**프로젝트 중간 보고서**

| **프로젝트명** | K-Chatbot | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **이행기간** | 2023.05.01 ~ 2023.05.26 | | | | | |
| **멘토** | 고대영 | | **팀원** | | 석은수, 손지현, 이소정 | |
| **1. 프로젝트 설계** | 1. 학습데이터    1. 웰니스 대화 스크립트 데이터셋 (AI hub(<https://aihub.or.kr/>)    2. 송영숙님의 챗봇 데이터 (https://github.com/songys/Chatbot\_data) 2. 이용모델    1. KoGPT2(https://github.com/SKT-AI/KoGPT2) | | | | | |
| **2. 프로젝트 일정 진행 상태** | 1. 프로젝트 일정   * 일정 및 진행도     **진행사항**  5.10 - 데이터 변경  5.16 - 데이터 준비 완료  5.19 - 데이터 변경 및 전처리 완료  5.24-25 - 웹 구현 예상안 구성완료  5.26 - 중간 보고 발표 준비 완료(보고서, ppt) | | | | | |
| **3. 멘토링 결과** | 5.10 - KoAlpaca 이용 권장  5.16 - KoGPT2 이용 권장, 데이터셋 확정  5.18 - 5.26까지 모델 구현 이행 이후 웹서비스 구현 | | | | | |
| **4. 프로젝트 진행의 문제점 및 해결방안** | 1. 모델    1. 기존 : KoAlpaca    2. 문제점       1. KoAlpaca를 이용하려 했으나 자원 문제로 이용 불가능 (KoAlpaca모델이 거대 모델이어서 A100이상의 자원이 필요함)       2. 심리케어 챗봇은 감정에 맞는 답변이 요구됨 문장생성 모델 학습 시 감정 라벨 학습 필요    3. 해결 방법       1. KoGPT2 이용       2. 답변 생성 시 KoBert 모델로 입력에 대한 감정 분석 2. 데이터    1. 기존 : 주제별 텍스트 일상 대화 데이터, 웰니스 대화 스크립트 데이터    2. 문제점       1. 주제별 텍스트 일상 대화 데이터를 Q&A형식으로 전처리 불가능 (카카오톡, 페이스북 등 일상 대화 데이터는 Q&A형식으로 정제 불가, 세명이상 대화의 경우 질문이 여러번 중복되거나 답변이 없는 경우가 많아 전처리 하여도 데이터의 개수가 매우 적어 이용불가)    3. 해결 방법       1. 데이터셋 변경 => 웰니스 대화 스크립트 데이터셋과 송영숙님의 챗봇 데이터를 합하여 사용하기로 결정 | | | | | |
| **5. 향후 일정** | 1. 플랫폼 업로드  * T3Q.ai 서버에 데이터를 넣고 학습시킨 모델 파일 올리기  1. 웹서비스 구현  * 사용자와 챗봇 단둘이 대화를 주고받는 형식의 채팅방 구현 | | | | | |
| **6. 역할 변경** | 석은수 : 모델테스트 진행 및 평가, 문서작성, 웹서비스 구현  손지현 : 데이터 전처리, 문장생성 모델링  이소정 : 데이터 수집, PPT 제작, 감정분석 모델링, 웹서비스 구현 | | | | | |
| **7. 기타 문제** | 1. 자원 부족  * 코랩 프로를 사용해 데이터를 학습시켜도 자원이 부족하고 서버나 GPU등 자원을 지원받을 수 없어 모델 학습에 차질이 생김 * 강의장 인터넷 연결이 끊길 때가 종종 있어 학습이 중단되는 경우 발생 | | | | | |