|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **과제 보고서**  **(2024.07.11)** | | 결  재 | 작 성 자 | 선 임 | | 부 장 | | 위 원 |
|  |  | |  | |  |
| 작 성 자 | 서원형 | | | | 작 성 일 | | 24.07.11 | |

|  |  |  |  |
| --- | --- | --- | --- |
| 과제명 | AI 전자약 활용 서비스 자료조사 | 진행도(%) | 90 |
| 결론 | - 9가지의 정실질환 관련 데이터셋을 찾았고 활용 방안을 고안, 실현가능성 판단.  [남은 문제점]  데이터셋이 누구에게나 개방되어 있는 것이 아니라 활용신청이 필수기 때문에 활용하게 된다면 그에 따른 자료 준비 및 구체적인 계획이 요구됨 | | |
| 진행 과정 | **[개요]**  전자약과 관련된 정부과제가 많이 생기고 있고, 인증된 데이터를 통해 비싼 양질의 서비스를 제공할 수 있을 것이나 구체화된 내역이 없어, 가장 중요한 데이터셋부터 조사하였음.  **[조사 방법]**  정실질환과 관련하여 우선 국내의 공공데이터 포털, AI-HUB를 조사하였고 총 5가지의 데이터를 가져올 수 있었음.    1. 노인 정신건강 영상 데이터(음성 데이터, 오각형 이미지, 얼굴표현 영상)    2 정신건강진단 및 예측을 위한 멀티모달 데이터(음성 데이터)    3 고령인구 우울증 데이터(음성 데이터, 텍스트)    4. 파킨슨병 및 관련 질환 진단 음성데이터(음성데이터)    5. 소아청소년 정실질환 진단 안저 데이터    추가적으로 Kaggle을 통해 총 4개의 데이터를 추가적으로 확인하였다.  1. 알츠하이머병을 감지하는 필기 데이터  2. 알츠하이머 MRI 전처리 데이터 세트  3. Synthetic Therapy Conversations  4.정신 건강 대화 데이터  **[활용 방안]**   |  |  |  |  | | --- | --- | --- | --- | | 데이터 | 활용방안 | 개발 난이도 | 서비스 | | 노인 정신건강 영상 데이터 | 음성, 발화, 행동적 특성 등을 AI로 분석하여 치매 조기진단 모델 개발 가능 | 상 | 웹, 앱 서비스 | | 정신건강진단 및 예측을 위한 멀티모달 데이터 | 음성인식을 통한 우울증 진단 보조 시스템, 질문을 주고 대답을 통해 분석 PHQ-9 | 상 | 웹, 앱 서비스 | | 고령인구 우울증 데이터 | 60세 이상 노인의 맞춤형 우울, 불면, 인지기능에 대한 건강 상태 정보 서비스 | 중 | 웹, 앱 서비스 | | 파킨슨병 및 관련 질환 진단 음성데이터 | 파킨슨 병의 조기 발견 가능 | 중 | 웹, 앱 서비스 | | 소아청소년 정실질환 진단 안저 데이터 | ADHD, ASD 구현 가능 | 중 | 병원 시스템 | | 알츠하이머병을 감지하는 필기 데이터 | 알츠하이머 병 조기 발견 | 중 | 웹, 앱 서비스 | | 알츠하이머 MRI 전처리 데이터 세트 | 알츠하이머 병 조기 발견 | 상 | 병원 시스템 | | Synthetic Therapy Conversations | LLM 채팅으로 치료 | 상 | 웹, 앱 서비스 | | 정신 건강 대화 데이터 | LLM 채팅으로 치료 | 상 | 웹, 앱 서비스 |   위 9개 데이터 세트를 활용하는 방안은 위 표와 같다.  조금 더 응용된 활용방안으로는 모델을 2가지를 혼합하여 쓰면 난이도는 올라가지만 더 좋은 서비스로 발전 가능하다고 생각한다. (예시 : 노인의 정신 건강 영상 데이터를 활용하여 진단하고 LLM을 통하여 치료까지 연계해주는 서비스)  **[문제점]**  여러 요소를 고려해보았을 때, 2가지의 문제점이 크게 대두된다.  1. 데이터 활용 신청 및 선정의 복잡성  2. 모델 개발 환경의 부족  처음의 문제로는 AI-HUB의 경우 헬스케어 데이터는 활용신청 -> 심사 -> 활용 승인의 프로세스로 데이터를 사용할 수 있다. 이때, 필요한 항목으로는 사용목적, 사용목적 분류, 기관생명윤리위원회 심의 결과 통지서, 기관생명윤리위원회 승인된 연구계획서, 보안서약서가 필요하다. 이에 기간이 소요될 것이고 학습의 오랜 시간이 필요하다. 물론 이는 Kaggle의 데이터는 바로 다운로드가 가능하기 때문에 해결될 수 있는 에러이나 AI-HUB는 국내 병원, 연구소등에서 직접 수집한 양질의 데이터를 국민에게만 무료로 제공하기 때문에 데이터의 품질은 Kaggle이 떨어진다. 다만, 다양성 측면에서는 Kaggle의 데이터가 앞서는 모습을 보인다.  두번째 문제로는 컴퓨팅 할 수 있는 GPU 자원의 부재이다. 텍스트 데이터는 가볍게 학습시킬 수 있지만, 이미지, 영상 데이터는 많은 용량으로 학습시간이 굉장히 오래걸린다. 이때, AI-HUB에서 컴퓨팅 자원을 30일간 무료로 대여해주지만, 신청일 기준 14일 이후에 이용가능하다는 점과 정확한 컴퓨팅 자원이 어느정도인지 공지되지 않았다. 이에 2가지 해결책이 있다. 첫번째로는 AWS 환경의 구축이다. 이때 SSD보다 빠른 속도를 지원하는 S3와 Sagemaker를 통한 빠른 모델 학습이 가능하다. 다만, 이때 AWS의 서버 비용이 일정하게 부과된다는 단점이 존재한다. 물론 Colab도 머신러닝에 이용이 가능하지만 Colab은 실현 가능성을 판단하기 위해 라이트한 모델을 사전학습 시키는 용도로 사용하고 실제로는 더 큰 GPU 리소스로 학습시켜야 한다 생각한다. 두번째로는 로컬 GPU 서버의 구축이다. 비용은 일시에 지불하지만 지속적으로 나가는 비용을 줄일 수 있다. 하지만, 최신 GPU가 계속해서 업데이트 되고 SSD 같은 추가 장비의 구축 필요성도 있기 때문에 첫번째 방법이 유용하다고 판단한다.  **[결론]**  위와 같은 많은 어려움이 있음에도 불구하고 AI를 활용한 양질의 서비스를 개발한다는 것은 미래 먹거리를 빠르게 선점할 수 있다고 생각한다. 다만, 처음부터 올바른 계획을 세우고 차근 차근 협의된 프로세스대로 진행해야 할 것이다. | | |