우울 감정 진단 및 LLM을 활용한 커뮤니티 답글 생성 시스템

CUAI 6기 NLP 1팀
배현규(응용통계학과), 좌대현(응용통계학과), 허윤빈(응용통계학과)

[요약] 본 연구는 커뮤니티 글에서 우울 감정을 진단해내고 우울 감정으로 판단된 글에 대해 답글을 생성하는 것을 목표로 진행하였다. 연구는 크게 우울 감정 분류 모델과 답글 생성 모델 두개로 나눠진다. 학습 결과 성능이 가장 좋았던 KcELECTRA를 분류모델로, KoGPT-2를 최종 생성모델로 선택하였다. 본 연구는 공감과 위로를 필요로 하는 사람들에게 온라인 상에서 자연스럽게 치유의 말을 건넬 수 있다는 점에서 의미가 있다. 향후 더 다양한 데이터를 수집해 우울증 챗봇 혹은 대화에서 우울증을 판단해내는 등 다양한 연구로 발전시킬 수 있다.

1. 서 론

우울증이 패션이 될 정도로 우울하다는 말을 입에 달고 사는 한국인들에게, 우울증은 더이상 낯설게만 느껴지는 정신적 질환이 아니다. 실제로, 대학생들이 많이 이용하는 커뮤니티인 '에브리타임'의 게시판 중, 우울증 게시판이 존재한다는 사실만으로도 많은 사람들이 우울한 감정과 함께 살아가고 있다는 것을 알 수 있다.

누구보다도 공감과 위로가 필요한 이들이지만, 그럼에도 불구하고 우울증 게시판에서는 조롱하는 답글이나 장난치는 답글과 같은 불필요한 답글들이 수시로 작성되는 것을 확인할 수 있었다.

이에 본 연구는 텍스트에서 나타나는 우울, 불안, 슬픔, 상처 등의 감정을 인식하고 해당 글에 대해 따뜻하고 지지적인 답글을 생성함으로써 온라인 커뮤니티의 긍정적인 기능을 증진하고자 한다. 본 연구는 우울과 관련된 감정을 파악하고, 작성된 글에 적절한 공감과 위로의 글을 생성하는 것을 목표로 하고 있다.

2. 본 론

우울한 감정을 느끼는 사람들에게 적절한 답글을 생성하려면, 크게 두가지 모델이 학습되어야 한다. 하나는 여러 게시글 중 우울 감정을 분류하는 모델, 또다른 하나는 글의 문맥을 파악하여 적절한 답변을 제시하는 모델이다. 본론에서는 AI HUB에 공개된 데이터와 네이버 지식인 등 직접 수집한 데이터를 활용해 두 가지 모델을 학습한다. 아래 〈그림 1〉는 학습된 모델들을 활용해 적절한 답글을 생성해내는 답글 생성 AI의 흐름도이다.

〈**그림 1**〉 우울 감정 게시글 답글 생성 AI 흐름도

Ⅰ. 분류 모델

1) 데이터셋

학습 데이터로는 AI HUB에서 제공되는, 정신건강 상담 대화로 이루어진 웰니스 데이터[1], 감성 대화 말뭉치[2], 한국어 감정 정보가 포함된 단발성 대화 데이터[3]를 이용하여 우울, 비우울로 이진 분류를 하였다.

정신건강 상담 대화인 웰니스 데이터는 우울한 감정을 잘 나타낸다고 판단하여 1로 라벨링 하였고, 감성 말뭉치와 단발성 대화에서 중립, 기쁨, 행복의 감정만을 추출해 0으로 라벨링 하였다.

본 연구의 최종 목적은 실제 커뮤니티 상에서 우울 감정을 판단하는 것이므로, 에브리타임 우울증 게시판과 자유게시판에서 각각 200개씩 글을 수집해 테스트 데이터로 이용했다. 우울증 게시판에서 수집한 글은 1, 자유게시판에서 수집한 글은 0으로 라벨링 해주었다. 이후 분류가 모호한 데이터는 제거하여 총 390개의 데이터를 테스트 데이터로 사용하였다.

에브리타임의 게시글을 통해 test를 실시해본 결과 0.7 정도의 성능이 도출되었다. 성능 향상을 위해 학습에 사용될 데이터 셋을 새로 구성하였다.

새로운 학습 데이터셋은 감성 대화 말뭉치와 단발성 대화 데이터만을 이용하였다. 감정 대분류 중 불안, 분노, 상처, 슬픔에 대해서는 1로 라벨링을 하였고, 기쁨, 행복, 놀람, 중립에 대해 0으로 라벨링하여 새로운 학습 데이터셋을 구성하였다.

2) 전처리

학습 데이터셋에 대해 불용어 처리, 특수문자 및 공백 제거, 중복 문자 제거를 진행하였다. 먼저 구성된 데이터셋에 Okt 형태소 분석기를 사용하여 의미상 필요 없는 조사, 어미 등을 제거하였다. 다음으로 숫자, 영어, 특수문자 등 의미상 큰 영향을 주지 않는 단어들을 제거한 후, 'ㅋㅋ…ㅋㅋ'와 같이 반복되는 자음들은 두 번만 반복되도록 처리하여 주었다. 마지막으로 Bi- LSTM의 경우 추가적인 전처리로, 등장 빈도가 2회 이하인 단어들은 제거하여 주었다. 전처리 후 모든 문장이 포함되도록max-len은 최대문장 길이인 125보다 큰 128로 설정하였다.

3) 학습 모델

모델은 BiLSTM, GPT, 자연어처리 분야에서 좋은 성능을 보이는 BERT의 파생 모델인 ELECTRA 총 세가지 모델을 사용하였다. BiLSTM은 모델은 직접 학습하였고 ELECTRA, GPT의 경우는 Pretrained된 이용해 파인튜닝을 진행하였다.

BiLSTM은 정방향으로만 학습을 진행하던 LSTM을 발전시켜 마지막 노드에서 역방향으로 실행되는 다른 LSTM을 추가한 것이다. 기존 LSTM은 시퀀스를 정방향으로만 처리하여 문장의 전체적인 의미를 이해하는 데에 제약이 있었다. 그러나 BiLSTM은 양방향으로 정보를 수집하고 이를 결합하여 문장의 전체적인 의미를 파악할 수 있게 되었다.

ELECTRA[4]는 BERT의 파생모델이다. BERT에서 MLM task를 사전 학습에 사용하는 것과 달리 ELECTRA는 마스킹할 대상이 될 토큰들을 다른 토큰으로 변경한 뒤 이 토큰이 실제 토큰인지 교체된 토큰인지 판별하는 형태로 학습을 진행한다. BERT는 마스킹된 토큰이 무엇인지를 예측하기에 마스킹된 15%의 토큰만을 학습하지만 ELECTRA는 각 토큰의 원본 여부를 판단해야 하므로 모든 토큰을 대상으로 학습이 이루어진다는 점에서 더 효율적이다. 본 연구에서는 네이버 뉴스에서 댓글과 대댓글을 수집해 토크나이저와 ELECTRA 모델을 학습한 KcELECTRA[5]를 사용하였다.

GPT는 트랜스포머의 디코더와 비슷하게 이루어져 있으며 학습이 단방향으로 진행되기 때문에 문장 생성 분야에서 더 많이 사용된다. 하지만 자연어처리 분야에서 전반적으로 좋은 성능을 보이기에 분류 task를 시도해 보았다.

본 연구에서는 GPT 모델 중 한국어에 최적화 되어있는 KoGPT2[6] 모델을 사용하였다.

4) 성능 평가

성능은 1) 데이터셋에서 설명한 에브리타임에서 크롤링해온 데이터를 사용하여 계산하였고 결과는 아래와 같다.

〈표 1〉 분류모델 성능평가 표

	BiLSTM	KcELECTRA	KoGPT2
Accuracy	0.7923	**0.8949**	0.8462
Recall	0.7665	**0.9137**	0.8680
Precision	0.8118	**0.8824**	0.8341
F1-score	0.7885	**0.8978**	0.8507

위 표와 같이 전체적으로 가장 우수한 성능을 보인 Kc-ELECTRA를 최종 분류 모델로 선택하였다.

Ⅱ. 생성모델

1) 데이터셋

우울 감정으로 분류된 글에 대해서 적절한 답변을 생성하기 위해 질문과 대답으로 이루어진 대화쌍 데이터가 필요했다. 따라서 생성모델 학습을 위해 분류모델에 사용했던 데이터와 같은 감성 말뭉치 대화 데이터셋을 이용했다.

감성 말뭉치 대화 데이터셋은 사람의 손을 거쳐 직접 구축된 코퍼스이기 때문에, 해당 데이터셋을 이용해 학습한 결과가 SNS 글에 댓글을 생성하기에는 말투와 표현이 일부 어색할 수 있다는 우려가 있었다. 자연스러운 표현을 학습시키기 위해 네이버 지식인에서 직접 크롤링을 통해 우울 감정과 관련한 약 1000개의 질문글, 답글 쌍을 수집하였다.

2) 전처리

크롤링한 데이터셋의 답변에는 답변자를 소개하거나 관련 기관을 홍보하는 등 불필요한 정보가 다수 포함되어 있었다. 답변의 형식을 일정하게 하도록 처리하기 위해 다음 그림과 같이 openai의 api를 이용해 답변의 형식을 일부 보정하였다.

```
# GPT-3 API 호출 함수
def generate_text(text):
    response = openai.chat.completions.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": "입력받은 문장은, 전문의들이 우울증이 있는 사람들에게 답변한 문장들이다."},
            {"role": "system", "content": "우울증에 대한 해결책이 포함된 부분만 요약해서 문장으로 제시해라"},
            {"role": "system", "content": "한국어로 답변해라"},
            {"role": "system", "content": "입력받은 문장의 문체를 최대한 유지한 채로 문장을 제시해라."},
            {"role": "system", "content": "단순히 해결책에 대해 나열하지 말고, 연속된 문장들로 답변해라."},
            {"role": "user", "content" : text}
        ],
        temperature=0.5
    )
    return response.choices[0].message.content
```

〈그림 2〉 답변글의 형식보정을 위한 프롬프팅

프롬프팅을 통해 처리된 답변을 확인해보니, 여전히 '채택해주세요'와 같은 문구나 특정 기관들의 홍보성 답변이 남아있었다. 이는 생성모델의 적절한 댓글 생성을 방해하므로 삭제 처리하였다.

지식인 질문 글에 대해서는 'ㅋㅋㅋㅋ', 'ㅎㅎ' 와 같은 초성으로 된 단어를 삭제하고, 자살을 'ㅈㅅ'로 표현하는 등의 줄임말을 올바른 단어로 대체하였다.

모든 처리가 끝난 후, 데이터 어그멘테이션을 진행했다. 어그멘테이션은 문장 내 단어를 유사한 단어로 교체하는 replacement 방법과 문장에 새로운 단어를 추가하는 insertion 방법을 사용하였다.

먼저 replacement 방법은 랜덤으로 문장내 한 단어를 마스킹한 후 마스킹된 단어를 Masked Language Model로 채워넣는 방식을 사용했다. 사용된 모델은 fill-mask 방식으로 학습된 Ko-ELECTRABert 모델을 사용했다.

〈표 2〉 replacement을 이용한 데이터 증강

증강 전	이순신 장군은 매우 뛰어난 장군이다.
증강 후	이순신 장군은 매우 훌륭한 장군이다.

다음으로 insertion 방식은 랜덤으로 단어들 사이에 mask token을 삽입해 그 mask token을 채우는 방식을 사용했다. 모델은 replacement와 같은 Ko-ELECTRA Bert 모델을 사용했다.

〈표 3〉 insertion을 이용한 데이터 증강

증강 전	이순신 장군은 매우 뛰어난 장군이다.
증강 후	이순신 장군은 매우 뛰어난 조선의 장군이다.

3) 학습모델

답변 생성 모델로는 앞서 분류모델에 사용한 Ko-GPT2와 더불어, Transformer의 인코더와 디코더 구조를 모두 지닌 BART를 사용하였다.

BART는 Bidirectional AutoRegressive Transformer의 약자로, 양방향으로 시퀀스의 토큰들을 어텐션 메커니즘으로 인코딩하여 문맥을 잘 파악하는 BERT의 특성과, 주어진 연속된 단어를 맞추는 방식으로 학습되어 generation task에 적합한 GPT의 특성을 모두 지닌 모델이다.

보통 BART모델은 Summarization과 같은 다운스트림 태스크에 자주 사용되는 모델이지만, 인코딩된 게시글에 대해 적절한 이해를 바탕으로 디코더 구조에서 적합한 답글을 출력해낼 수 있을 것이라고 생각하여 질의응답을 위한 모델로 선정하였다.

BART모델중에서도 한국어 이해에 조금 더 적합하게 프리트레인된 모델인 KoBart[7] 모델을 사용해 게시글에 대한 답변을 출력하는 모델을 파인튜닝하였다.

4) 성능지표

text generation 모형의 일반적으로 자주 사용되는 성능지표인 Perplexity(이하 PPL)를 사용하였다.

PPL은 정규화된 문장에 대한 확률에 역수를 취한 값으로 다음과 같이 수식으로 표현할 수 있다.

$$PPL(W) = \sqrt[N]{\frac{1}{\Pi_{t=1}^{n} P(w_t|w_1, ..., w_{t-1})}}$$

이전 단어로 다음 단어를 예측할 때 평균적으로 몇 개의 단어 후보를 고려하는지를 의미하는 지표이다. 때문에, PPL값이 낮을수록 좋은 언어 모델이라 해석한다.

5) 모델 결과 비교

다음은 Ko-GPT2와 KoBART학습 결과를 PPL 지표로 비교한 표이다. 파인튜닝을 진행하기 전 base모델과 학습에 사용된 데이터 별로 성능을 비교해 보았다.

〈표 4〉 감성말뭉치 파인튜닝

Perplexity	Model	
	KoGPT-2	KoBART
base	118.36	78191.35
5-epoch	2497.39	72727.88
10-epoch	10580.58	77291.38

〈표 5〉 네이버 지식인 파인튜닝

Perplexity	Model	
	KoGPT-2	KoBART
base	118.36	78191.35
5-epoch	813.19	71973.62
10-epoch	1561.36	78371.74

앞서 밝혔듯 PPL로는 생성 모델의 대략적인 성능 지표를 알 수 있다. 하지만 PPL이 높더라도 어떤 모델의 문장이 자연스러운 경우도 여럿 있다. 이번 연구에서도 KoGPT2에 비해 KoBART의 PPL이 매우 높게 나온 것을 알 수 있다. 이는 학습에 사용한 KoBART모델이 summarization 태스크를 프리트레인했기 때문에 질의응답 태스크를 파인튜닝 했을 때 PPL수치가 높게 나온 것으로 추측한다.

따라서 본 연구에서는 PPL뿐만 아니라, 생성된 답변의 수준을 직접 확인하고 이를 모두 고려하여 문장 생성에 적합한 최종 모델을 선정한다. 학습된 모델들을 활용해 실제 에브리타임 우울증 게시판 게시글을 크롤링 한 뒤 우울로 분류된 글들에 대해 답글을 생성해 보았다. 우울로 분류된 글들의 일부에 대해 답글을 생성한 결과는 다음과 같다.

〈표 6〉 에브리타임 우울증 게시판 게시글

Post
살면서 이룬게없어 할수있는게 새벽에 술까는거밖에 없네

〈표 7〉 GPT-2 감성말뭉치 파인튜닝

5-epoch comment	10-epoch comment
무슨일 있으세요? 조금 더 자세히 말씀해주세요	지금 기분이 안좋으신가요?

〈표 8〉 GPT-2 네이버 지식인 파인튜닝

5-epoch comment	10-epoch comment
저도 예전에 비슷한 경험이 있어서 이렇게 댓글을 달게 되었습니다. 먼저, 정신건강의학과에 대한 상담 받아보시는 것을 추천드립니다. 그리고 그걸 통해 몇 가지 조언을 받을 수 있습니다. 우선, 잠이 오지 않더라도 꼭 취하도록 하세요. ... 혼자서 모든 걸 해결할 수는 없습니다. 가능한 빨리 전문가의 도움을 받아 건강한 삶을 살아가시길 바랍니다.	그리고 한 시간씩 술을 마실 것도 도움이 될 수 있습니다. 간혹은 우울증도 겪으셨다고 하셨군요. 우울증은 정말로 힘들고 지칠 수 있는 상태입니다. 이런 어려운 상황에서는 전문가의 도움을 받는 것이 중요합니다. ... 의사나 의사와 상담하여 적절한 조치를 취할 필요가 있습니다.

〈표 9〉 KoBART 감성말뭉치 파인튜닝

5-epoch comment	10-epoch comment
어떤게 가장 힘드신가요?	무슨게 제일 힘드세요?

〈표 10〉 KoBART 네이버 지식인 파인튜닝

5-epoch comment	10-epoch comment
우울증의 증상들이 나타나기 시작합니다. 이 증상을 개선하기 위해서는 먼저 약물치료가 필요합니다. 약물을 복용하는 것이 가장 중요합니다. 약물 치료는 약물에 대한 의존도를 낮추는 것이 좋습니다. 약을 복용하면 약물의 부작용이 나타날 수 있습니다. 약물은 약물과 함께 복용할 수 있으며, 약물 복용 시에는 약물이 복용될 수 있으므로 약물 복용시에는 약물 복용을 중단해야 합니다. 약물 치료를 받는 것이 도움이 됩니다.	우울증의 증상을 완화하기 위해 정기적인 검진을 받는 것이 좋습니다. 정기적으로 검진 받는 것도 좋은 방법 중 하나입니다. 정기검진을 통해 증상 완화에 도움을 줄 수 있습니다. 정기 검진과 정기적 검진은 신체적, 정신적, 사회적 건강, 심리적인 안정에 도움이 될 수 있으므로 정기검사를 통해 정확한 진단과 적절한 치료 방법을 찾는 것이 중요합니다.

대체로, 감성말뭉치를 통해 학습된 모델은 상대방의 상태와 기분을 묻는 질문들이 많이 등장한다. 또한 네이버 지식인 대화를 통해 파인튜닝된 모델들은 우울증의 증상 완화에 초점이 맞춰진 답변을 생성한다. 이는 지식인 답변을 프롬프팅을 진행할 때, 증상의 해결책을 위주로 텍스트를 보존했기 때문인 것으로 보인다.

GPT기반 모델과 BART기반 모델을 비교해보면, GPT기반의 답글 모델이 조금 더 풍성한 표현과 자연스러운 답변을 제공하는 것을 확인할 수 있다. BART모델의 경우 '약물을 복용하는 것이 가장 중요합니다. 약물 치료는 약물에 대한 의존도를 낮추는 것이 좋습니다.'와 같이 특정 단어를 반복해서 답변을 생성하는 경향이 있었다.

문장 생성에 조금 더 표현이 자연스럽고, PPL수치가 낮았던 GPT-2 기반 모델을 답변 생성 모델로 선정하였다. 학습하는 epoch가 커질수록 ppl이 증가함을 확인했고, 답변의 질적 차이가 뚜렷히 나타나지 않았기 때문에 5-epoch의 GPT2 모델을 최종 답글생성 모델로 선정하였다.

3. 결 론

본 연구의 목적은 인터넷 상에 올라온 글의 감정을 우울과 비우울로 분류한 후, 우울로 분류된 글에 위로가 되는 댓글을 생성해주는 모델을 학습하는 것이다. 연구 결과 우울 감정을 분류하는 모델로는 KcELECTRA모델을, 답글 생성모델로는 Ko-GPT2 모델을 통해 파인튜닝 된 모델을 최종 답글 생성 모델로 선정하였다.

본 연구의 한계점 및 보완할 점은 다음과 같다. 먼저, 양질의 데이터셋을 통해 답글 생성모델을 학습시키지 못했고, 컴퓨팅 자원의 한계로 모델을 충분히 학습시키지 못하였다. 답글모델 학습에 사용된 감성 말뭉치는 챗봇 학습을 위한 데이터로 대부분 문장의 길이가 짧고, 네이버 지식인 데이터는 일반적인 글의 뉘앙스가 타 커뮤니티 글과는 사뭇 분위기가 달랐다. 그 결과, 본 연구 목적에서 제안했던 '따뜻한 위로의 말을 건네주는 답글'보다는 우울증의 해결방안에 초점을 둔 답글이 생성되거나, 상태를 묻는 정도에 그치는 답글이 많이 생성되었음을 확인할 수 있다. 또한 다양한 모델 및 성능지표들을 추가로 활용했다면 더욱 적절한 모델을 선별할 수 있었을 것이라 기대한다. 다양한 대화상황의 데이터들을 수집하여 학습하고 LLama2와 같은 추가적인 생성형 LLM모델들을 활용해 학습한다면 원래 연구 목적에 더욱 적합한 답글 생성 모델을 얻을 수 있을 것이다.

일부 사람들은 인터넷 상에서 익명이라는 그림자 뒤에 숨어, 타인에게 상처주는 말들과 욕설을 서슴없이 쏟아내곤 한다. 그런 점에서 본 연구는 생성형 LLM의 파인튜닝을 통해 답글을 생성하여, 서로를 공감하고 위로할 수 있는 따뜻한 온라인 커뮤니티 문화 형성에 일조할 수 있다는 점에서 의의가 있다.

참고 문헌

[1] AI-HUB 데이터, 웰니스 대화 스크립트 데이터셋, https://aihub.or.kr/aihubdata/data/view.do?currMenu=120&topMenu=100&aihubDataSe=extrldata&dataSetSn=267

[2] AI-HUB 데이터, 감성 대화 말뭉치, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=86

[3] AI-HUB 데이터, 한국인 감정정보가 포함된 단발성 대화 데이터셋, https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=270

[4] Clark, Kevin, et al. "ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS", ICLR 2020, 2020

[5] beomi/KcELECTRA-base-v2022. Available:beomi/KcELECTRA-base-v2022 · Hugging Face

[6] skt/kogpt2-base-v2. Available: SKT-AI/KoGPT2: Korean GPT-2 pretrained cased (KoGPT2) (github.com)

[7] gogamza/kobart-base-v2. Available: https://huggingface.co/gogamza/kobart-base-v2