질문에 대한 적절한 응답을 작성해야 합니다.
8. 상대방이 자신의 이름을 밝히지 않았다면, 이름이 "익명"이라고 가정하고 응답을 작성하세요.

내용 :

답변 :

Fig. 7. Prompt using in AI model for counseling.

다음 내용에 대한 요약을 생성하세요.
다음 요구 사항을 준수해야 합니다.

요구 사항은 다음과 같습니다.

1. 내용의 구성은 사람이름:사람이 말한내용 으로 되어 있습니다. 구성을 참조해야 합니다.
2. 요약은 사람 이름을 포함하여 작성해야 합니다.
3. 요약은 4문장 이내로 작성해야 합니다.
4. 요약은 과거형으로 작성해야 합니다.
5. 요약을 제외한 다른 문장을 작성하지 마세요.
6. 요약에는 미래에 대한 예측을 작성하지 마세요.

내용 :

요약 :

Fig. 8. Prompt using in AI model for summarization.

Fig. 8은 상담요약을 위해 작성한 프롬프트로 주어진 내용에 대하여 요약을 진행하도록 AI에게 설명하고 있다. 주어지는 내용은 “사람이름: 작성한 내용”으로 가정하며 AI가 요약을 작성하는데 있어서 요약이 원본 상담내용보다 길어지지 않도록 길이에 대한 제약을 추가한다. 그리고 요약에 발화자가 포함되지 않아 어떤 사람이 어떤 말을 했는지 알 수 없는 상황을 방지하기 위해, 사람의 이름을 같이 포함하도록 작성한다.

모든 프롬프트 튜닝이 끝나고 나면, KoELECTRA 모델의 파인튜닝을 진행한다. 이 모델에서는 전체 가

중치를 모두 수정하는 풀 파인튜닝을 진행하며, 이 과정이 끝나면 입력된 문장에 대하여 6가지 감정(분노, 슬픔, 불안, 상처, 당황, 기쁨) 중 한가지로 분류하는 분석 AI모델을 만들어야 한다. 다음 Table 5는 KoELECTRA 모델의 파인튜닝에 사용한 하이퍼파라미터다. 사용한 모델을 보면 Base-v3-Discriminator 모델을 사용하는데, 이 모델은 기존 v1, v2 모델에 비해 사전학습에 사용된 데이터의 수가 20GB로 6GB가 더 많다. 파인튜닝 설정에 있어서 에포크는 4회로 진행하며, 학습률은 0.00003을 사용한다. 이렇게 파인튜닝이 모두 끝난 KoELECTRA 모델

Table 5. Hyperparameters for KoELECTRA model fine tuning.

Model	Batch Size	Epoch	Learning Rate	Training	Validation	Test
KoELECTRA-Base-v3-Discriminator	16	4	3e-5	41,302	10,326	6,640

은 분석 AI모델로써 학생의 문장으로부터 감정을 분류할 수 있게 된다.

4. 실험 및 평가

본 논문에서 제안하는 AI모델의 성능을 확인하기 위해 상담, 분석, 요약의 세가지 AI모델에 대하여 성능을 검증하는 실험을 하였다. 그런데, 상담과 요약 AI는 각각 상담과 상담 전체 내용에 대한 요약을 만들기 때문에 직접적인 평가가 어렵다. 따라서, Fig. 9와 같이 자유로운 평가 지표를 설정할 수 있도록 ChatGPT를 사용하는 평가방식인 G-Eval 프레임워크를 통해 모델을 평가하였다. 그리고 평가용 데이터 세트로 감성 대화 말뭉치와 한국어 대화 요약 데이터 세트를 사용하였다[17,18,19]. ELECTRA 기반의 분석 AI는 앞서 분할한 감정 데이터 세트의 테스트 데이터를 사용하여 분류 능력을 평가했다.

4.1 실험환경

다음 Table 6은 실험을 진행하는데 있어 사용한 서버의 사양과 서버 내에 설치한 소프트웨어의 버전이다. Table 6을 보면, 실험에서 NVIDIA의 RTX A5000 24GB GPU를 사용하였다.

상담, 요약 AI모델의 성능평가 실험에서는 감성 대화 말뭉치와 한국어 대화 요약 데이터 세트의 일부분을 추출하여 사용한다. 다음 Table 7은 평가에 사

Table 6. Specification of server.

System	Specification
CPU	Intel i9-13900
RAM	DDR5 128 GB
GPU	RTX A5000 24GB
OS	Rocky Linux 9.3
conda	23.3.1
pytorch	2.1.2
transformers	4.37.2
peft	0.8.2

용하는 데이터로, “감성 대화 말뭉치”와 “한국어 대화 요약”의 두가지가 있다. 감성 대화 말뭉치 데이터 세트에서는 인간이 작성한 처음 두 개의 문장을 추출하였다. 그리고 “언니가 있는데 내 친언니가 아니었어. 아니. 언니가 얘기해줬어.”와 같이 상담 AI모델에 입력할 수 있도록 하나의 문장으로 병합하였다.

한국어 대화 요약 데이터 세트는 주제별로 다양한 한국어 대화 원문이 포함되어 있는 대화 기반의 데이터 세트다. 이 데이터 세트에서 대화하는 화자의 수가 2명이면서, “개인 및 관계” 카테고리에 해당하는 데이터를 선정하였다. 그리고 대화와 대화에 따른 요약을 추출하여 이를 사용하였다.

평가에 사용하는 G-Eval 프레임워크는 ChatGPT와 프롬프트를 통해 모델의 성능을 평가할 수 있다 [19]. 이때, 평가에 사용할 평가 지표는 인간에 의해

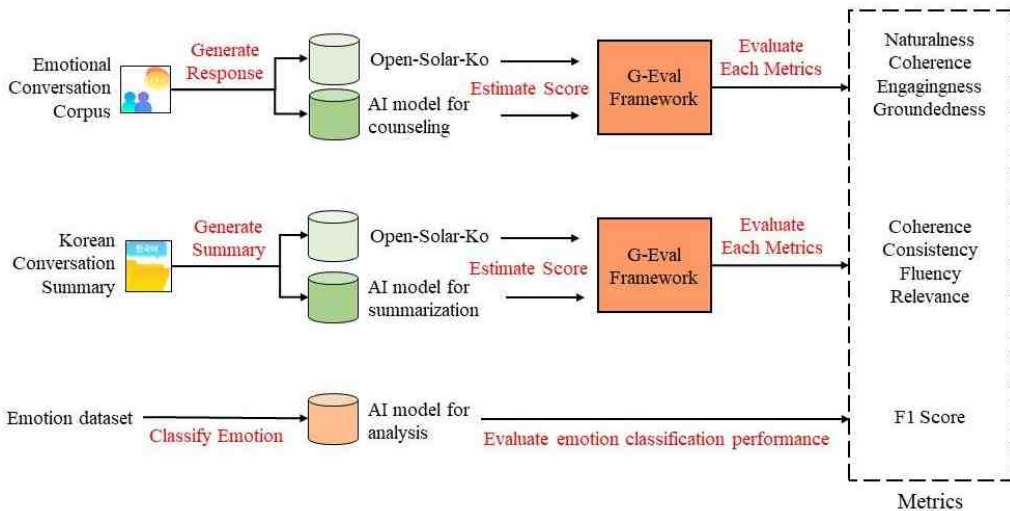


Fig. 9. The overview of the experiment.

Table 7. Datasets for experiment.

Model	Dataset	Extracted part	Amount
counseling	Emotional conversation corpus	The first two sentences written by human	100
summarization	Korean conversation summary	conversation, summary	100

선정되며, 선정된 평가지표를 ChatGPT에 입력하는 것으로 평가기준, CoT(Chain of Thoughts) 평가 단계가 포함된 프롬프트를 생성한다. 그리고 이 프롬프트를 평가할 데이터와 함께 ChatGPT 모델에 입력하는 것으로 LLM(Large Language Model) 모델에 의해 생성된 텍스트의 품질을 평가할 수 있다. 그리고 생성된 점수와 로그 확률을 사용한 점수의 가중합을 최종 평가점수로 사용하기 때문에 세밀하고 연속적인 점수를 얻을 수 있다.

G-Eval 프레임워크는 ChatGPT를 사용하여 평가하기 때문에, 평가에 사용한 ChatGPT 설정값도 중요하다. 다음 Table 8은 상담, 요약 AI모델의 평가에 사용한 ChatGPT 설정값이다. 평가 시점에서 GPT-4-Turbo 모델도 사용가능하였으나, 한글 프롬프트에 정상적인 응답이 이루어지지 않아, GPT-3.5-Turbo 모델을 사용하였다.

분석 AI모델의 성능평가에서는 학습에 사용하지 않은 테스트데이터를 사용한다. 다음 Table 9는 성능 측정에 사용한 전체 데이터 세트의 범주별 데이터 개수로, 전체 6,640개가 있다. 이 중 “상처”가 1,007개로 제일 적은 데이터이며, 제일 많은 데이터는 “분노”로 모두 1,257개가 있다.

4.2 상담, 요약 AI모델 평가

명령어와 프롬프트 튜닝이 끝난 상담과 요약 AI모델의 성능을 G-EVAL 프레임워크를 통해 평가하는 실험을 진행하였다. 이 실험은 크게 세가지 순서로 진행되었다. 첫 번째는 테스트 데이터와 그에 따른

모델의 응답을 수집한다. 두 번째로는 수집한 데이터를 프롬프트와 함께 ChatGPT에 입력한다. 마지막으로 ChatGPT로부터 반환받은 점수의 가중합을 계산하고, 전체 점수의 평균을 계산한다.

이 평가에서 원본 모델과 제안하는 모델의 차이는 튜닝 여부이기 때문에 원본모델은 요구사항을 제외한 기본적인 프롬프트를 입력하여 요약이나 상담을 진행할 수 있도록 구성하였다. 또한, 두 모델 모두 문장 생성 옵션에 있어서 repetition_penalty값을 1.5를 부여하는 것으로 문장의 반복 생성을 방지하였다.

상담 AI모델의 평가는 자연스러움, 일관성, 흥미로움, 답변의 근거 수준의 4가지 기준을 두고 평가하였다. 각각의 기준은 최소 0점부터 최대 5점까지 부여될 수 있는데, 이러한 기준에 대해서는 Fig. 10과 같이 프롬프트를 통해 GPT 3.5 모델에게 기준과 평가방식을 알려주었다.

다음 Table 10은 무작위로 추출한 100개의 평가 데이터에 대하여 ChatGPT가 평가 프롬프트에 따라 점수를 부여한 것이다. 평가 요소는 답변의 자연스러움, 일관성, 흥미로움, 신뢰도(Naturalness, Coherence, Engagingness, Groundedness)이다. 제안하는 상담 AI모델이 신뢰도와 흥미로움 부분에서 원본 모델보다 더 높은 평가를 받았다. 반대로 자연스러움과 일관성 부분에서는 비교적 낮은 평가를 받았다.

제안하는 상담 AI모델이 자연스러움에서 낮은 평가를 받은 평가를 받은 이유는 자연스러움의 평가 기준이 “답변이 인간과 비슷하게 느껴질 경우”이기 때문이다. 원본 모델은 답변이 “당신은 실패하지 않

Table 8. ChatGPT inference settings.

Model	Temperature	max_tokens	logprobs	top_logprobs	n
GPT-3.5-Turbo	0	5	True	20	1

Table 9. Test dataset to validate AI model for analysis.

Emotion	Angry	Sadness	Anxiety	Wound	Panic	Happiness
Amount	1,257	1,003	1,112	1,007	1,048	1,213

응답을 주의 깊게 읽고, 응답이 고민의 내용을 얼마나 잘 이해하고 반영하고 있는지 분석하세요.
응답이 제공하는 해결책이나 조언이 실제적이고 실행 가능한지 평가하세요.

평가 기준:

1점 (매우 낮은 근거): 응답이 고민과 전혀 관련이 없거나 맥락에 맞지 않는 해결책을 제안합니다.
2점 (낮은 근거): 응답이 고민을 부분적으로 이해하고 반영하지만, 제시된 해결책이 매우 불완전합니다.
3점 (보통 근거): 응답이 고민을 어느 정도 이해하고 있으며, 제시된 해결책이나 조언이 부분적으로 적절합니다.
4점 (높은 근거): 응답이 고민을 잘 이해하고 있으며, 대부분 실제적이고 실행 가능한 조언을 제공합니다.
5점 (매우 높은 근거): 응답이 고민을 완벽하게 이해하고, 매우 효과적이고 실행 가능한 조언을 제공합니다.

평가 단계:

1. 대화 읽기: 제공된 문장과 응답을 주의 깊게 읽으세요. 대화의 전체 맥락과 AI 응답의 내용을 이해하려고 노력하세요.
2. 응답 분석: AI의 응답을 분석하여, 고민에 대한 이해도와 제시된 해결책이나 조언이 얼마나 실제적이고 적절한지 평가하세요.
3. 점수 부여: 위의 기준에 따라 근거 있는 정도에 대한 점수를 1점부터 5점 사이에서 부여하세요.

문장:

일이 잘 마무리가 안 되어 걱정이야. 얼마 남지 않은 것 같은데 끝이 보이지 않고 지지부진해.

응답:

"걱정이 많으시군요." 하지만 너무 조급하게 생각하지 마세요! 모든 일은 시간이 필요합니다. 지금부터라도 천천히 시작해보세요. 그리고 무엇보다도 중요한 것은 자신감을 가지고 도전하는 것입니다. 실패하더라도 다시 일어나서 시도할 수 있는 용기가 중요해요. 그러니 포기하지 말고 끝까지 노력해주세요!"

평가 양식 (점수만 작성):

근거 있는 정도:

Fig. 10. Example of counsel prompt assessing groundedness.

Table 10. Evaluation results about ai model for counsel. (min = 0, max = 5)

Model	Naturalness	Coherence	Engagingness	Groundedness
Open-Solar-Ko-10.7b	3.724	3.658	4.122	3.521
AI model for counseling	3.025	2.600	4.273	3.599

있어, 당신에게는 아직 많은 시간이 남아있어!"와 같이 인간의 입장에서 작성된 반면, 제안하는 모델은 "저는 인공지능(AI)상담사로서.."와 같이 자신이 AI 상담사임을 밝히기 때문에 낮은 평가를 받는 것으로 추측하였다.

또한 이전 대화 맥락과의 연관성을 평가하는 일관성 부분에서도 낮은 점수를 부여받았는데, 평가 데이터가 질문에 대한 단발성 응답으로 구성되어 있어, 제안하는 모델이 상대방의 고민을 해결하기 위해 역으로 질문을 하는 과정에서 불필요한 질문이 많이 포함되며 점수가 낮아졌을 가능성이 있다.

예를들어, "우리 부모님은 나를 사랑하지 않는 것 같아서 마음이 아파. 항상 나보다 동생들을 더 챙기셔. 나도 사랑받고 싶어."와 같은 질문에 "먼저 여러 분께서 말씀하신 내용을 이해하기 위해 몇 가지 질문을 드리고자합니다. 첫째로 부모님께서 왜 형님을 덜 챙겨주시는 걸까요? 둘째로는 어떤 이유로 인해

그런 생각이 들게 되었을까요? 셋째는 이러한 생각 때문에 어떻게 느끼셨나요? 마지막으로 이런 상황에서 무엇을 원하시는지 궁금하네요."와 같이 답변한다면 답변의 일관성이 부족하다고 평가될 가능성이 있다.

요약 AI모델은 평가지표로 응집성, 일관성 유창성, 연관성(Coherence, Consistency, Fluency, Relevance)의 4가지 기준을 두고 평가하였으며, 평가 결과는 Table 11과 같다. Table 11을 보면, 요약 AI모델이 원본 모델과 비교하여 모든 부분에서 우수한 점수를 부여받았다. 원본 모델과 비교하여 모든 부분에서 우수한 평가를 받은 이유로는 평가에 사용한 데이터인 "한국어 대화 요약"이 뉴스와 같은 문어체로 구성된 것이 아닌 생략이나 변형이 많은 일상대화체로 구성되어 원본 모델이 요약하는데 부족한 성능을 보였을 가능성이 있다.

실험을 통해 상담 AI모델은 원본 모델과 비교하여

Table 11. Evaluation results about ai model for summarization. (min = 0, max = 5)

Model	Coherence	Consistency	Fluency	Relevance
Open-Solar-Ko-10.7b	2.708	3.039	3.869	3.571
AI model for summarization	3.281	3.657	3.956	4.005

두가지 평가지표에서 우수한 점을 보였고, 요약 AI모델은 모든 부분에서 우수한 성능을 보여준 것을 확인할 수 있었다. 따라서 상담, 요약 AI모델의 성능을 검증하였다.

4.3 분석 AI모델 평가

감정 데이터 세트로 파인튜닝을 한 분석 AI모델의 감정 분류성능을 검증하기 위해 테스트데이터를 사용하여 F1 점수를 측정하였다. Fig. 11은 분석 AI모델을 통해 분류한 테스트데이터의 F1 스코어를 그래프로 나타낸 것이다. 그래프를 보면 데이터의 수가 제일 적은 “상처”가 F1 점수가 약 0.563으로 제일 낮은 점수임을 확인하였다. 제일 높은 점수를 보인 “행복”은 약 0.931의 점수를 확인할 수 있었다. 튜닝한 모델의 전체 클래스 평균 F1 점수는 약 0.691로 분류 AI모델이 주어진 문장의 감정 분류에 있어서 유효함을 검증하였다.

그런데 각 클래스별 F1 점수의 편차가 약 0.3 정도로 큰 경우도 있었는데, 이는 모델이 학습하는데 사용한 데이터의 수가 적은 것이 영향을 끼쳤을 가능성이 있다. 본 연구에서는 데이터 세트의 전체 대화 문장을 사용하지 않고, 감정이 태깅된 사람의 첫 번째 문장만 학습데이터로 사용하였다. 이러한 방식은 실제 원본 데이터의 수(270,000)에 비해 훨씬 적은 수인 약 60,000개의 학습데이터를 구성하게 되어 학습 성

능이 낮아졌을 가능성이 높다.

5. 결 론

본 논문에서는 감정분석 기반의 상담용 생성형 AI 모델을 제안하였다. 이 모델은 한국어 기반 데이터로 사전학습이 되어있는 Open-Solar-Ko-10.7b 모델과 KoELECTRA-Base-v3-Discriminator 모델을 서로 다른 방법으로 파인튜닝을 시도하였다. 파인튜닝을 위해 학습데이터를 수집하였는데 Instruct-following 모델의 형태로 튜닝할 수 있도록, Instruction에 따른 Output 형태로 구성된 KoAlpaca 데이터 세트를 수집하였다. 또한 문장에 따라 감정을 분류할 수 있도록 AI Hub에서 “감성 대화 말뭉치” 데이터 세트를 수집하고 전처리 과정을 거쳤다.

파인튜닝 과정에서는 Solar 모델을 QLoRA 방식으로 명령어 튜닝을 진행하였다. 그리고 프롬프트 튜닝을 진행하였는데, 서로 다른 프롬프트를 입력하는 것으로 상담 AI모델과 요약 AI모델을 생성하였다. 이러한 프롬프트 튜닝은 인간 상담사의 상담자별 맞춤 지시도 입력할 수 있다면, 개인화된 상담 모델을 구성할 수도 있다. ELECTRA 모델은 다중분류가 가능하도록 분류기를 추가하고 전체 가중치를 수정하는 풀 파인튜닝을 통해 분석 AI모델을 생성하였다.

생성한 상담, 요약, 분석 AI모델을 평가하는 실험을 진행하였다. 먼저, 상담과 요약 AI모델은 Chat GPT 기반의 평가방법인 G-Eval 프레임워크를 통해 원본 모델과 비교하여 높은 평가를 받은 것을 확인하였다. 그리고 분석 AI모델은 학습에 사용하지 않은 테스트데이터를 통해 분류 성능을 평가하는 것으로 성능을 검증하였다.

본 논문에서 제안하는 감정분석 기반의 상담용 생성형 AI모델은 기존 연구와 비교하여 상담 내용을 요약하는 요약 AI모델의 도입으로 인간 상담사에게 전달할 데이터의 크기를 감소시키는 시도를 하였다. 또한, 요약 AI를 도입하는데 있어 새로운 모델을 학습시키지 않고 기존 모델을 주어진 명령어를 따르게

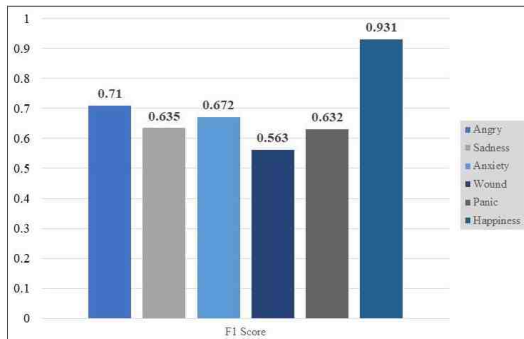


Fig. 11. Evaluation results about test dataset.

튜닝하고 다시 이를 프롬프트 튜닝을 통해 변형을 주는 방법으로 상담과 요약의 서로 다른 목적을 가진 모델로 구성할 수 있었다.

그러나 데이터 확보에 있어 실제 상담을 바탕으로 한 데이터를 확보하지 못해, 상담이라는 도메인을 직접적으로 모델에 적응시키지 못했다는 한계점이 있다. 이러한 한계점을 극복하기 위해 향후 연구에서는 실제 상담데이터를 확보하여 모델을 학습시키는 시도를 할 예정이다.

REFERENCE

- [1] J.Y. Han, Y.J. Kim, H.Y. Joh, J.Y. Lee, and U.R. Oh, "MoodChartBot: Design and Implementation of a Mood Chart Application Using AI Chatbot," *Journal of Korea Multimedia Society*, Vol. 26, No. 3, pp. 503-508, 2023.
- [2] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare," *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7184-7190, 2022.
- [3] Z. Xu, Y. Xu, F. Cheung, M. Cheng, D. Lung, Y.W. Law, et al., "Detecting Suicide Risk Using Knowledge-Aware Natural Language Processing and Counseling Service Data," *Social Science & Medicine*, Vol. 283, pp. 114176, 2021.
- [4] K.K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial," *JMIR Mental Health*, Vol. 4, Issue. 2, pp. 1-11, 2017.
- [5] S.H. An and O.R. Jeong, "A Study on the Psychological Counseling AI Chatbot System based on Sentiment Analysis," *Journal of Information Technology Services*, Vol. 20, Issue 3, pp. 75-86, 2021.
- [6] S.H. Kim, H.E. Park, and D.H. Lee, "Trends and Implications of Digital Therapeutics (DTx) in the Mental Health Field," *The Korean Journal of Counseling and Psychotherapy*, Vol. 34, No. 4, pp. 1401-1430, 2022.
- [7] S.H. Lee and J.Y. Lee, "Development of Artificial Intelligence-based Counseling Chatbot for Elementary School Students and Verification of Effectiveness," *Journal of The Korean Association of Information Education*, Vol. 27, No. 4, pp. 425-432, 2023.
- [8] A.N. Vaidyam, H. Wisniewskit, J.D. Halamka, M.S. Kashavan, and J.B. Torous, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape," *The Canadian Journal of Psychiatry*, Vol. 64, Issue 7, pp. 456-464, 2019.
- [9] H.S. Kim, M.H. Choi, and K.M. Kim, "A Study on the Design and Effectiveness of a AI-based Hyper-Personalized Mental Health Counseling System," *The Journal of Humanities and Social Science 21*, Vol. 14, Issue 3, pp. 3777-3790, 2023.
- [10] D.H. Kim, C.J. Park, S.H. Kim, W.S. Lee, W.H. Song, Y.S. Kim, et al., "SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling," *arXiv Preprint*, arXiv:2312.15166, 2023.
- [11] Solar-Ko-10.7b(2024), <https://huggingface.co/beomi/SOLAR-KO-10.7B> (accessed May 24, 2024).
- [12] K. Clark, M.T. Luong, Q.V. Le, and C.D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," *The Eighth International Conference on Learning Representations*, pp. 1-18, 2020.
- [13] monologg/koelectra-base-v3-discriminator (2020), <https://huggingface.co/monologg/koelectra-base-v3-discriminator> (accessed May 24, 2024).
- [14] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: Efficient Finetuning of Quantized LLMs," *arXiv Preprint*, arXiv:2305.

14314, 2023.

- [15] J. Wei, M. Bosma, V. Zhao, K. Guu, A.W. Yu, B. Lester, et al., "Finetuned Language Models are Zero-Shot Learners," *The Tenth International Conference on Learning Representations*, pp. 1-46, 2022.
- [16] KoAlpaca_v1.1.jsonl(2023), https://github.com/Beomi/KoAlpaca/blob/main/KoAlpaca_v1.1.jsonl (accessed May 24, 2024).
- [17] The Open AI Dataset Project (AI-Hub, S. Korea) Emotional conversation corpus(2020), <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=86> (accessed May 24, 2024).
- [18] The Open AI Dataset Project (AI-Hub, S. Korea) Korean conversation summary(2020), <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSetSn=117> (accessed May 24, 2024).
- [19] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511-2522, 2023.



김 석 준

2018년 3월~2022년 8월 가천대학교 IT융합대학 컴퓨터공학과 학사 수료
2023년 9월~현재 가천대학교 IT융합공학과 석사과정
관심분야: AIoT, AI



이 병 문

1988년 2월 동국대학교 전자계산학과 학사
1990년 2월 서강대학교 전자계산학과 석사
2008년 2월 인천대학교 컴퓨터공학과 박사

1990년~1997년 (주)LG전자 중앙연구소 네트워크 연구실 선임연구원
1998년 3월~현재 가천대학교 IT융합대학 컴퓨터공학과 교수
관심분야: 사물인터넷(IoT), AIoT, 블록체인, 시스템 보안