SK네트웍스 **Family AI** 과정 **12**기

모델링 및 평가 수집된 데이터 및 전처리 문서

산출물 단계	모델링 및 평가
평가 산출물	수집된 데이터 및 전처리 문서
제출 일자	2025.07.20
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-6TEAM
작성 팀원	남의헌, 조성지

1.데이터 수집 개요

- 데이터 수집 기간:2025.07.07 ~ 2025.07.16
- 수집 방식: (크롤링 / 수기입력)

1.1. 수집 목적

- 사용자의 그림검사결과에 내포된 감정을 분석하고, 챗봇 페르소나인 5가지 유형(추진형 / 내면형 / 관계형 / 안정형 / 쾌락형)에 대해 분류하는 AI 모델 개발을 위한 학습 데이터셋을 구축하는 것을 목적으로 함

1.2 데이터 출처

데이터명	출처	수집 방식	형식
감성 대화 말뭉치	AI-HUB	파일 다운로드	JSON
멘탈케어 챗봇의 의도분석 학습데이터 구축을 위한 심리상담 대화의 감정표현 분류 연구	DBPIA	웹 사이트 검색 및 다운로드	PDF
성격유형별 선호도서 추천을 위한 서평 키워드 유효성 연구	RISS	웹 사이트 검색 및 다운로드	PDF

1.3 수집 대상 및 범위

- 대상: 일상적인 대화 및 전문적인 심리 상담 대화에 나타나는 감정 표현 텍스트 데이터
- 총 수집 건수: 669건 사용
- 개인정보 포함 여부: 예

- 민감 정보 여부 및 조치 사항: 민감 정보사항 익명화 처리(존재할경우)

1.4 법적·윤리적 고려 사항

- 개인정보보호법 준수 여부: O
- 데이터 수집 사전 동의 유무: O (서비스 약관에 명시됨)
- 수집 데이터의 활용범위 명시 여부: O (데이터 수집시 이미 동의된 데이터 수집)

2. 저장 방식

- 저장 경로: ./backend/data/personality_keywords_dataset_v2.json
- 저장 형식: JSON
- 일관성 확보 방법:
    - 데이터를 저장 전 Label, Keyword 등 필수 키(필드)가 모두 존재하는지 검사
    - 필수 키에 해당하는 값이 비어있지 않은지 확인하여 데이터 누락을 방지

2.1 중복 및 정합성 검증

- 중복 제거 기준:
    - 성격유형(Label)과 감정 키워드(Keyword)의 조합을 기준으로, 완전히 동일한 데이터는 중복으로 간주하고 제거
- 정합성 확보 방법:
    - 의미를 가지기 어려운 한 글자 단어나 특수기호만으로 이루어진 키워드는 필터링하여 제거

3. 데이터 전처리 절차

3.1 이상치 탐지 및 처리

- 이상치 기준:
    - 무관한 정보: PDF에서 추출 시 포함된 페이지 번호, 참고문헌, 표 제목 등 연구 내용과 직접 관련 없는 텍스트
    - 라벨링 불가능 데이터: `extract_keywords.py`로 추출했으나, 특정 성격 유형으로 분류하기 모호한 문장
    - 저장된 json파일의 text부분에 특수기호가 포함된 경우

- 처리 방법:
    - 무관한 정보는 일괄 제거
    - 에니어그램 9가지 유형 중 유형 번호가 명시된 블록만 필터링
    - 성격 유형 분류가 모호한 데이터는 모델 학습의 노이즈로 작용할 수 있어 학습 데이터에서 제외 처리함
    - Json파일 내에서 text항목의 특수기호 제거
- 품질 기여 설명:
    - 무관 정보 제거 및 라벨링 불가능 데이터 제외 처리로 학습 정확도 개선

3.2 결측치 처리

- 결측 필드:
    - 원천 데이터를 정제한 후 직접 구성한 데이터셋으로, 별도의 결측치는 존재하지 않음

3.3 데이터 변환 및 재현 가능성

- 텍스트 정제:
    - 특수문자 및 숫자 제거: 한글, 영문, 공백을 제외한 모든 불필요한 문자를 정규표현식으로 제거함
    - 공백 정규화: 문장 내 여러 개의 공백을 단일 공백으로 변환하여 일관성을 확보함
- 토큰화:
    - 별도의 형태소 분석기를 사용하지 않음. 대신, 정규표현식(re.findall)을 사용하여 텍스트에서 '...다' 형태로 끝나는 단어들을 키워드로 추출하는 방식을 사용함
- 변환 코드 예시:

```
keywords = set()
for i, line in enumerate(lines):
    # 감정명(예: 4.1.1기쁨(JOY))이 있는 줄 찾기
    if re.match(r"\d+\.\d+\.\d+[가-힣]+\([A-Z]+\)", line.strip()):
        # 다음 1~3줄에서 작은따옴표(‘ ’)로 감싸진 단어 추출
        for j in range(1, 4):
            if i + j < len(lines):
                found = re.findall(r"‘([^’]+)’", lines[i + j])
                for word in found:
                    keywords.add(word.strip())
```

3.4 전처리 절차 요약 (프로세스 플로우)

1. 데이터 수집
2. 이상치 및 결측치 확인
3. 중복 제거
4. 텍스트 정제
5. 토큰화 및 형용사 키워드 추출(정규표현식)
6. 성격 유형별 라벨 매핑
7. 라벨링 불가능한 데이터 제거
8. 저장 (json)

3.5 데이터 선정

- 선정 데이터 : '성격유형별 선호도서 추천을 위한서평 키워드 유효성 연구(2020, 차윤희)'
    - 해당 데이터는 성격 유형(Label)과 개인이 사용하는 감정 키워드(Keyword) 사이의 관계를 명확하게 정의하고 있음. 이는 그림 검사 결과로 나타난 표현(Text)을 분석하여 사용자의 성격 유형을 분류하려는 본 프로젝트의 목표와 직접적으로 일치함.

- 제외 데이터 : '감성 대화 말뭉치', '멘탈케어 챗봇의 의도분석 학습데이터 구축을 위한 심리상담 대화의 감정표현 분류 연구(2023, 박온유▪남지순)'
    - 두 데이터는 일상적인 대화나 전문 심리상담과 같은 구어체 텍스트에 초점을 맞추고 있음. 이는 문어체 기반의 서술적 표현이 주로 나타나는 그림 검사 결과 텍스트의 특성과 다소 차이가 있어 사용하기에 부적합함

3.5.1 데이터 불균형 해소 및 증강

- 선정한 '성격유형별 선호도서 추천을 위한 서평 키워드 유효성 연구' 데이터만을 사용할 경우, 특정 성격 유형(Label)에 대한 감정 키워드(Keyword)가 부족하거나 라벨 간 데이터 분포가 불균일할 수 있는 한계가 있음
- 이러한 데이터 불균형 문제는 모델이 특정 유형에 과적합이 되는 원인이 될 수 있음. 이를 해결하고 모델의 일반화 성능을 높이기 위해, 각 성격 유형(Label)별 데이터 비율을 고려하여 데이터를 669개로 증강함. 이를 통해 모델이 모든 성격 유형을 편향 없이 균등하게 학습하도록 데이터셋을 구성함.

3.6 데이터셋 분리

- 학습/테스트 비율:

구분	건수	비율

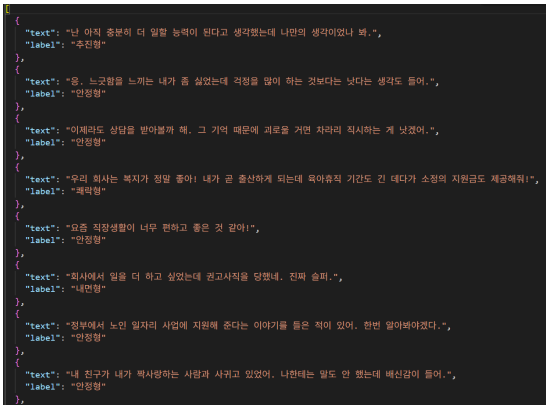
Train	468	70%
Test	201	30%

- 분리 기준: 무작위(Random State=42), 클래스 비율 균형 유지 (Stratified Split)

4. 부록

- 전처리 전/후 비교 예시

  전 / 후




```
[
  {
    "text": "난 아직 충분히 더 일할 능력이 된다고 생각했는데 나만의 생각이었나 봐.",
    "label": "주진형"
  },
  {
    "text": "응. 느긋함을 느끼는 내가 좀 싫었는데 걱정을 많이 하는 것보다는 낫다는 생각도 들어.",
    "label": "안정형"
  },
  {
    "text": "이제라도 성담을 받아볼까 해. 그 기억 때문에 괴로울 거면 차라리 직시하는 게 낫겠어.",
    "label": "안정형"
  },
  {
    "text": "우리 회사는 복지가 정말 좋아! 내가 곧 출산하게 되는데 육아휴직 기간도 긴 데다가 소정의 지원금도 제공해줘!",
    "label": "패락형"
  },
  {
    "text": "요즘 직장생활이 너무 편하고 좋은 것 같아!",
    "label": "안정형"
  },
  {
    "text": "회사에서 일을 더 하고 싶었는데 권고사직을 당했네. 진짜 슬퍼.",
    "label": "내면형"
  },
  {
    "text": "정부에서 노인 일자리 사업에 지원해 준다는 이야기를 들은 적이 있어. 한번 알아봐야겠다.",
    "label": "안정형"
  },
  {
    "text": "내 친구가 내가 싹사랑하는 사람과 사귀고 있었어. 나한테는 말도 안 했는데 배신감이 들어.",
    "label": "안정형"
  },
```


- 전처리 코드 전체 링크 또는 주요 코드 첨부

```
        re.DOTALL
    )
    match = pattern.search(full_text)
    return match.group(1).strip() if match else ""

# PDF 경로
pdf_path = r"C:\SIMPLESTOCK\final-team6-main\word\성격유형별 선호도서 추천을 위한 서평 키워드 유효성 연구.pdf"

# 표 부분 텍스트 추출
section_text = extract_section_from_pdf(
    pdf_path,
    "<표 8> 성격유형별 형용사 키워드",
    "3.2 도서의 선정"
)

# 라벨별 키워드 저장용
label_keywords = {}

# 한 줄씩 처리
lines = section_text.split("\n")
current_label = None

for line in lines:
    line = line.strip()
    # 1~9 숫자형 라벨 시작되는 줄인지 확인
    match = re.match(r"^(\d)번", line)
    if match:
        type_num = match.group(1)
        current_label = type_to_label.get(type_num)
        if current_label and current_label not in label_keywords:
            label_keywords[current_label] = []
        line = line[len(match.group(0)):]  # 숫자부분 제거 후 키워드 처리

    if current_label:
        keywords = re.findall(r"[가-힣]{2,}하다", line)
        label_keywords[current_label].extend(keywords)
```

- 참조 문헌 및 API 문서 (공공 데이터, 크롤링 페이지 등)

  - 박온유, 남지순, 멘탈케어 챗봇의 의도분석 학습데이터 구축을 위한 심리상담 대화의 감정표현 분류 연구, (2023), 145~160p
  - 차연희, 조현양, 성격유형별 선호도서 추천을 위한 서평 키워드 유효성 연구, (2020), 44~48p