에브리타임 댓글봇 개발

질문글을 자동 인식하고 답변을 생성하는 **댓글봇**

2025114794 이주형

Table of contents

Introduction

- 학교 커뮤니티 특성상 학교에 대한 **질문글**이 많이 올라옴
- 특히 학기초 같은 경우 새내기들의 **반복적인** 질문글이 많음
- 이 질문글에 대해 **자동으로** 질문에 대한 댓글을 달아주는 **AI**가 있다면 유용할 것
- 따라서 **질문글을 자동으로 분류하고 그에 맞는 답변을 생성하는 댓글봇**
을 만들어보고자 함

Project Plan

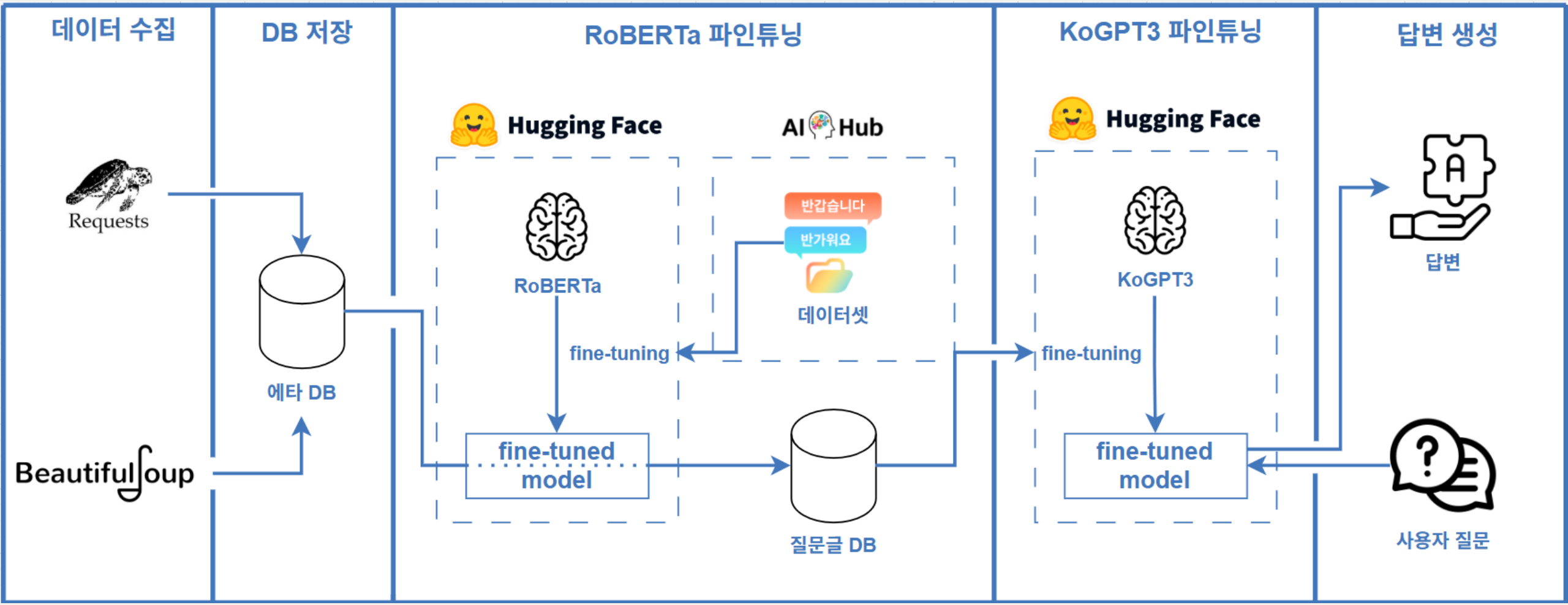
1. 에브리타임 게시물과 댓글 **크롤링** 및 **데이터베이스 저장**

2. 질문글만 분류할 수 있는 **모델 구축** (RoBERTa Fine-tuning)

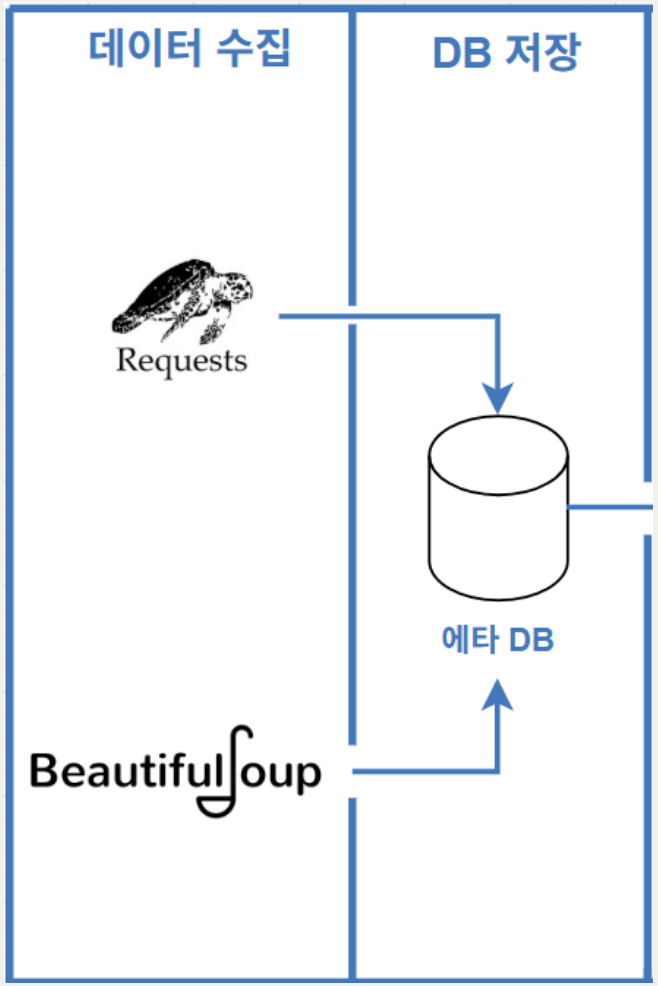
3. 댓글 생성 **모델 구축** (KoGPT3 Fine-tuning)

4. 한계점 및 보완계획

Project Architecture



1. 에브리타임 게시물 크롤링



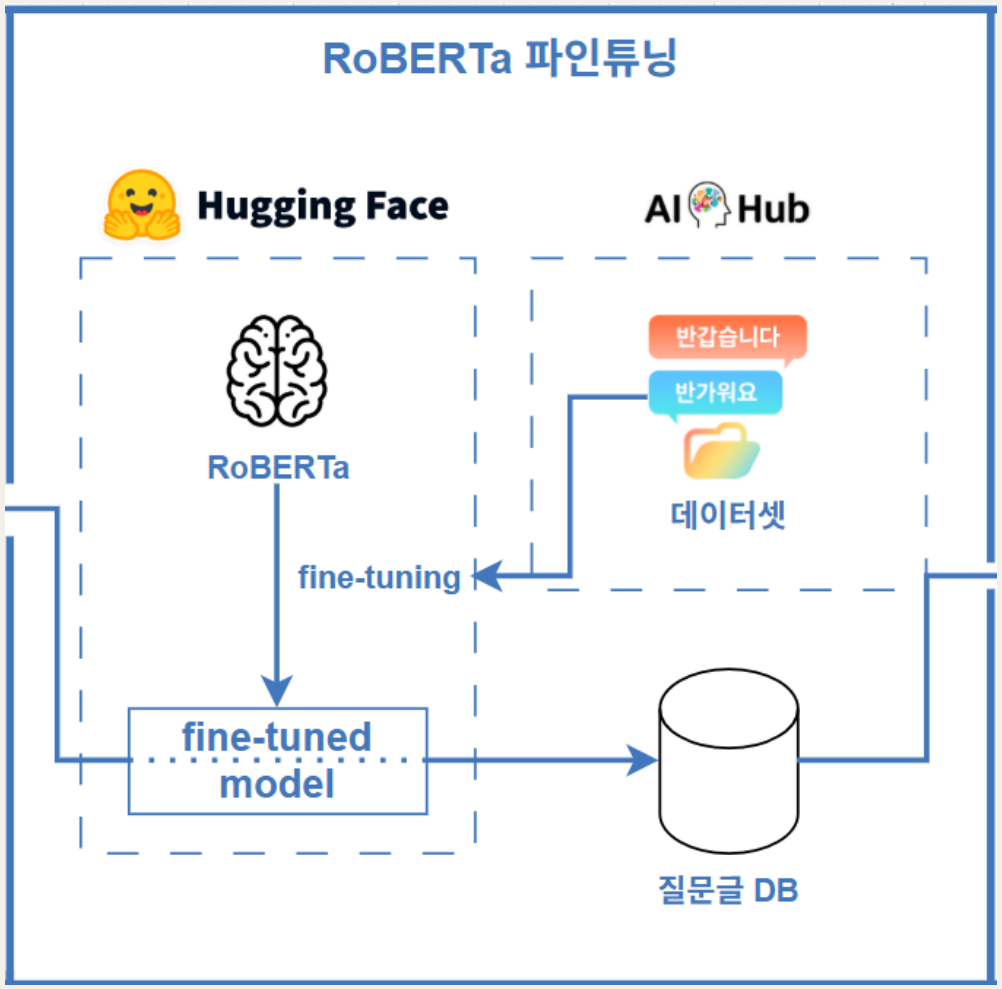
에브리타임 자유게시판과 새내기 게시판에서 게시글과 댓글을 수집 및 전처리

DB 테이블 예시

article
게시물 번호
게시글 내용

comment
댓글 번호
게시물 번호
Parent 댓글 번호
댓글 내용

2. 질문글 분류 모델 구축

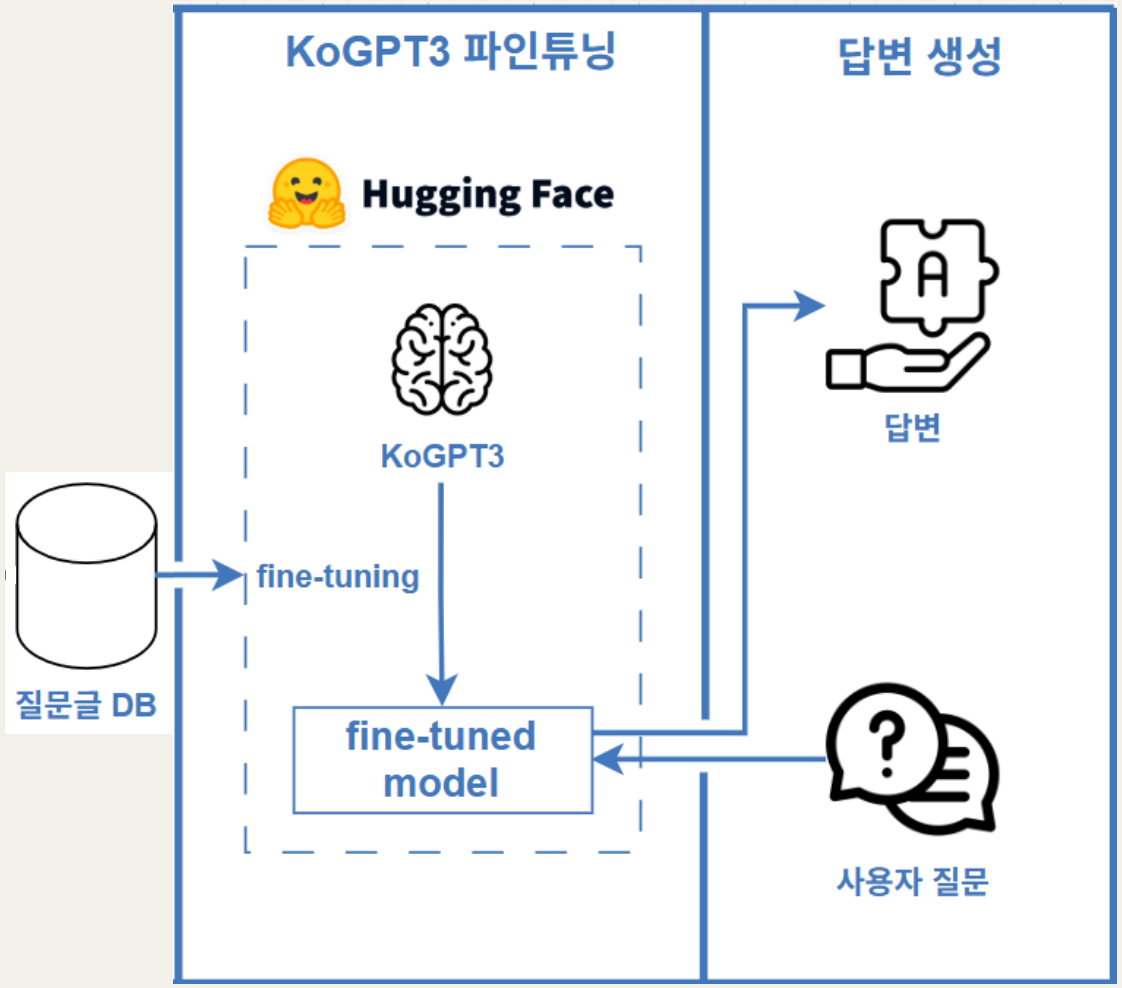


모델 : RoBERTa 사용해 파인튜닝
https://huggingface.co/klue/roberta-base

데이터 : AI Hub의 "주제별 텍스트 일상 대화 데이터" 와 "용도별 목적대화 데이터" 사용

```
{
  "id": 9,
  "text": "A.휴대폰 번호는 혹시 죄송하지만 어떻게 되세요",
  "norm_text": "A.휴대폰 번호는 혹시 죄송하지만 어떻게 되세요",
  "speaker": {
    "id": "A",
    "sex": "여성",
    "age": "1그룹(10대~20대)"
  },
  "speechAct": "질문하기",
  "morpheme": "A/SL+./SY+휴대폰/NNG+번호/NNG+는/JX+혹시/MAG+
},
```

3. 댓글 생성 모델 구축



모델 : **KoGPT3** 사용해 파인튜닝
https://huggingface.co/kakaobrain/kogpt

데이터 : RoBERTa를 이용해 뽑아낸 질문글 중 ChatGPT에 질의 후 가장 적절한 답변만 골라 데이터셋 생성

4. 한계점 및 보완계획

한계점

1. **데이터 품질 문제** - 커뮤니티 글 특성상 줄임말, 비표준어 많아 노이즈 데이터 다수 존재
2. **질문글 분류 정확도 한계** - AI Hub 데이터는 일상 대화 중심 -> 커뮤니티 질문글에 부적합 가능성
3. **생성된 댓글의 신뢰성 부족** - koGPT3 환각으로 사실과 다른 문장 생성 위험

4. 한계점 및 보완계획

보완계획

1. **비표준 텍스트 교정** - py-hanspell 라이브러리를 사용해 맞춤법 및 띄어쓰기 자동 교정

2. **질문글 분류 성능 개선**
    1) AI hub 데이터 + 커뮤니티 질문글 데이터 혼합 파인튜닝
    2) 상용 LLM으로 질문글 자동 추출
    3) 수작업 라벨링으로 데이터 품질 향상

3. **댓글 생성 신뢰성 향상 -** RAG기반 학교 공식 정보 참조 -> 환각 현상 감소, 사실 기반 답변 생성

Q&A