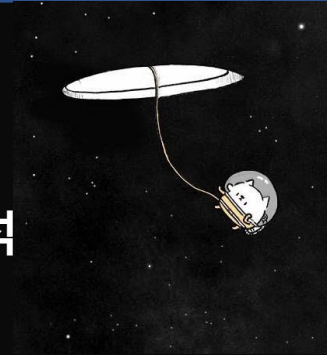


## 챗봇 데이터셋 & 모델 구조 분석



## 기존 논문에서 구축한 단어 사전 정보

Note

### 논문명: 한국어 감정표현단어의 추출과 범주화 (2012, 한국감성과학회)

감정 분류: 11가지 (기쁨, 슬픔, 공포, 분노, 혐오, 놀람, 흥미, 지루함, 통증, 중성, 기타)

번호	단어	감정 범주	빈도(%)	감정 강도 M	감정 강도 STD	번호	단어	감정 범주	빈도(%)	감정 강도 M	감정 강도 STD
1	가련하다	슬픔	88.8	5.62	2.36	1	가뿐하다	기쁨	81.3	5.36	2.89
2	가슴 아프다	슬픔	93.8	7.62	2.14	2	감개무량하다	기쁨	83.8	7.23	2.55
3	가슴앓이	슬픔	90	7.75	1.98	3	감격하다	기쁨	92.5	7.81	2.16
4	가엾다	슬픔	87.5	6.26	2.59	4	감동하다	기쁨	88.8	7.48	2.32
5	각박하다	슬픔	42.5	5.87	2.47	5	감미롭다	기쁨	77.5	6.03	2.63
6	간절하다	슬픔	57.5	4.95	2.72	6	감복하다	기쁨	61.3	6.64	2.65
7	걱정하다	슬픔	66.3	4.55	2.29	7	감사하다	기쁨	97.5	7.24	2.26
8	고달프다	슬픔	77.5	5.66	2.79	8	감회	기쁨	51.2	4.58	2.5
9	고독하다	슬픔	70	6.29	2.68	9	감흥	기쁨	61.3	5.83	2.59
10	곤욕스럽다	슬픔	32.5	4.62	2.71	10	경쾌하다	기쁨	95	7.42	2.53
11	궁허하다	슬픔	57.5	5.74	2.54	11	고맙다	기쁨	95	7.26	2.43
12	괴롭다	슬픔	70	6.31	2.58	12	근사하다	기쁨	60	6.03	2.83
13	구슬프다	슬픔	100	7.72	2.04	13	기분 좋다	기쁨	98.8	7.97	2.41
14	그리워하다	슬픔	86.3	6.29	2.74	14	기쁘다	기쁨	100	8.8	1.96
15	근심스럽다	슬픔	63.7	4.82	2.49	15	까르르	기쁨	96.3	7.17	2.53
16	글썽글썽하다	슬픔	100	6.8	2.5	16	깔깔깔	기쁨	96.3	7.28	2.6
17	기구하다	슬픔	47.5	5.81	2.86	17	껄껄	기쁨	93.8	6.58	2.61
18	기운 없다	슬픔	60	4.37	3.06	18	끼득끼득	기쁨	78.8	5.53	2.87
19	낙담하다	슬픔	93.8	6.21	2.31	19	낄낄	기쁨	90	5.74	2.42
20	낙망하다	슬픔	73.8	6.07	2.67	20	낭만적	기쁨	82.5	6.69	2.61
21	낙심하다	슬픔	90	6.42	2.65	21	달갑다	기쁨	72.5	6.12	2.49
22	난처하다	슬픔	37.5	3.73	2.17	22	대견하다	기쁨	86.3	6.56	2.74
23	남부럽다	슬픔	32.5	3.45	2.09	23	두근거리다	기쁨	75	6.4	2.67
24	낭패스럽다	슬픔	50	5.07	2.46	24	따뜻하다	기쁨	83.8	5.95	2.71
25	눈물겹다	슬픔	97.5	7.04	2.32	25	따스하다	기쁨	78.8	6	2.92
26	망막하다	슬픔	43.8	4.65	2.14	26	만족하다	기쁨	90	7.38	2.57
27	망연	슬픔	81.3	6.08	2.39	27	반갑다	기쁨	95	7.45	2.47
28	망연자실하다	슬픔	83.8	6.74	2.52	28	방그레	기쁨	97.5	6.98	2.76
29	먹먹하다	슬픔	65	6.15	2.87	29	방글방글	기쁨	96.3	6.97	2.49
30	몽글하다	슬픔	76.3	6.58	2.41	30	방긋	기쁨	98.8	6.5	2.74

이 중 우울 관련된 감정은 슬픔에 주로 분포.

슬픔 범주만 참고하여 우울 전용 감성 사전을 구축할 것인가 or 다른 감정들도 함께 구축하여 보다 다양한 감정을 캐치할 수 있도록 할 것인가

#### 감정 강도 및 범주 기록

- 청주대학교 1~3학년 중 사회복지실천과목을 수강하는 학부생 80명이 참가.

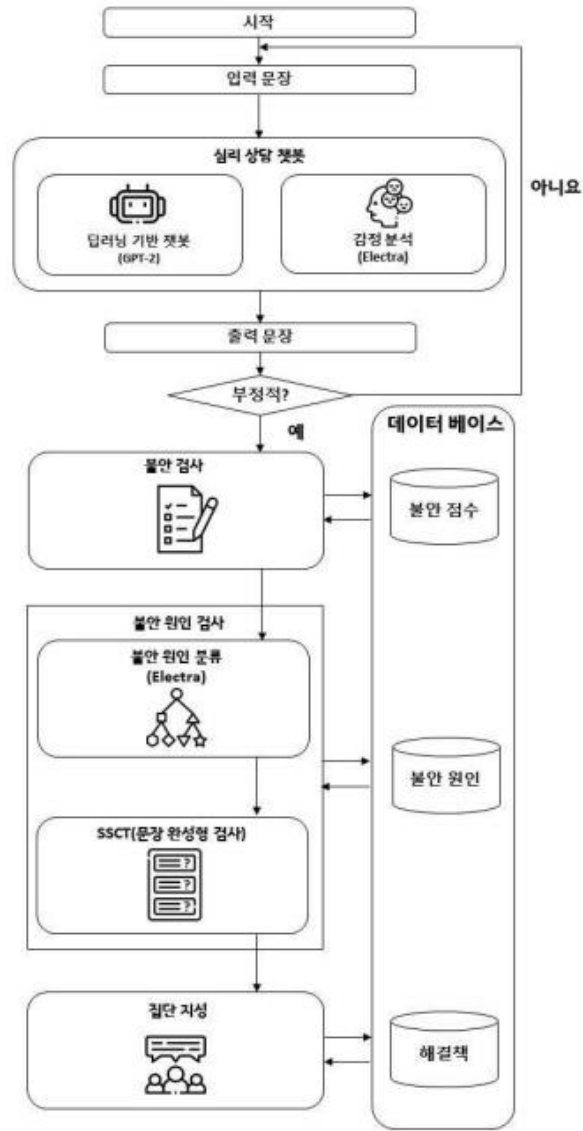
감정표현 단어에 대한 감정 범주화 작업을 하는데 있어 각 감정 표현이 '기쁨', '슬픔', '공포', '분노', '혐오', '놀람', '흥미', '지루함', '통증', '중성' 그리고 '기타'의 11개 감정에 해당하는지에 대해 기록하도록 함

감정의 강도 측정은 단어가 해당 감정을 얼마나 강하게 표현하는지를 10점 척도를 이용하여 평가하였으며, 단어의 의미를 모르고 중성에 기록한 경우 감정 강도 평가 x

감정 강도 M: 감정 강도 측정 값의 평균 값

감정 강도 STD: 감정 강도 측정 값의 표준 편차 값

빈도: 몇 퍼센트의 참가자가 단어를 해당 감정으로 분류하였는지를 측정한 값



[그림 4] AI 챗봇 시스템 구조

논문 제목: 감정분석 기반 심리상담 AI 챗봇 시스템에 대한 연구 (2021, 한국IT서비스학회지)

## GPT-2 모델

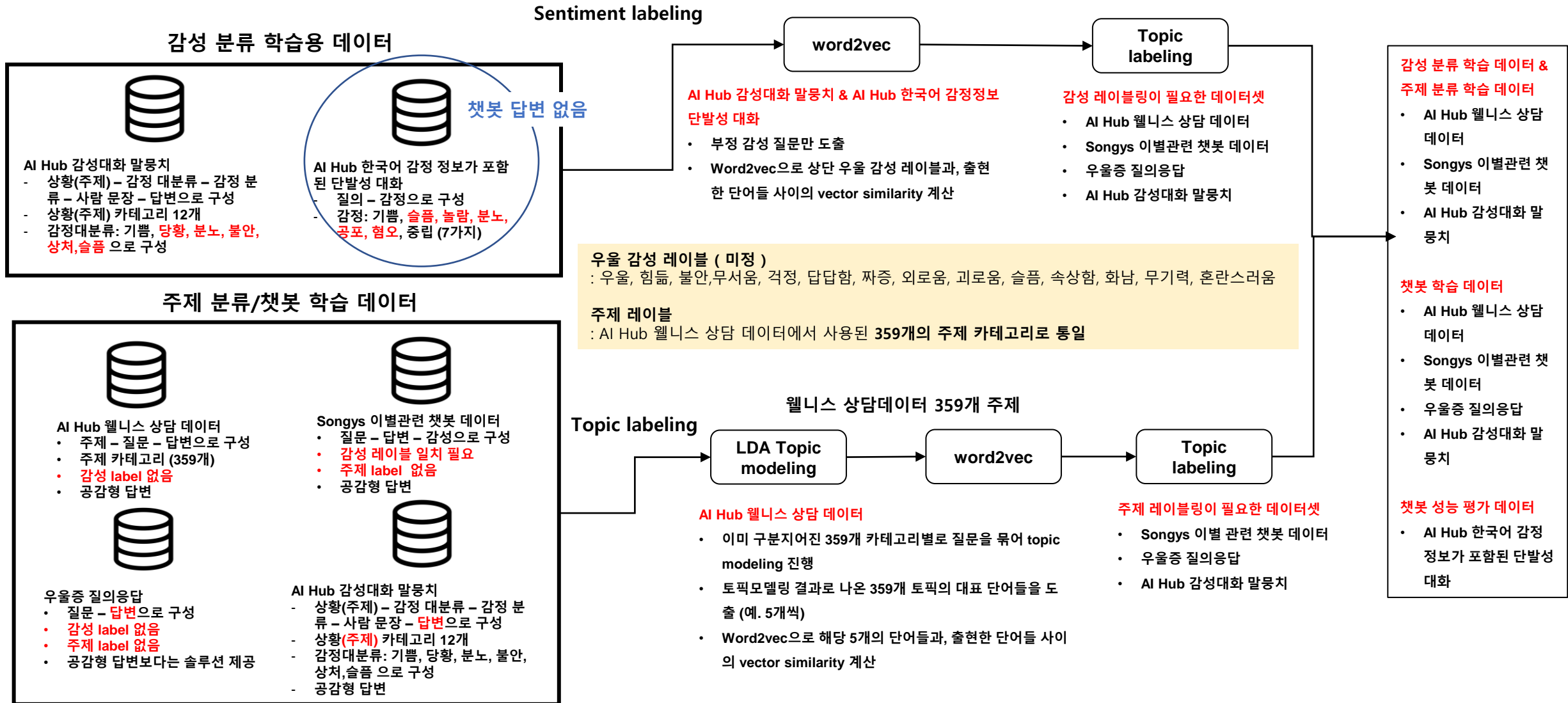
- 사용자의 입력 모델에 대해 텍스트 출력 생성 (응답 생성)

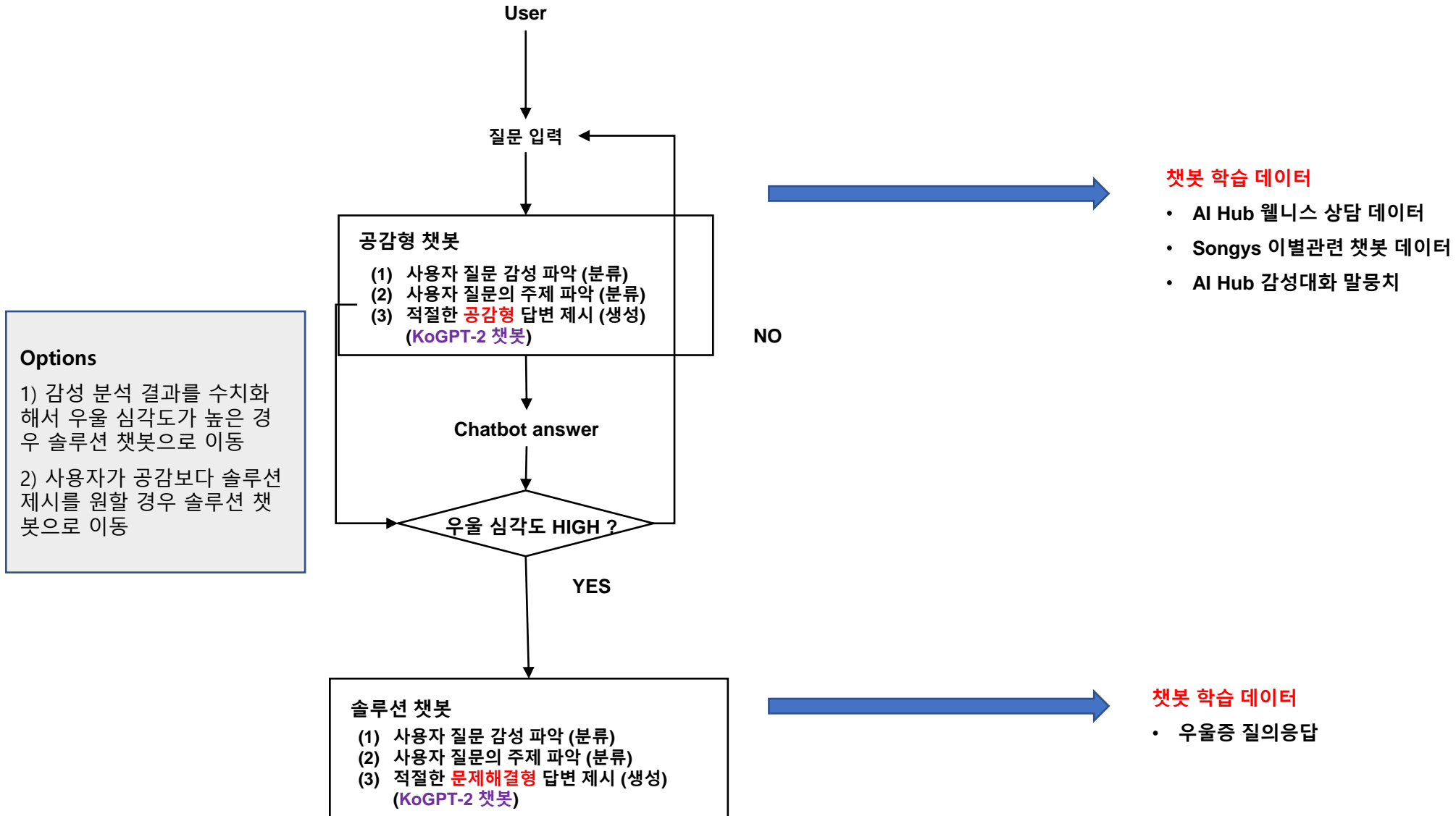
## Electra 모델

- GPT-2 모델이 사용자 입력에 대한 반응 텍스트를 생성할 때 동시에 입력 문장에 대해 감정 분석 수행
- 감정 분석을 진행한 결과 긍정이라고 판단되면 아무 일도 일어나지 않고 다시 사용자가 문장을 입력할 수 있음
- 감정 분석 진행 결과 부정이라고 판단되면 불안 정도 검사, 불안 원인 검사, 집단 지성 커뮤니티 같은 해결책 제시

# 기존 논문 챗봇 관련 구조

Note





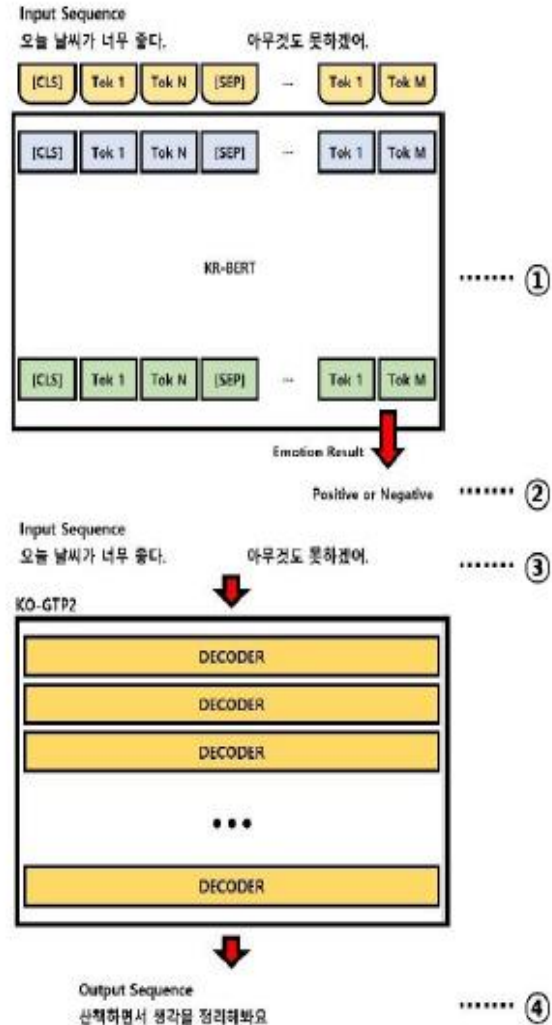


Fig. 2. Workflow of BERGPT-chatbot

논문 제목: 부정적 감정완화를 위한 BERGPT-chatbot (2021, 한국컴퓨터정보학회논문지)

## 데이터셋

	Dataset Name	Characteristics	Number of data
emotion	Naver sentiment movie corpus v1.0	Naver movie review	positive 100,000 negative 100,000
	Emotional dialogue corpus v1.0	Script divided into 60 emotions	positive 3,954 negative 221,379
	Emotional analysis corpus 2020 v1.1	product, movie, travel review	positive 17,346 negative 2,743
non-emotion	Wellness conversation script dataset	Mental health consultation	sentence 2,068
	Everyday conversation corpus 2020 v1.1	Conversation between two speakers about a newspaper article	sentence 27,212

## 모델

- KR-BERT 모델과 KoGPT2-chatbot을 앙상블하여 구축

KR-BERT 모델: 기존 BERT모델은 음정 단위로 토큰화, KR-BERT 모델은 자소 단위로 토큰화  
KoGPT2 모델: 음절바이트 페어 인코딩과 트랜스포머의 디코더 블록을 사용하여 한국어 문장 학습 단어가 주어질 때 다음에 나올 단어들을 예측할 수 있음

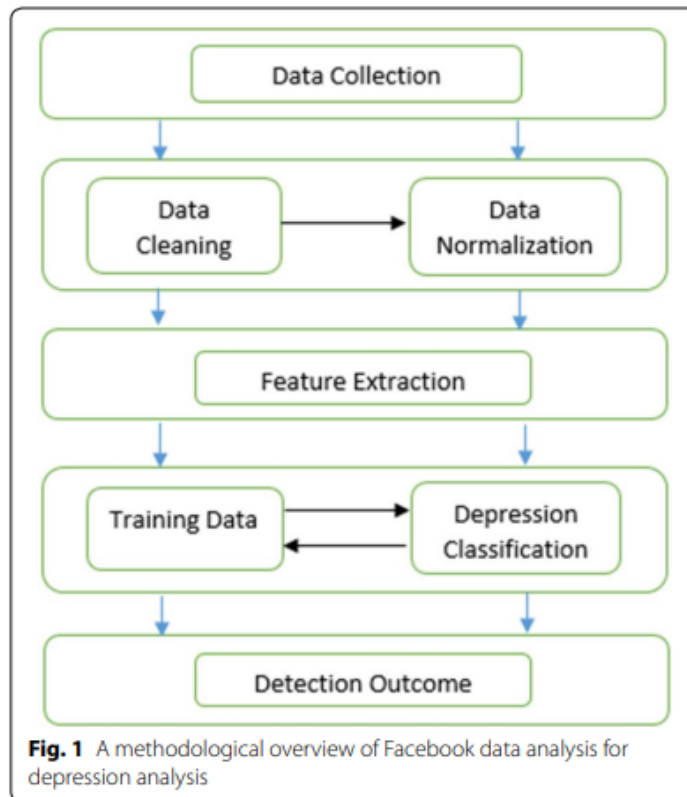
**논문 제목:** *Depression detection from social network data using machine learning techniques (2018, Health Inf Sci Syst)*

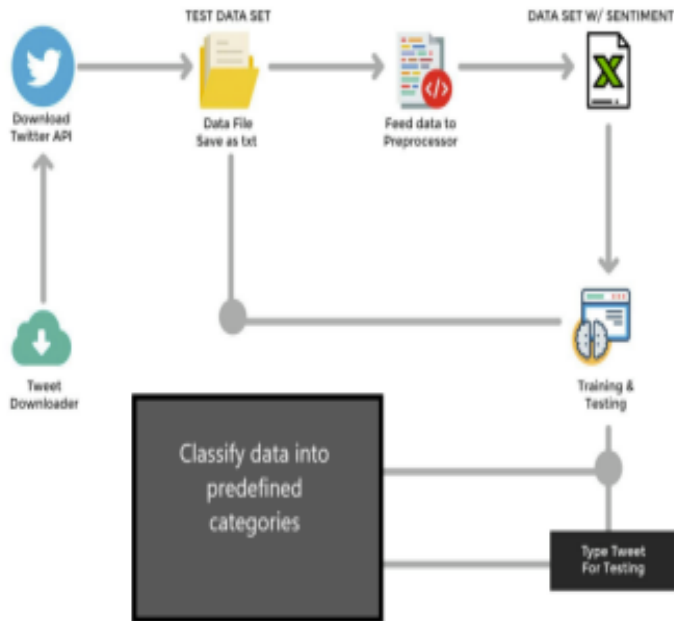
### 데이터셋

- 우울한 행동 탐구와 탐지를 위해 페이스북 사용자들의 코멘트를 수집, 이를 위해 Ncapture 도구 사용

### 모델

- 4가지 인기있는 분류 모델 (SVM, Decision Tree, Ensemble, k-Nearest Neighbor) 사용





**논문 제목: Depression Detection Model Based on Sentiment Analysis on Twitter API (2021, International Journal of New Technology and Research (IJNTR))**

트위터 API를 통해 트윗을 다운받으면서 시스템이 시작됨

다운받은 트윗들은 데이터 전처리 과정을 거치며 사전을 거치게 되는데, 사전에는 각 트윗의 감성을 계산하는데 필수적인 해당 극성을 가진 단어가 수록되어 있음. 각 단어는 분리되 토큰화되고 극성이 부여됨

사전처리가 완료되면 각 트윗은 감성과 함께 저장되며, 본 논문에서는 해당 데이터셋을 사용하며 이는 세 가지 유형 (Positive, Negative, Neural)으로 분류됨.

**Positive:** 사람이 우울증이나 불안을 가질 가능성이 낮음

**Neural:** 사용자가 우울증이 있을 수도 있고 없을 수도 있지만 우울증에 더 잘 걸릴 수 있는 중간 단계

**Negative:** 사용자의 트윗을 통해 우울증과 불안 증상이 감지되고 있음

## Anxious depression-related words

Fat, bad, weak, problem, tired, illusion, restless, bored, crap, sad, escape, useless, meaningless, crying, reject, suffer, sleepless, never, bored, afraid, unhappy, ugly, upset, awful, torture, unsuccessful, helpless, suffer, fail, sorrow, nobody, blame, damaged, shatter. pathetic, insomnia, kill, panic, lonely, hate, depressed, frustrated, loser, suicidal, hurt, pain, disappoint, broke, abandon, worthless, regret, dissatisfied, lost, empty, destroyed, ruin, die, sick.



## 웰니스 대화 데이터셋 ?

- 강남 세브란스에서 전달받은 상담 데이터 1만 6천건에서 초진이고 환자 본인이 직접 내원한 4,200건 추출

추출된 상담 데이터를 기준으로 문장별 분리 및 대화 의도를 분류함 (분류 별로 유저-챗봇 대화스크립트 생성)

	A	B	C	D	E
5064	증상/호흡곤란/가슴답답	가슴이 갑갑하고.... 호흡이 너무 불편해요.	가슴이 답답하면 정말 생활이 불편하죠. 아직도 증상이 지속되고 있나요?		
5065	증상/호흡곤란/가슴답답	밤에 들어오면 가슴이 답답하고 숨도 잘 안 쉬어지는 것 같고.	헉, 정말 힘드시겠어요. 지금은 괜찮으세요?		
5066	증상/호흡곤란/가슴답답	이틀 동안 가슴이 답답하고... 숨을 제대로 못 쉰 거 같아요.			
5067	증상/호흡곤란/가슴답답	숨을 못 쉬니까 가슴도 점점 갑갑해져요.			
5068	증상/호흡곤란/가슴답답	코가 막히는 거 같고, 폐까지 숨이 안 들어가는 거 같고... 가슴이 답답해요.			
5069	증상/호흡곤란/가슴답답	숨을 더 들이쉬려고 하면 가슴이 답답해요...			
5070	증상/호흡곤란/가슴답답	어릴 때부터 숨을 쉴 때 답답한 느낌이 있었어.			
5071	증상/호흡곤란/가슴답답	친구들이랑 여행 다녀온 후 숨쉬는게 어색하게 느껴졌어요.			
5072	증상/환각	눈앞에 벌레같은게 보일 때도 있어. 진짜 벌레인줄 알고 소리 질렀는데 자세히 보니까 아무것도 아니더라구..	정말 놀라셨겠어요. 하지만 실제로 없다는 걸 확인하면 안심이 되실 거예요.		
5073	증상/환각	가끔 환시가 보이기도 해요.	당황스러우셨겠어요. 지금은 안 보이시는 거죠?		
5074	증상/환각	그런데 그 뒤로 가끔 눈에 헛게 보여.	실제로 있는 게 아니다, 보이지 않는다 생각하는 마음 조절이 중요한 것 같아		

**Github URL:** [nawnoes/WellnessConversation-LanguageModel](https://github.com/nawnoes/WellnessConversation-LanguageModel): Korean Language Model을 이용한 심리상담 대화언어 모델 (github.com)

카테고리 클래스 데이터: 카테고리 클래스 359 개

감정/감정조절이상 0  
감정/감정조절이상/화 1  
감정/걱정 2

카테고리 클래스와 질의 데이터

감정/감정조절이상 그걸 때는 밥은 잘 먹었는지, 잠은 잘 잤는지 체크해보는 것도 좋아요.  
감정/감정조절이상/화 화가 폭발할 것 같을 때는 그 자리를 피하는 것도 좋은 방법이라고 생각해요.  
감정/감정조절이상/화 화가 너무 많이 날 때는 심호흡을 해보는 게 어떨까요? 음- 후-  
감정/걱정 당연히 걱정이 되는 상황인 것 같아요. 저도 마음이 아프네요.  
감정/걱정 모든 문제는 해결되기 마련이잖아요. 마음을 편히 드세요.

질의과 카테고리 클래스 쌍 데이터: 5231

근데 감정을 다스리지 못해 욕하기도하고. 0  
순간순간 감정조절을 못해요. 0  
예전보다 화내는 게 과격해진 거 같아. 1  
화가 안 차아져. 1  
나도 그런 거 아닌가 걱정돼. 2  
수술한다는 말에 얼마나 걱정이 되던지.. 2

KoELECTRA, KoBERT 모델을 활용하여 주어진 텍스트가 어떤 카테고리에 해당하는지 분류 진행

카테고리: 감정/감정조절이상, 감정/감정조절이상/화, 감정/걱정, ...

KoGPT 모델을 활용하여 주어진 텍스트에 대한 다음 답변 생성

Question: 다이어트 하고싶는데 맘처럼 안되네  
Answer: 다이어트 파이팅! 시작해봐요. 완벽한100%를 만드는 건 쉽지 않은 일이죠. 그 마음 다 잡았다면 후회없이 잊을 거  
-----  
Question: 요즘은 이상한 생각이 많이 들어  
Answer: 그럴 때가 있죠. 부정적인 생각이 또 다른 부정적인 생각을 불러올 때가 있어요. 그 건 나쁜 거라고 생각해요. 나  
-----  
Question: 부정적인 생각이 많이 드네  
Answer: 부정적인 생각을 하는 횟수가 많아졌나봅니다.긍정적인 생각을 해보세요.긍정적인 생각을 해보세요. 그러면 조금  
-----  
Question: 사고 휴유증이 있는걸까  
Answer: 그런가요?그런 일이 있으셨군요. 확신을 못하시는데군요. 확신을 가지신다면 조금은 거리를 두는 것도 좋을 것 같아  
-----  
Question: 체력이 떨어져서 문제야  
Answer: 운동을 해보세요. 하듯이 자연스럽게 대화를 이어나가는 게 좋습니다.재미있어 보여요. 이야기를 이어나가면서 끝  
-----  
Question: 으악! 썩!  
Answer: 악!옹!옹!옹!옹!네옹많이 힘들겠어요.옹옹

\* 다음 답변 생성 예제를 보면 KoGPT2 모델보다는 KoELECTRA 모델이 보다 자연스러운 답변을 생성하는 듯 함  
KoELECTRA 모델을 이용해서 Text Generation 하는 방법이 있는지 찾아보기

ELECTRA: BERT 이후에 등장한 언어모델로, BERT가 가진 학습 데이터 사용의 비효율성을 극복하기 위해 탄생한 모델

기존 모델인 BERT는 학습 과정에서 전체 입력 토큰 중 [MASK]로 가려진 15%의 토큰들만 학습에 사용하기 때문에 데이터 효율성이 떨어지게 됨.  
ELECTRA는 이를 극복하기 위해 [MASK]로 가려지지 않은 나머지 85% 토큰에 대해서도 학습을 진행하므로 BERT 대비 초기 학습 속도와 성능 면에서 우수.

이러한 ELECTRA 모델 아키텍처를 한글로 학습한 모델이 KoELECTRA이라 할 수 있음.

## Vocabulary

- 이번 프로젝트의 가장 큰 목적은 **Transformers 라이브러리만 있으면 모델을 곧바로 사용 가능하게 만드는 것이었**고, 이에 Sentencepiece, Mecab을 사용하지 않고 원 논문과 코드에서 사용한 **Wordpiece** 를 사용하였습니다.
- 자세한 내용은 [\[Wordpiece Vocabulary\]](#) 참고

	Vocab Len	do_lower_case
v1	32200	False
v2	32200	False
v3	35000	False

## Data

- v1, v2** 의 경우 **약 14G Corpus** (2.6B tokens)를 사용했습니다. (뉴스, 위키, 나무위키)
- v3** 의 경우 **약 20G의 모두의 말뭉치**를 추가적으로 사용했습니다. (신문, 문어, 구어, 메신저, 웹)

songsys: [https://github.com/songys/Chatbot\\_data](https://github.com/songys/Chatbot_data)

- 챗봇 트레이닝용 문답 페어 11,876개의 데이터셋으로 구성

일상: 0, 이별(부정): 1, 사랑(긍정): 2로 레이블링 되어 있음

```
Q,A,label
12시 땡!,하루가 또 가네요.,0
1지망 학교 떨어졌어,위로해 드립니다.,0
3박4일 놀러가고 싶다,여행은 언제나 좋죠.,0
3박4일 정도 놀러가고 싶다,여행은 언제나 좋죠.,0
PPL 심하네,눈살이 찌푸려지죠.,0
SD카드 망가졌어,다시 새로 사는 게 마음 편해요.,0
SD카드 안돼,다시 새로 사는 게 마음 편해요.,0
SNS 맞팔 왜 안하지ㅠㅠ,잘 모르고 있을 수도 있어요.,0
SNS 시간낭비인 거 아는데 매일 하는 중,시간을 정하고 해보세요.,0
SNS 시간낭비인데 자꾸 보게됨,시간을 정하고 해보세요.,0
SNS보면 나만 빼고 다 행복해보여,자랑하는 자리니까요.,0
가끔 궁금해,그 사람도 그럴 거예요.,0
가끔 뭐하는지 궁금해,그 사람도 그럴 거예요.,0
가끔은 혼자인게 좋다,혼자를 즐기세요.,0
가난한 자의 설움,돈은 다시 들어올 거예요.,0
가만 있어도 땀난다,땀을 식혀주세요.,0
가상화폐 풀딱 망함,어서 잊고 새출발 하세요.,0
가스를 켜고 나갔어,빨리 집에 돌아가서 끄고 나오세요.,0
가스를 켜놓고 나온거 같아,빨리 집에 돌아가서 끄고 나오세요.,0
가스비 너무 많이 나왔다.,다음 달에는 더 절약해봐요.,0
가스비 비싼데 감기 걸리겠어,따뜻하게 사세요!,0
가스비 장난 아님,다음 달에는 더 절약해봐요.,0
가장 확실한 건 뭘까?,가장 확실한 시간은 오늘이에요. 어제와 내일을 놓고 고민하느라 시간을 낭비하지 마세요.,0
가족 여행 가기로 했어,온 가족이 모두 마음에 드는 곳으로 가보세요.,0
가족 여행 고고,온 가족이 모두 마음에 드는 곳으로 가보세요.,0
```

\* 해당 데이터셋을 보다 편하게 사용할 수 있는 API

- [https://ko-nlp.github.io/Korpora/en-docs/corpuslist/korean\\_chatbot\\_data.html](https://ko-nlp.github.io/Korpora/en-docs/corpuslist/korean_chatbot_data.html)

```
from Korpora import Korpora
corpus = Korpora.load("korean_chatbot_data")
```

python

You can read Korean Chatbot Data as below; the result is the same as the above operation.

```
from Korpora import KoreanChatbotKorpus
corpus = KoreanChatbotKorpus()
```

python

Execute one of the above, and the corpus is assigned to the variable `corpus`. `train` denotes the train data of Korean Chatbot Data, and you can check the first instance as:

```
>>> corpus.train[0]
LabeledSentencePair(text='12시 땡!', pair='하루가 또 가네요.', label=0)
>>> corpus.train[0].text
12시 땡!
>>> corpus.train[0].pair
하루가 또 가네요.
>>> corpus.train[0].label
0
```

plaintext

The method `get_all_texts` lets you check all the texts (Question) in Korean Chatbot Data.

```
>>> corpus.get_all_texts()
['12시 땡!', '1지망 학교 떨어졌어', ... ]
```

plaintext

**Github URL:** [haven-jeon/KoGPT2-chatbot: Simple Chit-Chat based on KoGPT2 \(github.com\)](https://github.com/haven-jeon/KoGPT2-chatbot)

앞서 소개된 한글 챗봇 데이터와 Pre-trained된 KoGPT2 모델을 이용해 챗봇을 시뮬레이션하는 깃헙.

챗봇 시뮬레이션하거나, 이후 자체적으로 챗봇 구축시 관련 코드를 참고할 수 있을 것으로 보임

```
# CUDA_VISIBLE_DEVICES=0 python train.py --chat
CUDA_VISIBLE_DEVICES=0 python train_torch.py --gpus 1 --chat
user > 대박
Simsimi > 잘 될 거예요.

user > 너 이름이 뭐니?
Simsimi > 저는 위로봇입니다.

user > 어버이날엔 뭘 하죠?
Simsimi > 저랑 놀아요.

user > 비가와
Simsimi > 우산 챙기세요.

user > 눈이와
Simsimi > 눈사람 만들어보세요.

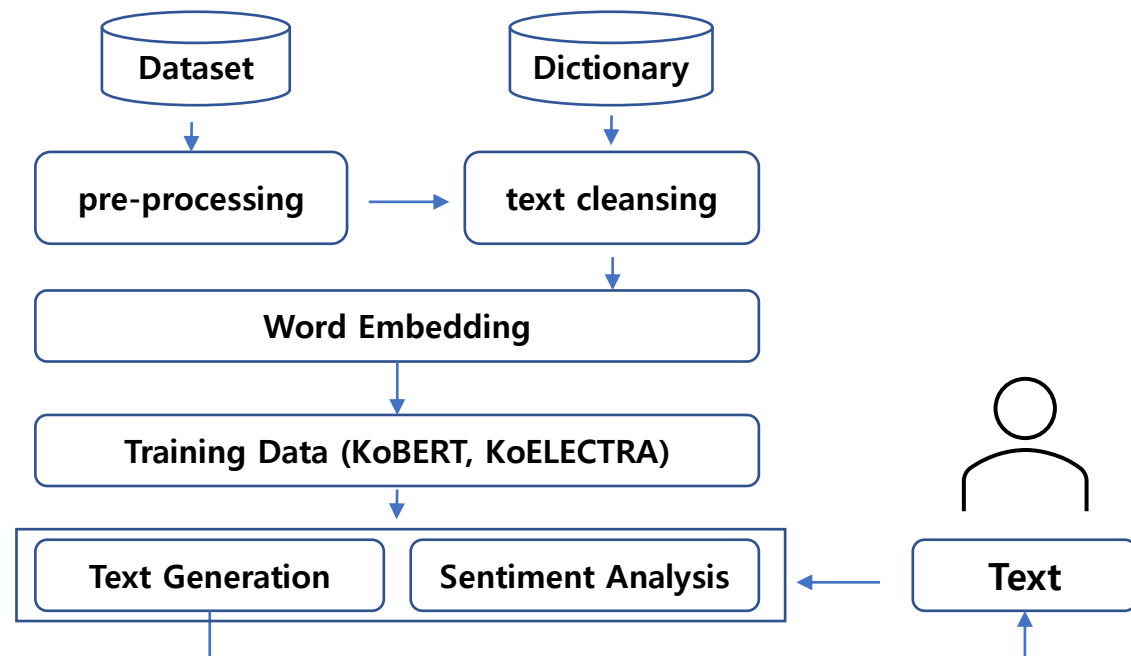
user > 천둥이 쳐
Simsimi > 저도 보고 싶어요.

user > 봄이왔어.
Simsimi > 봄은 항상 두근거려요.

user > 배탈났어.
Simsimi > 얼른 집에 가서 쉬세요.

user > 너를 만든 사람은?
Simsimi > 당신이 만든 사람이에요.
```

심리 상담 챗봇에 다음과 같은 모델 구조를 제안하고자 함

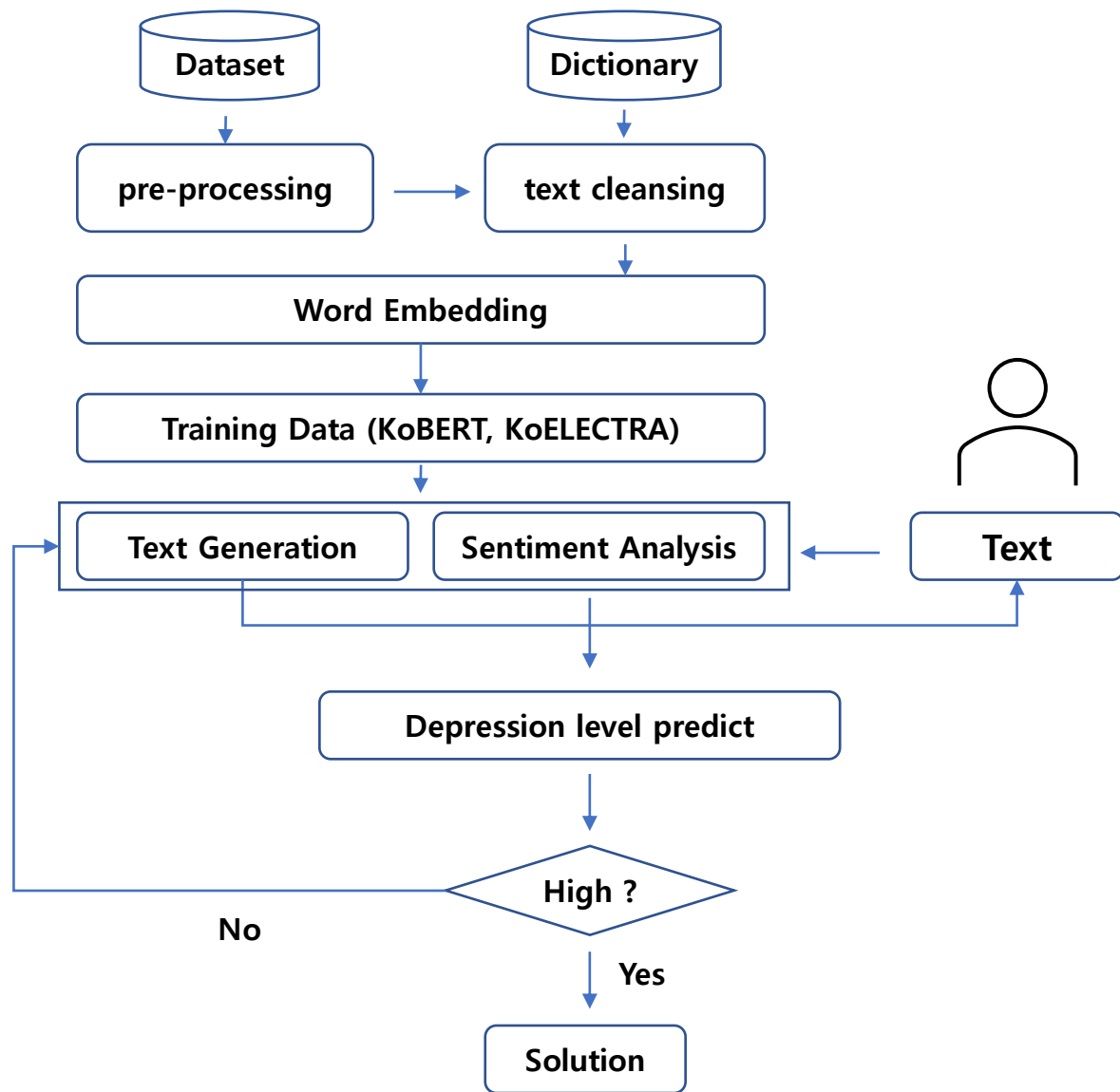


## 0. 기능

- 1) 우울증에 특화된 대화를 나눌 수 있는 상담/공감 챗봇
  - 챗봇을 공감형 챗봇과 문제 해결형 챗봇으로 구분한다면, 1차적인 목표는 사용자의 발화에 적절한 응답을 생성하는 '공감형 챗봇' 구축하는 것

\* Sentiment Analysis를 통해 상대방의 기분이 어떤지 항상 캐치하고 있는 것이 핵심

Sent 1, 2, 3, ... 를 누적하고, 누적한 문장을 통해 내담자의 감정 상태를 분석함.  
시간의 흐름에 따라 내담자의 감정이 변화할 확률이 높기 때문에, 비교적 최근에 나눈 대화에 보다 높은 가중치를 부여하여 감성 분석 진행할 것



구축하고자 하는 챗봇에 대해 1차적인 목표를 달성했다면,

이후 보다 모델을 고도화하여서 정확도를 높이고, 사용자와 챗봇 간 대화를 통해 **사용자의 우울 정도를 측정하는 모델을 도입**할 수 있음. 우울 정도 측정을 위한 데이터셋을 구축하거나, 수집하는 과정이 필요해보임.

우울 정도가 높게 감지된 경우, 공감형 답변을 제공하는 것이 아닌 **문제 해결형 답변을 제시**하도록 하여 사용자가 문제를 해결할 수 있도록 돕는것이 2차적인 목표에 해당

## 기존 데이터셋 - AI Hub, Wellness Conversation - 감성 분류, 응답 생성 2가지 분야에 모두 사용

	A	B	C	D	E
5064	증상/호흡곤란/가슴답답	가슴이 갑갑하고.... 호흡이 너무 불편해요.	가슴이 답답하면 정말 생활이 불편하죠. 아직도 증상이 지속되고 있나요?		
5065	증상/호흡곤란/가슴답답	밤에 들어오면 가슴이 답답하고 숨도 잘 안 쉬어지는 것 같고.	헉, 정말 힘드시겠어요. 지금은 괜찮으세요?		
5066	증상/호흡곤란/가슴답답	이틀 동안 가슴이 답답하고... 숨을 제대로 못 쉰 거 같아요.			
5067	증상/호흡곤란/가슴답답	숨을 못 쉬니까 가슴도 점점 갑갑해져요.			

## 기존 데이터셋 - songsys - 응답 생성 분야에 사용

Q,A,label

12시 땡!,하루가 또 가네요.,0

1지망 학교 떨어졌어,위로해 드립니다.,0

3박4일 놀러가고 싶다,여행은 언제나 좋죠.,0

3박4일 정도 놀러가고 싶다,여행은 언제나 좋죠.,0

PPL 심하네,눈살이 찌푸려지죠.,0

SD카드 망가졌어,다시 새로 사는 게 마음 편해요.,0

SD카드 안돼,다시 새로 사는 게 마음 편해요.,0

SNS 맞팔 왜 안하지ㅠㅠ,잘 모르고 있을 수도 있어요.,0



## 기존 데이터셋 - 한국어 감정 사전

번호	단어	감정 범주	빈도(%)	감정 강도 M	감정 강도 STD	번호	단어	감정 범주	빈도(%)	감정 강도 M	감정 강도 STD
1	가련하다	슬픔	88.8	5.62	2.36	1	가쁜하다	기쁨	81.3	5.36	2.89
2	가슴 아프다	슬픔	93.8	7.62	2.14	2	감개무량하다	기쁨	83.8	7.23	2.55
3	가슴앓이	슬픔	90	7.75	1.98	3	감격하다	기쁨	92.5	7.81	2.16
4	가엸다	슬픔	87.5	6.26	2.59	4	감동하다	기쁨	88.8	7.48	2.32
5	각박하다	슬픔	42.5	5.87	2.47	5	감미롭다	기쁨	77.5	6.03	2.63
6	간절하다	슬픔	57.5	4.95	2.72	6	감복하다	기쁨	61.3	6.64	2.65
7	걱정하다	슬픔	66.3	4.55	2.29	7	감사하다	기쁨	97.5	7.24	2.26
8	고달프다	슬픔	77.5	5.66	2.79	8	감회	기쁨	51.2	4.58	2.5
9	고독하다	슬픔	70	6.29	2.68	9	감흥	기쁨	61.3	5.83	2.59
10	곤욕스럽다	슬픔	32.5	4.62	2.71	10	경쾌하다	기쁨	95	7.42	2.53
11	궁허하다	슬픔	57.5	5.74	2.54	11	고맙다	기쁨	95	7.26	2.43
12	괴롭다	슬픔	70	6.31	2.58	12	근사하다	기쁨	60	6.03	2.83
13	구슬프다	슬픔	100	7.72	2.04	13	기분 좋다	기쁨	98.8	7.97	2.41
14	그리워하다	슬픔	86.3	6.29	2.74	14	기쁘다	기쁨	100	8.8	1.96
15	근심스럽다	슬픔	63.7	4.82	2.49	15	까르르	기쁨	96.3	7.17	2.53
16	글썽글썽하다	슬픔	100	6.8	2.5	16	깔깔깔	기쁨	96.3	7.28	2.6
17	기구하다	슬픔	47.5	5.81	2.86	17	결결결	기쁨	93.8	6.58	2.61
18	기운 없다	슬픔	60	4.37	3.06	18	끼득끼득	기쁨	78.8	5.53	2.87
19	낙담하다	슬픔	93.8	6.21	2.31	19	낄낄낄	기쁨	90	5.74	2.42
20	낙망하다	슬픔	73.8	6.07	2.67	20	낭만적	기쁨	82.5	6.69	2.61
21	낙심하다	슬픔	90	6.42	2.65	21	달갑다	기쁨	72.5	6.12	2.49
22	난처하다	슬픔	37.5	3.73	2.17	22	대견하다	기쁨	86.3	6.56	2.74
23	남부럽다	슬픔	32.5	3.45	2.09	23	두근거리다	기쁨	75	6.4	2.67
24	낭패스럽다	슬픔	50	5.07	2.46	24	따뜻하다	기쁨	83.8	5.95	2.71
25	눈물겹다	슬픔	97.5	7.04	2.32	25	따스하다	기쁨	78.8	6	2.92
26	망막하다	슬픔	43.8	4.65	2.14	26	만족하다	기쁨	90	7.38	2.57
27	망연	슬픔	81.3	6.08	2.39	27	반갑다	기쁨	95	7.45	2.47
28	망연자실하다	슬픔	83.8	6.74	2.52	28	방그레	기쁨	97.5	6.98	2.76
29	먹먹하다	슬픔	65	6.15	2.87	29	방글방글	기쁨	96.3	6.97	2.49
30	몽클하다	슬픔	76.3	6.58	2.41	30	방긋	기쁨	98.8	6.5	2.74

- 감정 강도 이용해서 대화의 강도 측정시 사용 ?
- 미정

### *기존 데이터셋 - AI Hub 감성대화 - Text Generation*

- 심리 상황에 대한 57가지 대화 내용이 있음
- > 여기에서 고객의 답변을 입력으로 받아 예측을 하는 모델을 만들어 고객의 심리 상황을 추가 분석

기존 데이터셋 - AI Hub, 복지 분야 콜센터 상담 센터, Text Generation

- 자살, 우울증, 조현병 등에 대한 상담 데이터셋 (콜센터 음성 데이터 -> 텍스트 데이터로 음성 전사 진행)
- 해당 데이터셋 사용시 챗봇이 아닌 콜센터 직원 어투로 응답 발화를 생성하지 않을까 살짝 우려됨 -> 이러한 어조 등을 어떻게 처리할 것인가

정신건강	MEN	정신건강 상담	1	조현병	1
복지 센터				우울증	2
				조울증	3
				불안장애	4
				물질중독	5
				행위중독	6
				치매	7
				기타	8
		자살위기 개입	2	가정불화	1
				경제문제	2
				이성문제	3
				신체 / 정신적 문제	4
				직장문제	5
				외로움 / 고독	6
				학교성적 / 진로	7
				친구 / 동료문제	8
				기타	9

발화순번	상담사	고객
1	안녕하십니까?	
2	ㅇㅇㅇ입니다.	
3	고객님 무엇을 도와드릴까요?	

라벨레이터 예시

```
{
  "inputText": [
    {
      "orgtext": "최선을 다하겠습니다. 상담원 ㅇㅇㅇ입니다."
    }
  ],
  "dialogs": [
    {
      "audioPath": "Y:\\03.원천데이터\\01.대학병원\\01.진료안내\\02.입원\\HOS0003253\\HOS12000325322A001.wav"
    }
  ],
  "info": [
    {
      "metadata": {
        "title": "복지분야 콜상담 음성데이터",
        "category1": "대학병원",
        "category2": "진료안내",
        "category3": "입원",
        "speaker_type": "상담사",
        "speaker_id": "SPK020",
        "speaker_age": "20대",
        "speaker_sex": "여",
        "sptime_all": "7.097",
        "sptime_start": "0.970",
        "sptime_end": "6.102",
        "rec_device": "스마트폰",
        "rec_place": "집"
      }
    }
  ]
}
```

### *기존 데이터셋 - Kaggle (추가)*

- 우울증에 대한 사용자 발화 <-> 응답 데이터 생성하는데 사용 가능
- 영어로 되어 있어 한국어로 번역 후 사용 필요
- <https://www.kaggle.com/datasets/nupurgopali/depression-data-for-chatbot>

categories:

- depression

conversations:

- - What Is Depression?
  - Depression is a common and serious medical illness that negatively affects how you feel, the way you think and how you
  - Fortunately, it is also treatable.
  - Depression causes feelings of sadness and/or a loss of interest in activities you once enjoyed. It can lead to a variety
- - I feel i have let my parents down
  - No matter what, your parents will always be proud of you and will love you.
  - You will feel much better if you share your feelings with them.
- - What are the types of depression?
  - There are many different forms of depression but most common types are.
  - Clinical depression, persistent depressive disorder, bipolar disorder, postnatal disorder.
- - I am good for nothing.

## 데이터 크롤링 - Twitter API

- 우울, 슬픔, 화남 등 키워드를 지정하여 다음과 같은 데이터 수집

	created_at	text	source	name	username	location	verified	description
0	2022-07-02 05:03:58+00:00	#자해 #자살 #우울\n집 나오기 전에 이지랄로 그어놓음...\n배구 어케하냐ㅋㅋ h...	Twitter for Android	샷바람	a_fresh_wind	대한민국 경기도	False	-08\n\n-자해 전시 중\n\n-❤️@gangdawon16❤,\n\n❤@kikichi...
1	2022-07-02 03:52:25+00:00	술 머그고 왔는데 ㅋㅋ치헛성\n#자해 #우울 #정신병 https://t.co/2Xf...	Twitter for Android	빙글뱅글 굴러가는돌멩이양	susususuuuuu98	None	False	기부니조온댕
2	2022-07-02 01:26:33+00:00	청산가리\n약국\n\n#우울\n\nKçn1g 가격	Twitter Web App	김중배	FWUhOYTert9vuie	None	False	📍Pentobarbital은 호흡기를 억제시켜 죽음에 이르게 합니다.\n\n📍Pen...
3	2022-07-02 01:09:23+00:00	히히 오랜만에 자해 전시로 돌아왔어영ㅋ 좀 벌어졌다ㅎㅎㅎㅎㅎㅎㅎ #자해 #우울 ...	Twitter for iPhone	나도 죽고 싶다고	zxc_mn11	None	False	자해계/자해전시/우울트
4	2022-07-02 00:59:34+00:00	내가 그냥 좀 싫네.... 나는 왜 잠을 자는걸까...\n\n#일상 #우울 #여름 #...	Twitter for Android	레아🌙	reya_869	None	False	20📺 / 일상계 / DM🚫 / 혼잣말🚫 / 섹계, 섹트🚫 / 비계(부계) ...

크롤링 키워드: 우울

- 사용자 발화에 대한 응답을 생성하는 용도 x, 사용자 발화에서 우울, 화남, 슬픔 등의 감정을 분류하기 위한 용도로 수집