●○ 자연어 영역

감성 대화 말뭉치

●○ 개요 : 감성 AI 코퍼스 데이터셋이란?

본 프로젝트는 AI 기반 감성 챗봇용 세대별 감성대화 텍스트 데이터의 구축이 주목적이며, 궁극적으로는 딥러닝 기반 언어 이해(NLU) 엔진인 ALBERT 모델링을 위한 데이터를 수집하는 것이다. 따라서 본 프로젝트 구축 데이터는 텍스트 파일(json) 형식으로 구축된다.

본 구축 데이터를 통해 한국인의 감성대화 엔진 또는 챗봇을 개발하려는 기업, 연구기관, 연구자 등에게 기술 개발의 리소스를 제공하며, 나아가 독거노인 돌봄 서비스 등을 제공하고 있는 기관 역시 궁극적인 수혜자가 될 수 있다.

본 데이터 구축은 시나리오 설계, 데이터 수집, 데이터 태깅, 데이터 검수의 과정으로 진행되며, 중간 결과물을 통한 NLU 엔진 모델링을 수행하여 모델 성능을 확인한다. 이어 대규모 데이터 구축을 수행한 뒤 공인 인증 기관의 품질 검증 과정을 거쳐 구축 데이터 및 NLU 모델을 공개한다.

단계	수행주제		내용
구축	어노테이션	크라우드 소싱	어노테이션과 라벨링은 동일한 개념이며, 페르소나 및 감정 상황의 조합이 설계된 후 코퍼스 수집에 착수
	라벨링		
	코퍼스 수집		미리 설계되고 어노테이션된 케이스에 대한 데이터수집
	1차검수		기초적인 문법 검사
	2차검수	주관기관	설계 상황에 대한 내용 검수
	태깅	크라우드 소싱	별도의 도구를 사용하여 검수
	3차 검수	주관기관	모델링을 위한 적정성 검토
	모델링	주관기관	인공지능 학습 과정

●○ 데이터셋의 구성

본 데이터셋은 감성 AI 코퍼스 데이터에 대해 대본을 생성하고, 해당 대본에 대해 60가지 감정으로 다양하게 구성되어 있다. 이에 따른 전사 텍스트 데이터는 JSON 형식으로 이루는데, 이 전사 텍스트 안에 각종 메타 데이터를 기록한다. 감성 코퍼스 데이터셋 27만 문장은 연구자가 일반적으로 연구를 진행하기에 충분한 양이며, 상용화 수준에서는 의미 있는 ALBERT 모델을 만들 수 있는 양이다.

데이터를 통해 한국인의 감성대화 엔진 또는 챗봇을 개발하려는 기업, 연구기관, 연구자 등에게 기술 개발의 리소스를 제공하며, 나아가 독거노인 돌봄 서비스 등을 제공하고 있는 기관에 활용이 가능하다.

NLU 엔진의 성능을 개선하기 위한 ALBERT 모델의 알고리즘을 활용하면, 형태소 단위의 세부 태깅을 하지 않고 문장 단위의 정합성만 검수하면 데이터 모델링에 큰 문제가 없다는 점이 장점이며, 문장에서 의미와 의도를 추출하는 확률이 기존 통계 모델링 기법(CRF+ 등)에 비해 월등히 높은 성능을 보여준다.

데이터 종6류	포함 내용	제공 방식
감성 AI 코퍼스	60가지 감정 상태가 포함된 대화 문장 전사 텍스트 (27만 문장)	JSON 포맷 파일
테스트용 음성	10,000문장 분량의 음성 테스트 데이터	WAVE 포맷 파일

●○ 데이터 구조와 예시

감성 AI 대화 말뭉치 데이터 파일은 json 형식으로 저장되며, 아래의 구조를 따른다.

| json 형식 샘플 |

```
{
"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_00023"
                    },
          "content": {
                    "HS01": "이번 프로젝트에서 내가 발표 실수를 해서 우리 팀이 감점을 받아서 너무
                              미안해.",
                    "SS01": "실수하시다니 정말 죄송한 마음이 크겠어요.",
"HS02": "내 능력이 부족한 거 같은데 그만 다녀야 하려나 봐.",
"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."
                }
        }
 }
```


감성 AI 대화 말뭉치 데이터는 기본적으로 3턴의 대화를 기준으로 하며, 사용자와 시스템 응답까지 최대 6개의 문장으로 1개의 대화 묶음이 이루어진다. 이때 대화의 상황을 만드는 요인은 상황에 대한 세부 항목, 질병, 감정 상태 등에 따라 분류된다.

구분	항목
대화 턴 (최대 3턴 = 6문장)	1. 사람 대화 (감정 상태)
	2. 시스템 응답 (응답 호응)
	3. 사람 대화 2 (진전된 대화)
	4. 시스템 응답 2 (응답 호응)
	5. 사람 대화 3 (진전된 대화)
	6. 시스템 응답 3 (마무리 대화)

감정 상태는 6개의 기본 감정을 기준으로 각각 9개의 세부 감정이 포함된 총 60개 감정 상태가 반영된다. 이때 첫 번째 발화에는 이러한 감정이 드러나도록 문장을 생성하는데, 첫 문장 발화에 대한 시스템 응답 및 사용자의 자연스러운 후속 대화에서는 상황에 따라 감정 상태가 포함되지 않은 문장이 출현할 수 있다.

60가지 감정 분류						
기분	분노	슬픔	불안	상처	당황	기쁨
1	툴툴대는	실망한	두려운	질투하는	고립된	감사하는
2	좌절한	비통한	스트레스 받는	배신당한	남의 시선 의식하는	사랑하는
3	짜증나는	후회되는	취약한	고립된	외로운	편안한
4	방어적인	우울한	혼란스러운	충격 받은	열등감	만족스러운
5	악의적인	마비된	당혹스러운	불우한	죄책감	흥분되는
6	안달하는	염세적인	회의적인	희생된	부끄러운	느긋한
7	구역질 나는	눈물이 나는	걱정스러운	억울한	혐오스러운	안도하는
8	노여워하는	낙담한	조심스러운	괴로워하는	한심한	신이 난
9	성가신	환멸을 느끼는	초조한	버려진	혼란스러운	자신하는

감정의 상태는 위의 그림과 같이 총 60가지로 구분하며, 해당 감정 상태에 대한 개개인의 페르소나를 생성하여 대화 코퍼스를 수집한다.

페르소나는 연령과 성별로 크게 구분하며, 사람의 대화에 대한 챗봇 시스템의 응답으로 구성된다.

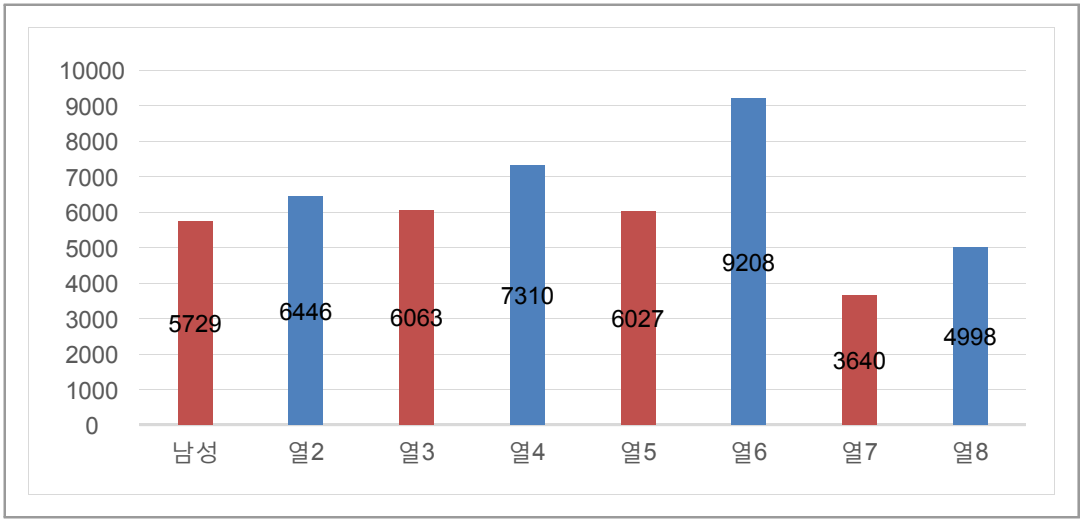
구분	항목	상세
페르소나	연령 (A)	청소년
		청년
		중년
		노년
	성별 (G)	남성
		여성
	시스템 응답 (C)	응답

●○ 데이터셋의 설계 기준과 분포

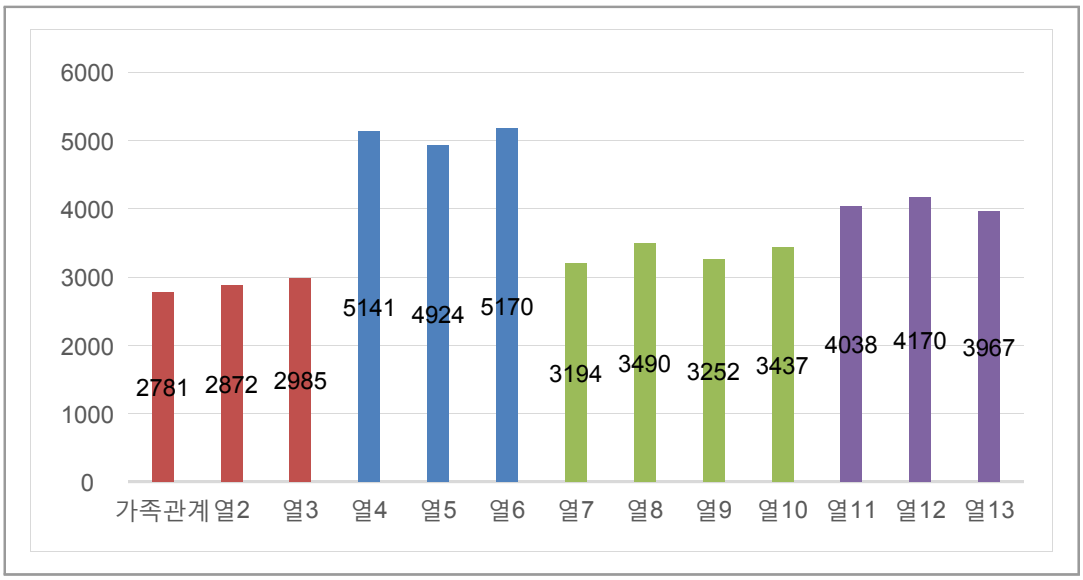
데이터셋을 설계할 때 가장 중요하게 고려했던 점은 데이터 밸런스이다. 본문과 질문, 정답을 구성할 때 적절한 분류기준을 만들었고, 해당 분류기준에 따라 골고루 데이터가 분포되도록 설계하여 학습 시 예상할 수 있는 데이터 편향성을 최소화하도록 했다.

감성 AI 말뭉치 데이터는 아래와 같은 분포로 구성된다.

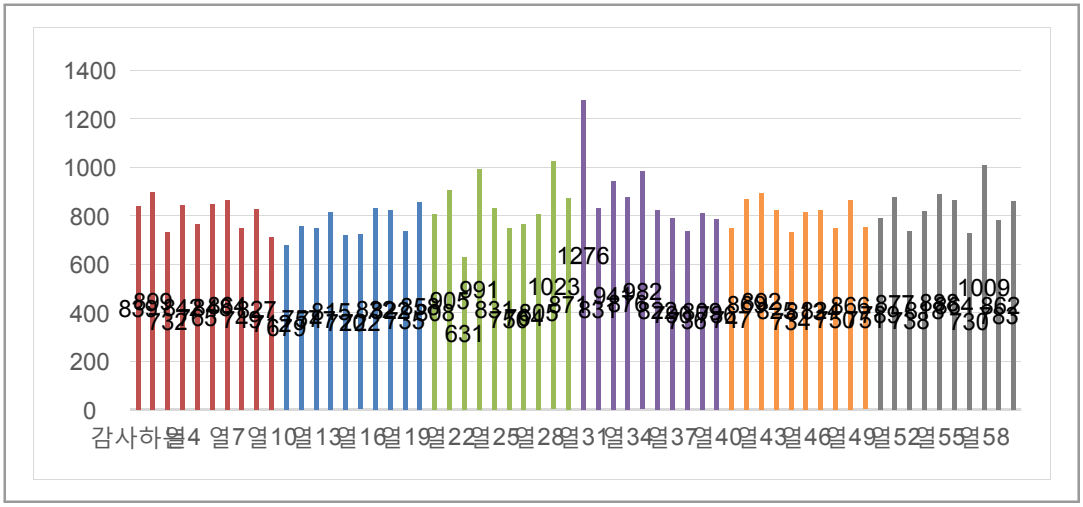
(1) 연령별 및 성별 분포



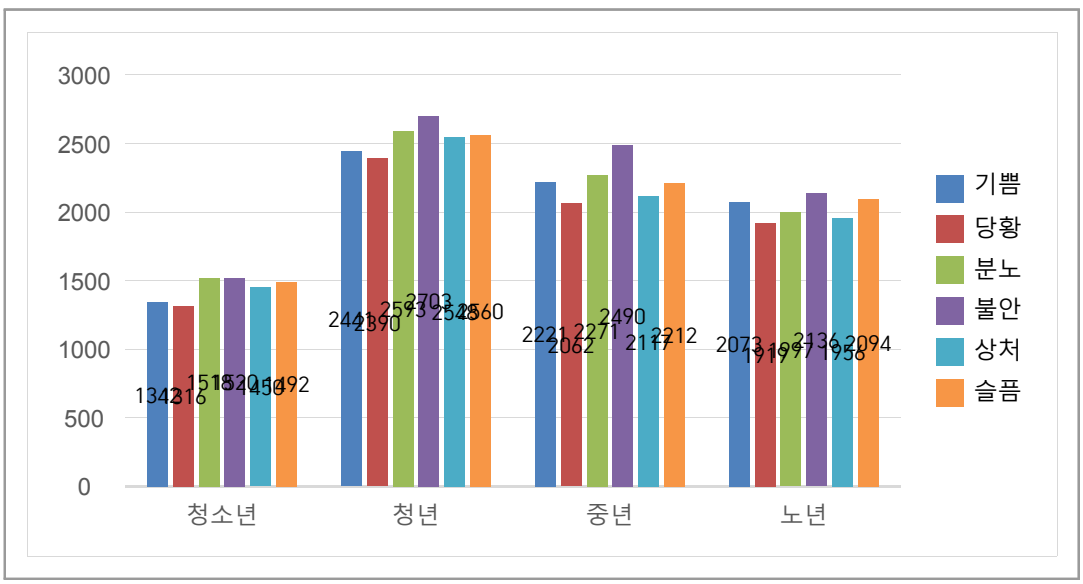
(2) 연령별 상황 키워드 분포



(3) 감정 상태별 분포

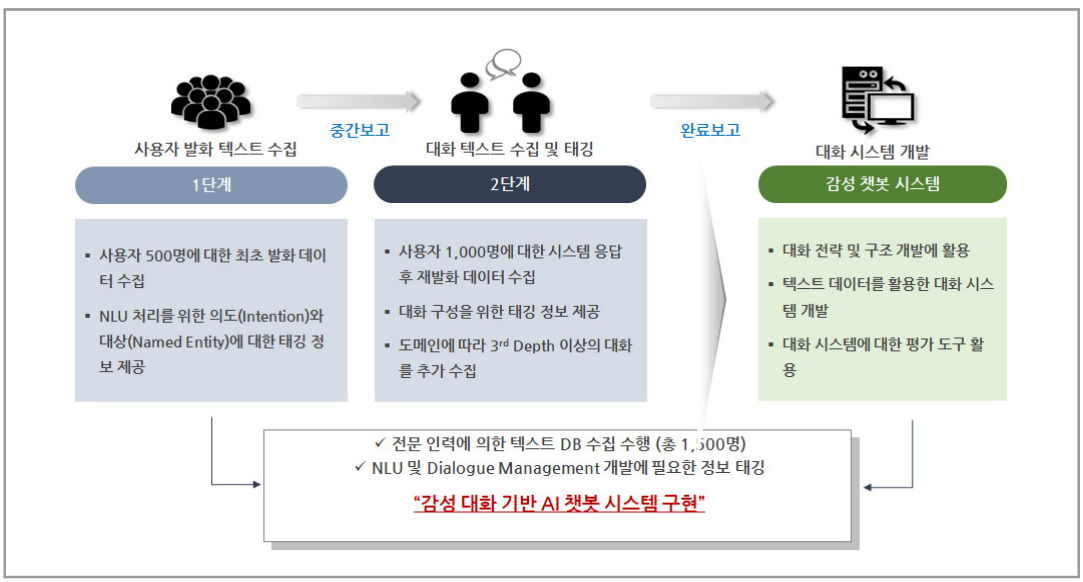


(4) 연령별 감정 분포도



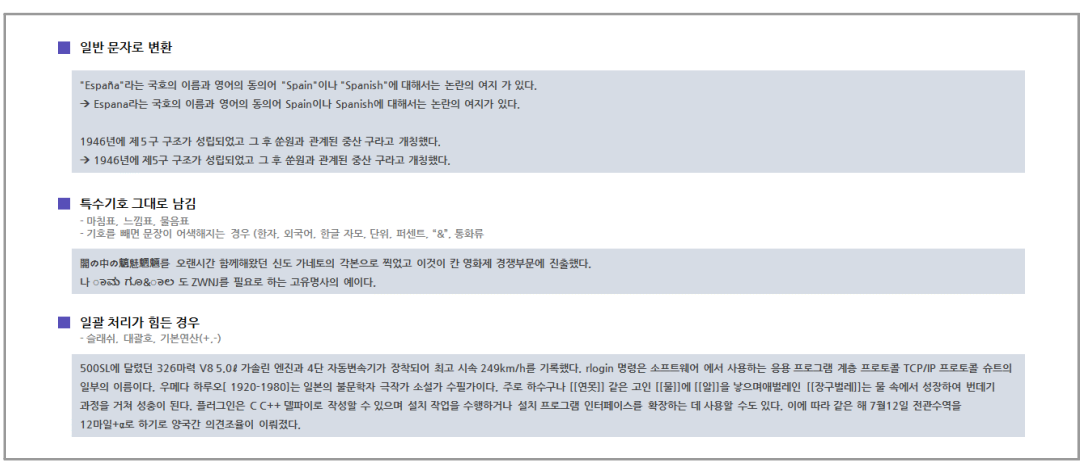
데이터 구축 과정

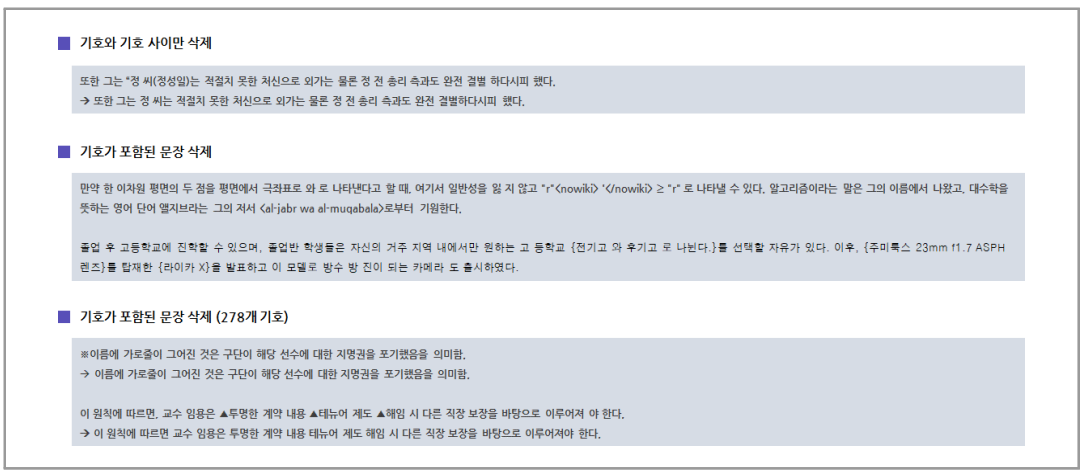
데이터 구축은 다음의 과정을 거쳐서 수행되었다.



데이터 수집

수집된 데이터는 다양한 후처리를 수행하여 정제한다. 문자의 변환, 특수기호 처리, 오탈자 수정 등 다양한 정제 과정을 거치는데, 본 프로젝트의 목적이 텍스트 코퍼스를 수집하는 것이기 때문에 주로 일반적인 텍스트 정제 방법론이 활용된다.





■ 기호와 기호 사이만 삭제

또한 그는 "정 씨(정성일)는 적절치 못한 처신으로 외가는 물론 정 전 총리 측과도 완전 결별 하다시피 했다.
→ 또한 그는 정 씨는 적절치 못한 처신으로 외가는 물론 정 전 총리 측과도 완전 결별하다시피 했다.

■ 기호가 포함된 문장 삭제

만약 한 이차원 평면의 두 점을 평면에서 극좌표로 와 로 나타낸다고 할 때, 여기서 일반성을 잃 지 않고 "r"<nowiki> '</nowiki> ≥ "r" 로 나타낼 수 있다. 알고리즘이라는 말은 그의 이름에서 나왔고, 대수학을 뜻하는 영어 단어 앨지브라는 그의 저서 <al-jabr wa al-muqabala>로부터 기원한다.

졸업 후 고등학교에 진학할 수 있으며, 졸업반 학생들은 자신의 거주 지역 내에서만 원하는 고 등학교 {전기고 와 후기고 로 나뉜다.}를 선택할 자유가 있다. 이후, {주미룩스 23mm f1.7 ASPH 렌즈}를 탑재한 {라이카 X}을 발표하고 이 모델로 방수 방 진이 되는 카메라 도 출시하였다.

■ 기호가 포함된 문장 삭제 (278개 기호)

※이름에 가로줄이 그어진 것은 구단이 해당 선수에 대한 지명권을 포기했음을 의미함.
→ 이름에 가로줄이 그어진 것은 구단이 해당 선수에 대한 지명권을 포기했음을 의미함.

이 원칙에 따르면, 교수 임용은 ▲투명한 계약 내용 ▲테뉴어 제도 ▲해임 시 다른 직장 보장을 바탕으로 이루어져 야 한다.
→ 이 원칙에 따르면 교수 임용은 투명한 계약 내용 테뉴어 제도 해임 시 다른 직장 보장을 바탕으로 이루어져야 한다.

본 프로젝트의 코퍼스 데이터는 ALBERT 알고리즘의 모델링을 목표로 하고 있기 때문에, 문장 단위의 정제가 진행되며, 의미의 분별이 분명하도록 띄어쓰기, 복합어 처리, 숫자 처리, 외국어 처리 등의 작업을 수행한다.

항목	내용
띄어쓰기	• 신문이나 언론사의 자료를 통해 지면 편집과 일반적인 관례를 따라서 띄어쓰기 오류 발생 – 관례('에베레스트산' (x)–〉'에베레스트 산'(o)) // 지면편집('김씨'(x) –〉'김 씨'(o))
복합어 처리	• 원시 데이터에 띄어 쓰지 않은 형태로 나타난 복합어 중 띄어 쓸 수 있는 곳에 "–"를 삽입 – 복합어가 사전에 등재된 경우는 분리하지 않음 – 접두어 또는 접미어가 붙어 형성 – 사이 시옷이 들어 있음 – '1 음절짜리' 단어로 구성 및 띄어쓰기를 하는 경우 중의성이 증가, 한 단어로 처리(타수, 결승골, 골세례 등) – 수 부류사가 '개' +"1음절짜리 명사" – 회사명, 브랜드명과 같은 고유명사는 원문 유지

숫자 처리	• 수를 적을 때 '만' 단위로 띄어쓰기 적용('만, 억, 조'및 '경, 해, 자')
	– 한글로 표시된 단위 등은 변환 범위에 포함하지 않음: 5분 –〉 {{오}}//{{5}}분
	– 한자어, 영문, 기호로 표시된 단위 등은 따로 변환: 3m –〉 {{삼}}//{{3}}{{미터}}//{{m}}
	– 고유명사나 연어로 분류될 수 있는 것은 동일 변환 범위에 포함: 샤넬 No.5 –〉 샤넬 {{넘버 파이브}}//{{No.5}}
	– 숫자의 읽는 방법은 번호독식과 봉독식으로 표현: 20개–〉{{스무, 이십}}//{{20}}개

외국어 처리	• 한글로 변화하되 기준은 국립국어 연구원의 '외래어 한글 표기법' 으로 표기 – '한글 음사' 다음 괄호 안에 적은 영문자: 빅맥(big-mac), 팅(ting) – 한국어 다음 의미의 명확성을 위해 적은 영문자: 생명 윤리(bioethics) – 2단어 이상의 영문 고유 명사 영어 구문 이상의 단위 – 이메일 주소, ID, 웹 URL, 컴퓨터 경로명, 파일명 등 결합된 고유명사는 전체를 한 단위로 처리

| 원시데이터 선정 |

원시데이터는 텍스트 형식으로 피험자에게서 바로 수집을 하기 때문에, 별도의 변환 과정은 필요하지 않다. 원시데이터는 피험자로부터 직접 수집을 원칙으로 하며, 해당 데이터는 크라우드 소싱 플랫폼을 활용하여 동시에 수집을 수행한다.

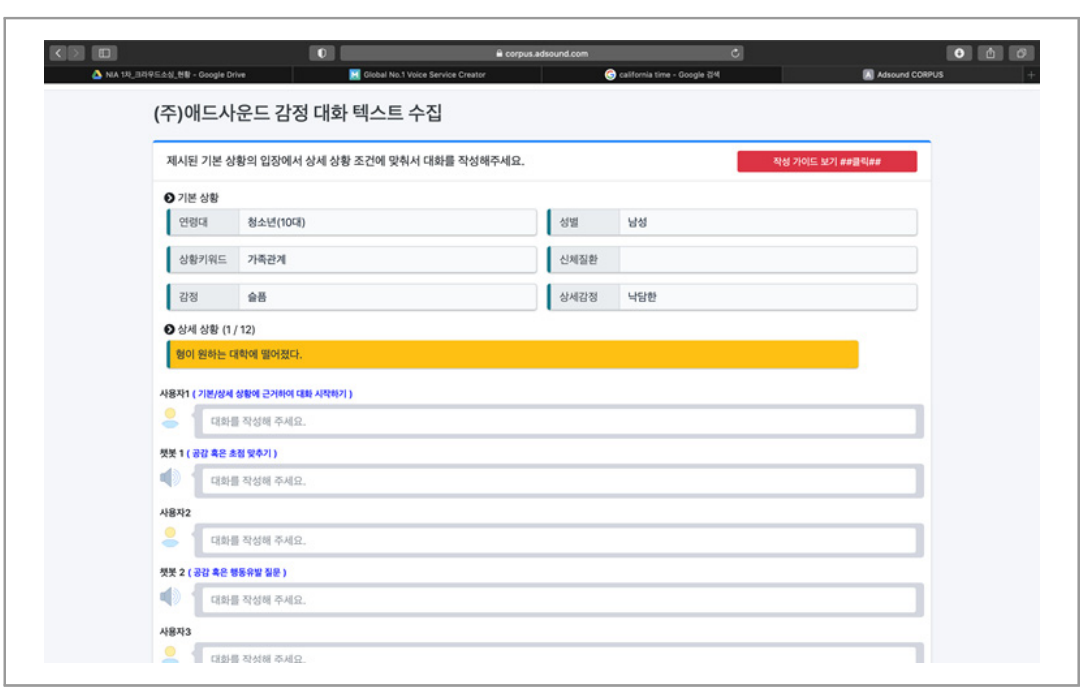
감정의 표현은 연령, 성별, 사회적 지위, 질병 이력, 학교 등 다양한 원인에 의해 달라질 수 있으므로, 다양한 상황을 제시하여 데이터 수집이 가능하도록 설계한다.

원시데이터 규모는 총 27만 문장의 텍스트 코퍼스이며, 상황 및 감정 상태를 태깅하여 의미 분류를 할 수 있도록 가공한다.

| 획득 · 정제 도구 |

원시데이터를 획득하는데 사용하는 도구는 직접 제작하여 사용한다. 기본 상황에서 연령, 성별, 상황 키워드, 신체 질환 등의 페르소나 정보가 주어지고, 이에 매핑되는 감정 상태와 상세 감정 상태를 활용하여 상세 상황에 대한 질문 및 응답을 작성한다.

이때 기본 상황이 수집자의 페르소나이며, 이 페르소나가 그대로 데이터 어노테이션의 정보가 된다. 또한 기본 감정 및 상세 감정 상태 정보 또한 어노테이션 정보로 곧바로 활용된다.



텍스트를 수집하면서 곧바로 어노테이션이 진행되는 형식이므로, 아래의 그림과 같이 상세 상황에 대해 순서대로 대화의 질의와 응답을 작성하는 절차로 데이터가 수집된다.

구어 데이터를 매우 효율적으로 수집할 수 있는 구조로 데이터 구축이 진행된다.

| 어노테이션/라벨링 : 어노테이션/라벨링 기준 |

사용자 프로파일과 감정 상태에 맞게 수집된 대화 데이터는 각각의 페르소나에 해당되는 ID와 수집 텍스트가 쌍으로 저장되어야 한다.

미리 사전에 사용자 페르소나와 감정 상태가 정의되어 있기 때문에, 이러한 페르소나와 감정 태그는 자동으로 부여되게 되므로 별도의 작업이 없이도 자연스러운 태깅이 수행된다.

이러한 태깅은 데이터 수집 툴을 통해 자동으로 수행되므로 별도의 태깅 작업이 없어도 무방하다.

Persona_Emotion ID는 아래와 같은 규칙으로 배정된다.

연령(A), 성별(G), 시스템응답(C)의 기본 ID가 배정된다.

구분	항목	상세	ID
페르소나	연령 (A)	청소년	A01
		청년	A02
		중년	A03
		노년	A04
	성별 (G)	남성	G01
		여성	G02
	시스템 응답 (C)	응답	C01

감정 상태(Emotion)의 경우 상황(Situation : S), 질병(Disease : D), 감정(Emotion : E)의 ID가 배정된다.

구분	항목	상세	ID
감정 상태	상황 (S)	상황 세부 항목	S01~S13
	질병 (D)	질병 세부 항목	D01~D02
	감정 (E)	감정 세부 항목	E10~E69

대화 턴에 대해서는 사람 대화(Human Speech : HS), 시스템 응답(System Speech), Turn (01~06) 값으로 ID가 지정된다.

구분	항목	상세	ID
대화 턴	사람 대화	Human Speech	HS
	시스템 응답	System Speech	SS
	대화 턴	Turn	01~06

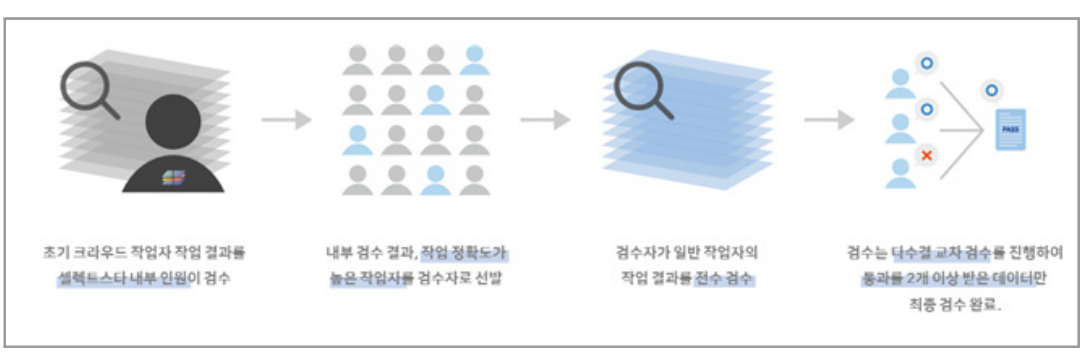
이에 따라 프로파일 ID는 아래와 같이 할당된다.

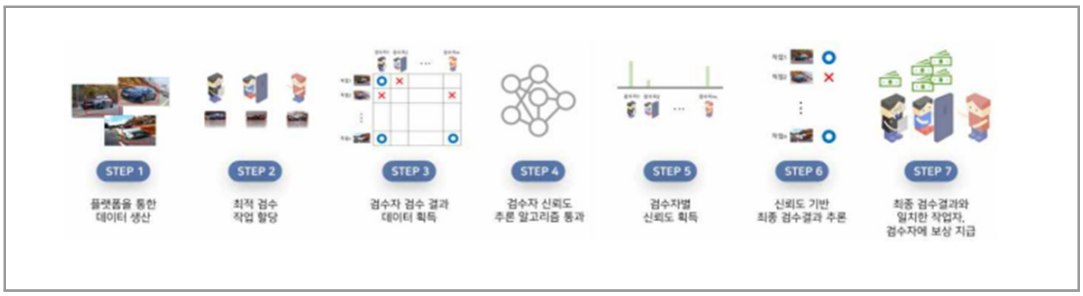
페르소나	프로파일	감정	연령	성별	상황	질병	감정
A01-G01-S01-D01-E01	Pro_01	A01-G01-S01-D01-E01	A01	G01	S01	D01	E01

●○ 검수: 검수 절차 및 기준

데이터의 검수는 크라우드 소싱 시스템을 통해 데이터를 수집한 뒤 크라우드 소싱 기업에서 1차 검수를 진행하며, 주관기관의 검수팀이 상용근로팀과 함께 2차 검수를 수행한다. 또한 최종 검수 결과는 전문가 초빙을 통해 최종 검수를 수행한다.

데이터의 규모가 방대하기 때문에, 크라우드 소싱 수집 단계의 1차 검수가 매우 중요하며, 이에 대한 절차는 다음과 같이 각 크라우드 소싱 기업의 자체 절차를 따른다.





데이터 검수는 완전성, 유효성, 정합성의 3가지 기준에 따라 수행된다. 다만, 본 프로젝트 산출물이 텍스트 코퍼스이다 보니 텍스트 처리 방법론을 따른다. ALBERT 알고리즘의 모델링 리소스는 문장 단위의 텍스트와 어노테이션 정보(감정)이다 보니 다른 빅데이터에 비해 검수의 과정이 직관적이며 다소 용이하다.

검수의 과정은 1차 검수, 2차 검수, 3차 검수의 3단계로 구성되어 있다.

검수	검수목적	검수 항목	검수 내용
1차 검수 (형태)	수집한 문장의 형태가 올바른지 확인	맞춤법	국립국어원 표준국어대사전 기준 검수
		특수 기호 표기	온점, 느낌표, 물음표를 제외한 특수 기호 제거
		어투	사람: 반말, 시스템: 해요체
2차 검수 (내용)	`Persona/Emotion에 맞는 문장인지 확인	연령	작성 내용이 제시한 연령대와 맞는지 확인
		성별	작성 내용이 제시한 성별과 맞는지 확인
		상황	작성 내용이 제시한 상황과 맞는지 확인
		감정	작성 내용이 제시한 감정과 맞는지 확인
		대화의 개연성	하나의 대화가 개연성 있게 진행되는지 확인
3차 검수 (모델링 적격성)	모델링 가능 여부 확인	개체명	개체명 태깅의 오류/적절성 확인
		중의적인 의미	한 문장 내에 있는 중의적인 표현의 의미 확인
		모호한 표현	한 문장 내에 있는 모호한 표현의 정확한 의미확인

1차 검수는 크라우드 소싱 기업에서 각각 원시 텍스트 데이터를 수집한 후 맞춤법 및 특수기호 표기 등에 대한 처리를 수행하며, 기본적인 말투 등 구어 대화 문장으로의 적절성을 검수한다.

2차 검수는 본 과제 수집 데이터의 특성 상 주어진 감정 및 상황 조건에 따른 내용의 검수를 위주로 진행된다. 연령, 성별, 상황, 감정, 대화의 개연성 등 주어진 페르소나 및 감정 상황에 맞는 내용의 대화가 이어지고 있는지를 검수한다.

3차 검수는 ALBERT 엔진 학습을 위한 대화 모델링 관점의 검수로, 문맥적 의미는 올바르지만 인공지능 시스템의 의미 값 해석에 영향을 줄 수 있는 사항들이 있는지 확인하는 과정이다. 중의적 표현이나 모호한 표현, 기타 모델링에 장애가 될 수 있는 사항들이 있는지 확인한다.

●○ 데이터 구축 담당자

수행기관(주관) : 송민규 총괄책임자
(이메일: minks@mediazen.co.kr)