JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 12: 1759~1768, Dec. 2022

사용자 입력 문장에서 우울 관련 감정 탐지

오재동[1] · 오하영[2*]

Detects depression-related emotions in user input sentences

Jaedong Oh[1] · Hayoung Oh[2*]

[1]Graduate Student, Artificial Inteligence Convergence, Sungkyunkwan University, Seoul, 03063 Korea
[2*]Associate Professor, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063 Korea

요 약

본 논문은 AI Hub에서 제공하는 웰니스 대화 스크립트, 주제별 일상 대화 데이터세트와 Github에 공개된 챗봇 데이터세트를 활용하여 사용자의 발화에서 우울 관련 감정을 탐지하는 모델을 제안한다. 우울 관련 감정에는 우울감, 무기력을 비롯한 18가지 감정이 존재하며, 언어 모델에서 높은 성능을 보이는 KoBERT와 KoELECTRA 모델을 사용하여 감정 분류 작업을 수행한다. 모델별 성능 비교를 위해 우리는 데이터세트를 다양하게 구축하고, 좋은 성능을 보이는 모델에 대해 배치 크기와 학습률을 조정하면서 분류 결과를 비교한다. 더 나아가, 사람은 동시에 여러 감정을 느끼는 것을 반영하기 위해, 모델의 출력값이 특정 임계치보다 높은 레이블들을 모두 정답으로 선정함으로써, 다중 분류 작업을 수행한다. 이러한 과정을 통해 도출한 성능이 가장 좋은 모델을 Depression model이라 부르며, 이후 사용자 발화에 대해 우울 관련 감정을 분류할 때 해당 모델을 사용한다.

ABSTRACT

This paper proposes a model to detect depression-related emotions in a user's speech using wellness dialogue scripts provided by AI Hub, topic-specific daily conversation datasets, and chatbot datasets published on Github. There are 18 emotions, including depression and lethargy, in depression-related emotions, and emotion classification tasks are performed using KoBERT and KOELECTRA models that show high performance in language models. For model-specific performance comparisons, we build diverse datasets and compare classification results while adjusting batch sizes and learning rates for models that perform well. Furthermore, a person performs a multi-classification task by selecting all labels whose output values are higher than a specific threshold as the correct answer, in order to reflect feeling multiple emotions at the same time. The model with the best performance derived through this process is called the Depression model, and the model is then used to classify depression-related emotions for user utterances.

키워드 : 우울증, 인공지능, 자연어처리, 언어 모델
Keywords : Depression, AI, NLP, Language model

Received 19 September 2022, **Revised** 25 September 2022, **Accepted** 4 December 2022

*** Corresponding Author** Hayoung Oh (E-mail: hyoh79@gmail.com, Tel:+82-2-583-8585)
Associate Professor, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.12.1759 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

1.1. 개요

오늘날 우리는 과거에 비해 물질적으로 풍요로운 삶을 살고 있지만 거주 및 근무 환경, 그리고 인간관계가 다양해지며 정신적 피로는 증가하였다. 과거보다 많은 정신적 스트레스를 받으며 나빠진 정신 건강은 우울증으로 이어질 수 있다[1]. 우울증은 정신 질환으로 간주 되지만 신체 건강에 영향을 미칠 수 있다. 우울증의 증상을 조기에 발견하는 한 가지 방법은 사람들이 어떤 이야기를 하고 있는지, 즉 그들의 대화를 분석하는 것이다[2].

우울과 관련된 증상을 조기에 발견하고, 이에 맞는 적절한 대화를 제공함으로써 이러한 감정을 완화해주는 과정은 중요하다고 할 수 있다. 이에 본 논문에서는 사용자 발화에서 우울한 감정을 조기에 발견하기 위해 감성 분석 기법을 적용한다. 더 나아가, 뇌는 컴퓨터처럼 논리에 따라 순차적으로 작동하지 않고, 여러 부위가 동시다발적으로 동작하기에 상충하는 여러 감정이 동시에 일어날 수 있다는 점에서 착안해[3], 우울 관련 감정 또한 다중으로 분류하고자 한다.

우리는 AI Hub에서 공개 데이터로 제공하는 웰니스 상담 데이터세트[4], 주제별 일상 대화 데이터[5], songys github에서 공개한 챗봇 데이터세트[6]를 사용한다. 해당 데이터세트에 대해 KoBERT, KoELECTRA등 여러 모델을 비교 분석하여 사용자 발화에서 우울한 감정을 분류해내는 최적의 모델을 제안한다.

본 논문의 구성은 다음과 같다. 제2장에서는 BERT, ELECTRA 모델의 개념을 간략히 설명한다. 제3장에서는 본 논문에서 사용한 데이터세트에 대한 개요, 데이터 전처리 과정, 데이터 전처리 후 분석 과정에 대해 다룬다. 제4장에서는 여러 모델을 이용해 우울한 감정을 탐지하는 과정에 대해 다룬다. 제5장에서는 본 논문에서 사용한 모델들에 대한 성능평가를 진행한다. 제6장에서는 결론, 개선할 점 및 이후 연구 방향을 소개한다.

1.2. 관련 연구

진효진 et al[7]은 2019년 1월 1일부터 12월 31일까지 심심이-사용자 대화 중 한글로 '우울'이라는 단어가 포함된 사용자 발화 문장 3,000개를 무작위로 추출하여 우울감에 대한 채팅 메시지 분류, 위험 탐지 목적으로 의료분야의 전문가와 공동으로 정량 및 정성 분석을 진행한다.

Yuki et al[2]은 사람들이 대화에서 우울증을 표현하기 위해 사용하는 모든 패턴을 찾기 위해 다양한 종류의

Table. 1 Comparison table of the previous work

title	published	characteristic	dataset	limitations
A study on the categories and characteristics of depressive moods in chatbot data [7]	2021	- Analysis of depression related discourse through chatbot-human interaction data - Categorize the types of depression into 5 categories	- Some of the Simsimi-User conversations from January 1 to December 31, 2019	- Use only one year's worth of analyzed data
Detecting Depression from Human Conversations [2]	2020	- analyzed the text using ten different machine learning algorithms - aim to find any patterns in spoken language people used to express depression	- interviews of various kinds of people male / female, different ethnicity, depressed /non-depressed, anxious/ calm, patients with PTSD (post-traumatic stress disorder) / non-PTSD.	
A Study on the Psychological Counseling AI Chatbot System based on Sentiment Analysis [9]	2021	- Use Dialogflow, a chatbot platform provided by Google, for anxiety testing - Simultaneous emotional analysis and sentence generation using two models - Provides a variety of solutions, including anxiety testing and collective intelligence communities	- Emotional dialgoue corpus - Wellness Conversation Script Dataset	- The Korean embedding-based model has yet to show better performance than the English embedding-based model
A BERGPT-chatbot for mitigating negative emotions [10]	2021	- Built by ensemble KR-BERT model and KoGPT2-chatbot - Text input based conversations	- Naver sentiment movie corpus v1.0 - Emotional dialogue corpus v1.0 - Emotional analysis corpus 2020 v1.0 - Wellness Conversation script dataset - Everyday conversation corpus 2020 v1.1	- Using a small size ELECTRA model
A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support [11]	2020	- 2 tasks to identify empathy in text and extract rationale - develop a multi-task RoBERTa-based bi-encoder model for tasks - show results of Ablation Study	- TalkLife dataset (contains 6.4M threads and 18M interactions) - Mental Health Subreddits dataset (1.6M threads and 8M interactions)	- model sometimes failed to identity short expressions of emotions in responses that otherwise contained a lot of instructions

사람들과 인터뷰한 데이터세트를 사용한다. 해당 데이터세트에 LSTM, SVM, 앙상블 방법, 랜덤 포레스트 및 의사 결정 트리를 포함한 10개의 다른 기계 학습 알고리즘을 사용하고 텍스트를 분석한다.

Sohn et al[8]은 일상생활에서 빈도 높게 사용하는 감정표현단어를 선정하기 위해 총 504개의 감정 표현단어들로 구성된 목록을 완성한다. 이렇게 완성한 감정표현단어들이 '기쁨', '공포', '분노' 등 10개 범주 감정 중 어느 감정과 관련 있는지 선택하도록 하여 감정인식 연구에 활용할 수 있도록 기반을 마련한다.

An et al[9]은 한국어를 기반으로 한 사용자 발화에 대한 응답을 생성하기 위해 KoGPT-2 모델을 사용한다. 더 나아가, 사용자의 발화문에 대한 감성을 분석하고 감성 점수에 따른 해결책을 제시하기 위해 KoELECTRA-small 모델을 사용한다. Song et al[10]은 해외의 "Replika" 같은 텍스트 입력 기반의 부정적 감정 완화가 가능한 챗봇을 구축하기 위해 KR-BERT와 KoGPT2 모델을 혼합한다.

Sharma et al[11]은 EPITOME 프레임워크를 구축하고 공감을 감정 반응, 해석 및 탐색 메커니즘으로 구분한다. 더 나아가, 각각의 메커니즘에 대해 사용자가 게시한 글(seeker post)의 답글(response post)에 공감이 얼마나 포함되어 있는지를 계산하여 적절한 응답을 제시하도록 피드백을 제공한다.

Ⅱ. 관련 기법

2.1. BERT

BERT는 Bidirectional Encoder Representations Transformers로, 모든 계층의 전후 문맥을 공동으로 컨디셔닝함으로써 레이블링 되어 있지 않은 텍스트에 대해 심층 양방향 표현을 사전학습 하도록 설계되었다[12]. BERT 모델에 대한 사전학습과 미세조정학습 절차는 다음 그림 1과 같다. BERT 모델에서 [CLS] 토큰은 모든 입력 예제들의 앞에 추가되는 특별한 토큰이며, [SEP] 토큰은 질의-응답과 같은 텍스트를 분리할 때 사용된다.

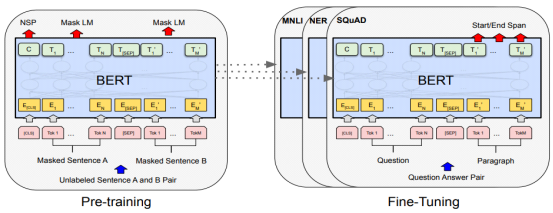

Fig. 1 pre-training and fine-tuning procedures for BERT

BERT 모델은 하나의 토큰 시퀀스에서 단일 문장과 한 쌍의 문장 모두를 명확하게 나타낼 수 있다. BERT 논문에서 사용된 시퀀스는 BERT에 대한 입력 토큰을 의미하며, 이는 한 문장 또는 두 문장을 동시에 일컫는다. BERT 모델은 다음 그림 2와 같은 토큰 임베딩, 세그먼트 임베딩, 위치 임베딩을 입력으로 받으며, 입력 표현은 이들의 총합으로 구성된다.

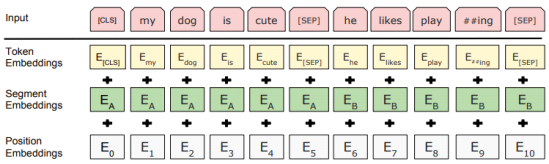

Fig. 2 BERT input representation

토큰 임베딩은 입력받은 모든 문장의 시작으로 [CLS] 토큰을, 문장의 구분을 위해 [SEP] 토큰을 부여한다. 세그먼트 임베딩은 하나의 시퀀스로 표현된 문장 쌍을 구분하기 위해 [SEP] 토큰을 활용한다. 더 나아가, 각 토큰이 문장 A에 속하는지 문장 B에 속하는지 표시하기 위해 0과 1로 구성된 임베딩을 추가한다. 위치 임베딩은 입력 토큰의 위치 정보를 알려주기 위해 사용되며[13], 이는 트랜스포머의 인코더가 사용하는 셀프 어텐션 모델이 입력의 위치를 고려하지 못하는 점을 보완하기 위함이다.

2.2. ELECTRA

BERT 모델은 주로 입력 시퀀스의 토큰 중 약 15% 정도를 마스킹하고, 이를 복원하는 MLM (Masked Language Modeling) 태스크를 통해 학습한다[13]. 이 경우 모델 학습 시 전체 토큰 중 마스킹 된 15%에 대해서만 손실이 발생하며, 비용이 많이 든다. 그뿐만 아니라, 학습에 사용한 [MASK] 토큰은 실제(inference)로는 존재하지 않는다는 한계점이 있다.

ELECTRA 논문[14] 연구진들은, 이러한 MLM 태스크의 한계를 극복하고자 새로운 사전학습 태스크인 RTD(Replaced Token Detection)를 제안한다. RTD는 생성기(generator)를 이용해 실제 입력의 일부 토큰을

그럴싸한 가짜 토큰으로 변경하고, 각 토큰이 실제 입력에 있는 진짜 토큰인지 생성기가 생성한 가짜 토큰인지를 분류기(discriminator)가 맞히는 이진 분류 문제이다. ELECTRA 모델은 대용량 코퍼스 X에 대해 생성기 손실 값과 분류기 손실 값의 합을 최소화하도록 학습한다.

ELECTRA 모델은 RTD 태스크를 통해 입력의 모든 토큰에 대해 학습하기 때문에 상당히 효율적이며 효과적이라 할 수 있다. ELECTRA 모델 구조는 다음 그림 3과 같다. GAN 모델[15]과 비슷한 구조를 보이지만, ELECTRA 논문의 연구진들은 생성기를 적대적 학습 방식이 아닌 최대가능도 추정법(Maximum Likelihood Estimation) 방식으로 학습한다.

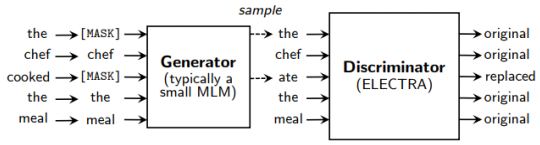

Fig. 3 An overview of replaced token detection

Ⅲ. 데이터세트

본 논문에서는 웰니스 대화 스크립트 데이터세트, 챗봇 데이터세트, 주제별 일상 데이터세트를 바탕으로 인텐트(intent), 사용자 발화(utterance), 레이블을 가진 데이터세트를 구축한다. 인텐트에는 '우울감, 슬픔, 외로움, 분노, 무기력, 감정조절 이상, 상실감, 식욕 저하, 식욕 증가, 불면, 초조함, 피로, 죄책감, 집중력저하, 자신감 저하, 자존감 저하, 절망감, 자살 충동, 불안, 일상' 총 20개가 존재한다. 이 중 일상 인텐트는 챗봇 및 주제별 일상 대화 데이터에서, 나머지 인텐트는 웰니스 데이터를 참고하여 구축한다.

3.1. 데이터 설명

웰니스 대화 스크립트 데이터세트는 강남 세브란스에서 전달받은 상담 데이터 16,000건 중 초진이며, 환자가 직접 내원한 4,200건을 추출하여 감정분석을 진행한 데이터이다. 해당 데이터세트는 발화 의도에 따라 총 359개의 레이블로 분류되며, 대분류는 '감정, 내원 이유, 모호함, 배경, 부가 설명, 상태, 원인, 일반대화, 자가 치료, 증상, 치료 이력, 현재 상태' 총 12개가 존재한다.

챗봇 데이터세트는 특정 커뮤니티에서 자주 나오는 이야기들을 참고하여 이별과 관련된 질문들에 위로의 목적으로 답변을 작성한 데이터세트이다. 사랑과 관련된 질의응답뿐만 아니라, 일상적인 발화에 대한 응답도 포함되어 있다. 본 연구에서는 이 중 일상적인 발화문만 사용한다.

주제별 일상 대화 데이터세트는 식음료, 주거와 생활, 교통, 교육, 가족 등 20개 주제에 대한 자유로운 일상대화 텍스트를 수집한 후, 각 대화의 주제와 참여 화자 정보, 화행이 나타나는 문장에 대한 라벨링을 진행한 데이터세트이다.

3.2. 데이터 분석

연구에 사용한 3가지 데이터세트 각각의 토큰 분포를 살펴보면 다음 그림 4와 같다. 왼쪽부터 순서대로 웰니스 대화 스크립트, 챗봇, 주제별 일상 대화 데이터세트이며, x축은 토큰 길이, y축은 데이터 개수를 의미한다. 웰니스 대화 스크립트와 주제별 일상 대화 데이터세트의 경우 토큰 분포가 서로 유사하며, 챗봇 데이터세트의 경우 다른 두 데이터세트보다 상대적으로 토큰 길이가 짧은 것을 볼 수 있다.

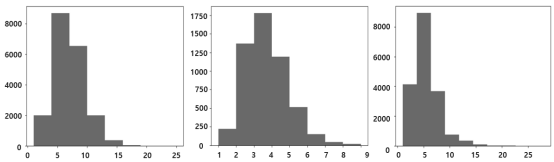

Fig. 4 Token distribution by data set

각 데이터세트의 사용자 발화에서 주로 등장하는 단어들을 알아보기 위해, 워드 클라우드 시각화를 진행해 본 결과는 다음 그림 5와 같다. 왼쪽부터 순서대로 웰니스 대화 스크립트, 챗봇, 주제별 일상 대화 데이터세트이며, 챗봇 데이터세트는 웰니스 대화 스크립트와 주제별 일상 대화 데이터세트의 중간적인 성향을 띄는 것을 볼 수 있다.

Fig. 5 Word Cloud by data set

3.3. 데이터 전처리

본 논문에서는 모델의 예측 성능에 데이터세트의 분포가 미치는 영향을 분석하기 위해, 웰니스 대화 스크립트, 챗봇, 주제별 일상 대화 데이터세트를 4가지 형태로 구축한다. 이와 같은 형태로 구축하는 목적은 웰니스 대화 스크립트에 존재하는 19개의 우울 관련 인텐트와 챗봇, 주제별 일상 대화 데이터세트로부터 얻은 일상 데이터의 적정한 비율을 찾기 위함이다.

데이터세트 1은 웰니스 대화 스크립트 데이터에서 얻은 우울 관련 데이터 19,673개, 챗봇 데이터와 주제별 일상 대화 데이터에서 얻은 일상 관련 데이터 23,000개로 구축한다. 데이터세트 2는 데이터세트 1에서, 챗봇 데이터를 제외하였으며 우울 관련 데이터 19,673개, 일상 관련 데이터 18,000개로 구축한다. 이를 통해 우울 관련된 19개 인텐트 총개수와 일상 관련된 데이터세트의 비율을 1:1로 구축하였으며, 일상 인텐트의 개수가 다른 인텐트들의 평균 개수보다 10배 많다.

데이터세트 3은 웰니스 대화 스크립트 데이터세어 얻은 우울 관련 데이터 19,673개, 주제별 일상 대화 데이터에서 얻은 일상 관련 데이터 1,000개로 구축한다. 데이터세트 4는 데이터세트 3에서, 챗봇 데이터를 추가하였으며 우울 관련 데이터 19,673개, 일상 관련 데이터 6,290개로 구축한다. 이를 통해 우울 관련된 19개 인텐트 평균 개수와 일상 관련된 데이터세트의 비율을 1:1이 되도록 하였으며, 이후 성능평가 부분에서 각 데이터세트 별 f1-score를 분석하고자 한다.

3.4. 가상 대화

학습한 모델의 인텐트 분류 결과를 보기 위해, 본 논문에서는 다음 그림 6과 같이 사람과 챗봇 간에 가상 대화 상황을 설정한다. 가상 대화는 대화 시작, 사용자의 우울 감정 토로, 챗봇과 대화를 통한 우울 감정 완화, 대화 종료 순서로 구성된다.

	turn_cnt	speaker_idx	text
0	1	User	안녕
1	2	Chatbot	안녕하세요 ~ 반가워요
2	3	User	이름이 뭐야 ?
3	4	Chatbot	저는 심리상담 챗봇, ~ 예요
4	5	User	아하 ㅎㅎ
5	6	Chatbot	좋은 아침이예요
6	7	User	맞아, 근데 나는 기분이 우울해
7	8	Chatbot	무슨 일 있어요 ?
8	9	User	그냥 요즘 너무 무기력한거 같아
9	10	Chatbot	무기력할 때는 몸을 움직여야 해요
10	11	User	그러기에는 너무 피로하고 무기력한걸
11	12	Chatbot	많이 지치신게 느껴지는거 같아요. 일이 많이 힘든가요 ?
12	13	User	번아웃이 온 거 같기도 해
13	14	Chatbot	그럴때는 잠시 마음의 여유를 가지고 쉬었다 가는 것도 좋은거 같아요
14	15	User	그래야겠어, 종종 산책하러 다녀야지
15	16	Chatbot	산책 좋조
16	17	User	들어줘서 고마워
17	18	Chatbot	네 ! 다음에 또 봐요

Fig. 6 Virtual Conversation between user and chatbot

Ⅳ. 모델링

본 논문에서 제안하는 전체적인 모델 아키텍처는 다음 그림 7과 같다. 우선 웰니스 대화 스크립트, 챗봇, 주제별 일상 대화 데이터세트에 대해 텍스트 정제, 결측값 제거 과정을 수행한다. 이후 우울 관련 감정 분류를 수행하기 위한 구조로 데이터를 전처리한다. 전처리한 데이터세트에 대해 KoBERT 및 KoELECTRA 모델을 이용한 우울한 감정을 분류하고, 분류 결과를 데이터베이스에 저장한다. 해당 과정은 사용자와 대화가 지속되는 동안 반복 수행되며, 대화가 지속됨에 따른 사용자가 느끼는 감정 변화는 후속 연구에 사용한다.

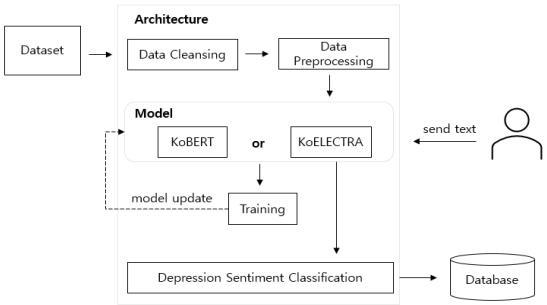

Fig. 7 Proposed scheme

4.1. 토크나이저

모델 학습에 앞서, 텍스트를 토큰 단위로 분할할 때 일부 인텐트들을 한 단어로 인식하도록 KoBERT, KoELECRA 토크나이저의 사용자 정의 사전을 수정한다. 다음 그림 8은 3.3절에서 구축한 학습 데이터세트 중 일부 사용자 발화 문장에 대해 토큰화를 수행한 결과이다. 사용자 정의 사전에 정의되어 있지 않은 인텐트들은 한 단어로 인식하지 못하는 것을 볼 수 있다.

	intent	text	bert_tokenize	electra_tokenize
0	우울감	하지만 우울감은 그대로야.	[_하지만, _우, 울, 감, 은, _그대로, 야, .]	[하지만, 우울, ##감은, 그대로, ##야, .]
1	무기력	어제도 오늘도 무기력해	[_어제, 도, _오늘, 도, _무기, 력, 해]	[어제, ##도, 오늘도, 무기, ##력, ##해]
2	외로움	외로움이 너무 끔찍해	[_외, 로, 움, 이, _너무, _, 끔, 찍, 해]	[외로, ##움이, 너무, 끔찍, ##해]
3	자신감저하	자신감이 바닥이야	[_자신감, 이, _바닥, 이, 야]	[자신, ##감이, 바닥, ##이야]
4	자존감저하	자존감이 바닥을 쳤어...	[_자, 존, 감, 이, _바닥, 을, _, 쳤, 어, ...]	[자존, ##감이, 바닥을, 쳤, ##어, [UNK]]
5	죄책감	하루 종일 죄책감에 시달려	[_하루, _종, 일, _죄, 책, 감, 에, _시, 달, 려]	[하루, 종일, 죄책, ##감에, 시달, ##려]
6	조조함	조조함이 너무 심해	[_조, 조, 함, 이, _너무, _심, 해]	[조, ##조, ##함이, 너무, 심해]

Fig. 8 Tokenization result before add token

한 단어로 인식되지 못하는 인텐트 관련 단어들은 각 토크나이저 사전에 사용자 정의 토큰으로 추가한다. 동일 입력 문장에 대해 토큰화를 진행한 결과는 다음 그림 9와 같다. 무기력, 우울감 등의 인텐트를 하나의 토큰으로 인식하는 것을 볼 수 있다.

	intent	text	bert_tokenize	electra_tokenize
0	우울감	하지만 우울감은 그대로야.	[_하지만, 우울감, _, 은, _그대로, 야, .]	[하지, ##만, 우울감, 은, 그대로, ##야, .]
1	무기력	어제도 오늘도 무기력해	[_어제, 도, _오늘, 도, 무기력, _해]	[어제, ##도, 오늘, ##도, 무기력, ##해]
2	외로움	외로움이 너무 끔찍해	[외로움, _이, _너무, _, 끔, 찍, 해]	[외로움, ##이, 너무, 끔찍, ##해]
3	자신감저하	자신감이 바닥이야	[자신감, _이, _바닥, 이, 야]	[자신감, ##이, 바닥, ##이, ##야]
4	자존감저하	자존감이 바닥을 쳤어...	[자존감, _이, _바닥, 을, _, 쳤, 어, ...]	[자존감, 이, 바닥, ##을, 쳤, ##어, ...]
5	죄책감	하루 종일 죄책감에 시달려	[_하루, _종, 일, 죄책감, _, 에, _시, 달, 려]	[하루, 종일, 죄책감, ##에, 시달려]
6	조조함	조조함이 너무 심해	[조조함, _이, _너무, _심, 해]	[조조함, 이, 너무, 심해]

Fig. 9 Tokenization result after add token

4.2. 인텐트 분류

제3장에서 소개한 데이터세트에는 20개의 인텐트가 존재한다. 우리는 해당 절에서 KoBERT, KoELECTRA 모델을 이용해 인텐트를 분류한다. 각 모델의 학습 파라미터는 기본 설정값을 사용하였으며, 이후 성능 비교 및 속도 측정을 위해 배치 크기, 학습률 및 문장의 최대 길이만 조정한다.

인텐트 분류 알고리즘은 다음과 같은 방식으로 이뤄진다. 우선 사전학습 모델, 토크나이저, 기본 설정 파일을 로드하고, 우울 분류 태스크 수행을 위한 형태로 이들을 조정한다. 이후 3.3절에서 전처리한 각각의 데이터세트를 학습, 검증, 테스트 데이터세트로 분할하고, 모델의 입력으로 전달하기 위한 형태로 변환한다. 모델 학습에 사용할 옵티마이저와 스케줄러를 정의하고, 모델을 학습 및 학습 완료한 모델로 감정을 분류한다. 해당 과정을 알고리즘으로 나타내면 다음과 같다.

Algorithm 1. Intent Classification

LOAD Pretrained Model, Tokenizer, Config file
ADJUST Model, Tokenizer, Config file
READ Data: Data includes Intent, Intent label, Context
SPLIT Data into Train, Validation, Test (size: 0.2)
PROCESS DATA (**RETURN** input_ids, attention_mask, token_type_ids)
DEFINE Data Loader, Optimizer, Scheduler
TRAIN MODEL (KoBERT, KoELECTRA) 30 epochs
TEST MODEL

해당 과정을 앞서 구축한 4개의 데이터세트에 대해 KoBERT, KoELECTRA 모델 각각에 대해 수행하며, 모델별 성능 비교를 위해 모델 학습 횟수는 30으로 통일하고 학습 파라미터도 기본 파라미터를 사용한다. 모델 학습 결과를 분석하여 가장 좋은 성능을 보이는 모델 및 데이터세트를 결정하고, 이후 해당 모델에 대해 배치 크기와 학습률을 조정하여 우울 분류를 위한 최종 모델을 선정한다. 해당 모델을 우리는 'Depression model'이라 부른다. 또한 모델 학습 과정 도중 손실 값에 대해 전체적으로 과적합 현상이 발견되었는데, 최종 모델 선정 시에는 과적합 현상을 방지하기 위해 조기 종료(early stopping callback)를 도입한다.

Depression model을 사용하여 3.4 절의 가상 대화 중 사용자 발화에 대한 우울 감정 분류를 수행한 결과는 다음 그림 10과 같다.

	turn_cnt	speaker_idx	text	predict
0	1	User	안녕	일상
2	3	User	이름이 뭐야?	일상
4	5	User	아하 ㅎㅎ	일상
6	7	User	맞아, 근데 나는 기분이 우울해	우울감
8	9	User	그냥 요즘 너무 무기력한거 같아	무기력
10	11	User	그러기에는 너무 피로하고 무기력한걸	피로
12	13	User	번아웃이 온 거 같기도 해	일상
14	15	User	그래야겠어, 종종 산책하러 다녀야지	일상
16	17	User	들어줘서 고마워	일상

Fig. 10 Classification result of Depression model

4.3. 인텐트 다중 분류

1장에서 언급한 바와 같이 사람은 보통 단일 감정만을 느끼기 보다는 복합적으로 감정을 느끼며, 이는 사용자 발화 문장에서도 표현될 수 있다. 복합적인 감정 분류를 위해, 우리는 Depression model의 예측 방식을 수정한다. 입력받은 사용자 발화에 대해, Depression model은 20가지 인텐트에 대한 로짓(logits)을 반환한다. 단일 감정 분류 모델은 해당 로짓 중 가장 값이 큰 하나의 인텐트를 정답 레이블로 선정한다. 반면, 다중 감정 분류 모델은 해당 로짓 값 중 특정 임계치를 넘는 인

텐트들은 모두 정답 레이블로 선정한다.

본 논문에서는 로짓 값의 임계치를 1, 3, 5로 변환해 가며 출력 결과를 비교하고, 적절한 예측 결과를 보이는 임계치를 설정한다. 다음 그림 11은 임계치를 1로 설정했을 때 모델의 출력 값으로, 우울한 감정이 포함된 사용자 발화에 대해 비교적 많은 레이블을 분류한 것을 볼 수 있다.

	turn_cnt	speaker_idx	text	predict
0	1	User	안녕	일상
2	3	User	이름이 뭐야?	일상
4	5	User	아하 ㅎㅎ	일상
6	7	User	맞아, 근데 나는 기분이 우울해	우울감, 일상
8	9	User	그냥 요즘 너무 무기력한거 같아	외로움, 무기력, 자신감저하, 일상
10	11	User	그러기에는 너무 피로하고 무기력한걸	우울감, 외로움, 무기력, 피로, 일상
12	13	User	번아웃이 온 거 같기도 해	일상
14	15	User	그래야겠어, 종종 산책하러 다녀야지	일상
16	17	User	들어줘서 고마워	일상

Fig. 11 Multi classification result, threshold 1

그림 12는 임계치를 3으로 설정했을 때 모델의 출력 값으로, 임계치를 1로 설정했을 때보다 출력 레이블의 수가 줄어든 것을 볼 수 있다.

	turn_cnt	speaker_idx	text	predict
0	1	User	안녕	일상
2	3	User	이름이 뭐야?	일상
4	5	User	아하 ㅎㅎ	일상
6	7	User	맞아, 근데 나는 기분이 우울해	우울감, 일상
8	9	User	그냥 요즘 너무 무기력한거 같아	무기력
10	11	User	그러기에는 너무 피로하고 무기력한걸	피로
12	13	User	번아웃이 온 거 같기도 해	일상
14	15	User	그래야겠어, 종종 산책하러 다녀야지	일상
16	17	User	들어줘서 고마워	일상

Fig. 12 Multi classification result, threshold 3

그림 13은 임계치를 5로 설정했을 때 모델의 출력 값으로, 그림 10과 같이 단일 분류 모델의 출력 결과와 같은 것을 볼 수 있다. 이러한 출력 결과를 바탕으로, 우리는 우울 감정 다중 분류 모델의 임계치를 3으로 설정한다.

	turn_cnt	speaker_idx	text	predict
0	1	User	안녕	일상
2	3	User	이름이 뭐야?	일상
4	5	User	아하 ㅎㅎ	일상
6	7	User	맞아, 근데 나는 기분이 우울해	우울감
8	9	User	그냥 요즘 너무 무기력한거 같아	무기력
10	11	User	그러기에는 너무 피로하고 무기력한걸	피로
12	13	User	번아웃이 온 거 같기도 해	일상
14	15	User	그래야겠어, 종종 산책하러 다녀야지	일상
16	17	User	들어줘서 고마워	일상

Fig. 13 Multi classification result, threshold 5

Ⅴ. 성능평가

이번 장에서는 앞에서 살펴본 모델들에 대한 성능평가를 진행한다. KoBERT 모델과 KoELECTRA 모델의 학습 속도를 비교하고, 데이터세트 별 모델들의 성능을 비교한다. 마지막으로, 좋은 성능을 보인 데이터세트와 모델에 대해 배치 크기와 학습률을 조정해가며 각 경우에 대한 성능을 평가한다.

5.1. KoBERT vs KoELECTRA 모델 속도 비교

본 논문에서는 KoBERT 모델과 KoELECTRA 모델의 학습 속도를 비교해보기 위해, 토크나이저의 토큰 최대 길이를 32와 512로 설정한 후 앞서 구축한 4개의 데이터세트 각각에 대해 모델 학습을 진행한다. 모델 학습 횟수는 모두 30으로 통일하였으며, 속도 측정 결과는 다음 표2와 같다.

서버 상태에 따라 측정 시마다 속도는 미세하게 차이가 났으며, 각 모델이 1번 학습할 때 걸리는 속도는 큰 차이가 없다. 이를 통해 ELECTRA 모델의 학습 속도가 BERT 모델보다 빠른 이유는 BERT 모델보다 적은 학습 횟수로 비슷한 성능을 낼 수 있기 때문이며, 1 epoch 당 학습 속도는 비슷한 것을 알 수 있다.

Table. 2 Training speed comparison

	Dataset1	Dataset2	Dataset3	Dataset4
KoBERT (max len: 32)	1306(s)	1124(s)	664(s)	793(s)
KoELECTRA (max len: 32)	1325(s)	1178(s)	632(s)	762(s)
KoBERT (max len: 512)	7359(s)	6425(s)	3475(s)	4382(s)
KoELECTRA (max len: 512)	7404(s)	1178(s)	632(s)	762(s)

5.2. 인텐트 분류 모델 성능평가

다음 표3은 Dataset 1~4에 대해 KoBERT, KoELECTRA 모델의 분류 성능을 비교한 것이다. 8개의 모델 중, 우리는 f1-score(F1)가 가장 높게 나온 5번 모델을 1차적으로 선택한다.

Table. 3 Classification result of each model

	Model	Dataset	Accuracy	Loss	F1
1	KoBERT	Dataset2	Train: 1.0 Val: 0.96	Train: 0.16 Val: 0.98	0.9
2	KoBERT	Dataset3	Train: 1.0 Val: 0.845	Train: 0.28 Val: 1.36	0.9
3	KoBERT	Dataset1	Train: 1.0 Val: 0.9	Train: 0.5 Val: 0.86	0.85
4	KoBERT	Dataset4	Train: 1.0 Val: 0.861	Train: 0.47 Val: 1.23	0.86
5	KoELECTRA	**Dataset2**	**Train: 1.0 Val: 0.897**	**Train: 0.3 Val: 1.08**	**0.91**
6	KoELECTRA	Dataset3	Train: 1.0 Val: 0.85	Train: 1.52 Val: 0.482	0.84
7	KoELECTRA	Dataset1	Train: 1.0 Val: 0.9	Train: 0.33 Val: 0.94	0.9
8	KoELECTRA	Dataset4	Train: 1.0 Val: 0.86	Train: 0.57 Val: 1.333	0.87

배치 크기와 학습률은 모델의 성능에 영향을 미친다 [16]. 우리는 앞서 선택한 5번 모델에 대해 배치 크기(Batch)와 학습률(LR)을 조정해가며 성능 비교를 진행하였으며, 결과는 다음 표4와 같다. 하이퍼파라미터를 바꿔가며 학습을 진행하였으나, 기본 설정값에 해당하는 모델의 성능이 가장 좋다.

Table. 4 Classification results by batch size and learning rate

	Batch	LR	Accuracy	Loss	F1
1	**default (32)**	**default (5e-5)**	**Train: 1.0 Val: 0.897**	**Train: 0.3 Val: 1.08**	**0.91**
2	x 1/4	x 1/4	Train: 1.0 Val: 0.9	Train: 5.5 Val: 1.2	0.9
3	x 4	x 4	Train: 1.0 Val: 0.89	Train: 0.2 Val: 0.929	0.9
4	x 1/4	x 4	Train: 0.478 Val: 0.478	Train: 6416 Val: 2.16	0.48
5	x 4	x 1/4	Train: 1.0 Val: 0.879	Train: 8.76 Val: 0.623	0.89

모델의 기본 설정값에서 배치 크기를 1/4로 줄이고 학습률을 4배 늘려서 학습할 때, 모델이 수렴하지 못한다. 다음 그림 14는 모델이 수렴하지 못할 때의 훈련 기록과 모델이 정상적으로 학습될 때의 훈련 기록을 보여준다. 수렴하지 못한 모델은 학습 도중 검증 데이터의 손실 값에 변화가 거의 없는 것을 볼 수 있다.

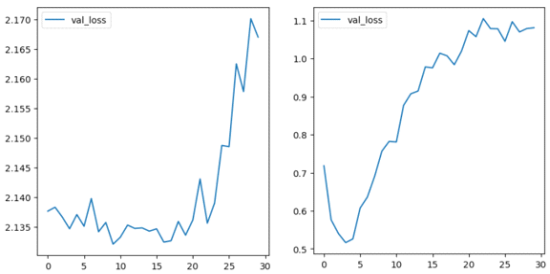

Fig. 14 Training logs of unconverged model(left) and converged model(right)

모델이 정상적으로 학습될 때의 손실 값 변화 기록을 살펴보면, 일정 학습 횟수 동안 감소하다 이후 꾸준히 증가하는 것을 볼 수 있다. 이에 우리는 5번 학습 동안 손실 값이 더 이상 감소하지 않으면 학습을 중단하도록 설정하였으며, 7번 학습 후 종료한 모델을 얻었다. Depression model과 조기 종료를 적용한 Depression model의 성능을 비교하면 다음 표5와 같다. 조기 종료한 모델의 검증 데이터세트 손실 값은 기존 모델보다 나아졌으나 정확도와 F1 점수는 감소한 것을 볼 수 있다.

Table. 5 Classification result of each model

	Model	Accuracy	Loss	F1
1	Depression model (default)	Train: 1.0 Val: 0.897	Train: 0.3 Val: 1.08	0.9
2	Depression model (Early stop)	Train: 1.0 Val: 0.87	Train: 53.84 Val: 0.697	0.89

본 논문에서는 조기 종료한 Depression model의 각 인텐트에 대한 F1 점수를 측정한다. 다음 표6은 F1 점수가 낮은 2개의 인텐트와 높은 2개의 인텐트에 대한 정보를 보여준다. 일상, 상실감 인텐트는 모델이 잘 분류하는 반면, 자신감 저하, 감정조절 이상 인텐트는 모델이 잘 분류하지 못하는 것을 볼 수 있다.

Table. 6 Classification result of each model

	Intent	precision	recall	F1
1	Daily life	0.98	0.97	0.98
2	Sense of loss	0.91	0.89	0.9
3	A decline in confidence	0.57	0.72	0.63
4	Emotionally controlled abnormality	0.51	0.46	0.49

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022 R1F1A1074696).

Ⅵ. 결 론

6.1. 연구 결과

본 논문에서는 웰니스 대화 스크립트, 주제별 일상 대화, 챗봇 데이터세트를 전처리하고 해당 데이터세트에 대해 KoBERT, KoELECTRA 모델을 이용한 인텐트 분류 작업을 진행한다. 우리는 분류 성능을 높이기 위해 데이터세트를 다양하게 구축하였으며, 배치 크기와 학습률을 조정해가며 최적의 모델인 Depression model을 얻는다. 이후, 사용자 발화에 대한 우울 관련 감정 분류 시 해당 모델을 사용하고자 한다.

6.2. 개선 사항

5.2절에서 언급한 바와 같이, 우리가 제안하는 Depression model은 일부 인텐트들에 대한 분류 성능이 좋지 못하다. 후속 연구에서는, 데이터세트의 레이블별 불균형 문제를 고려하여 가중치를 다르게 설정하거나, 모델에 우울 관련한 감정 정보를 입력으로 추가 전달하는 방법을 연구하고자 한다.

6.3. 확장 가능성

본 논문에서 학습한 KoBERT, KoELECTRA 우울 분류 모델을 바탕으로 사용자 발화에 대해 우울한 감정을 조기에 탐지하여 사용자 감정에 따라 맞춤형 답변을 제공하는 챗봇으로 확장할 수 있다. 더 나아가, 사용자와 챗봇 간의 대화 기록을 데이터베이스에 저장하여, 대화가 진행됨에 따라 모델의 예측에 기반한 사용자가 느끼는 우울 관련 감정의 변화를 이후 연구에 사용할 수 있을 것으로 기대된다.

Reference

[1] A red light for modern mental health and stress management are essential [Internet]. Available:http://www.medical-tribune.co.kr/news/articleView.html?idxno=100431.

[2] J. Q. Yuki, Md. M. Q. Sakib, Z. Zamal, S. H. Efel, and M. A. Khan, “Detecting Depression from Human Conversations,” in *Proceedings of the 8th International Conference on Computer and Communications Management (ICCCM'20)*, New York: NY, USA, pp. 14 – 18, 2020.

[3] A story about our brains [Internet]. Available: http://scienceon.hani.co.kr/436471.

[4] Wellness conversation dataset [Internet]. Available: https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-006.

[5] Subject-specific text Daily conversation [Internet]. Available: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=543.

[6] Korean Chatbot dataset [Internet]. Available: https://github.com/songys/Chatbot_data.

[7] H. Chin, G. Baek, C. Cha, J. Choi, H. Im, and M. Cha, “A study on the categories and characteristics of depressive moods in chatbot data,” in *Proceedings of the Korea Information Processing Society Conference*, Yeosu, Korea, pp. 993 – 996, 2021.

[8] S. J. Sohn, M. S. Park, J. E. Park, and J. H. Sohn, "Korean Emotion Vocabulary: Extraction and Categorization of Feeling Words," *Science of Emotion and Sensibility*, vol. 15, no. 1, pp. 105-120, Mar. 2012.

[9] 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, no. 3, pp. 75 – 86, Jun. 2021.

[10] Y. G. Song, K. M. Jung, and H. Lee, “A BERGPT-chatbot for mitigating negative emotions,” *Journal of the Korea Society of Computer and Information*, vol. 26, no. 12, pp. 53-59, Dec. 2021.

[11] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, “A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 5263-5276, 2020.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, Minneapolis: MN, USA, pp. 4171-4186, 2018.

[13] About BERT [Internet]. Available: https://happy-obok.tistory.com/23.

[14] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” *arXiv preprint arXiv:2003.10555*, Mar. 2020.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139-144, Oct. 2020.

[16] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT Express*, vol. 6, no. 4, pp. 312-315, Dec. 2020.

오재동(Jaedong Oh)

성균관대학교 인공지능융합학과 석사
※관심분야 : 추천시스템, NLP, 챗봇

오하영(Hayoung Oh)

서울대학교 컴퓨터공학과 박사
숭실대학교 조교수
아주대학교 부교수
U.C.Berkeley 방문연구원
성균관대학교 소프트웨어융합대학 부교수
※관심분야 : 소셜정보망 분석, 추천시스템, 데이터분석 및 인공지능