Data INt final

- The goal of this team project is to practice using problem-based methodologies to find data that match the selected topic and to get key messages and interpretations from the data

문제 기반 방법론의 관점에서 우리의 프로젝트는 적합할까?

1. 문제 정의

- **핵심 문제**: 기존 감정분석 방법론(예: 사전 기반, 머신러닝 기반)과 GPT 같은 생성형 AI를 활용한 감정분석 방법론이 성능과 결과 해석에서 어떤 차이를 보이는가?
- **세부 질문**:
  - Bible 데이터를 대상으로 GPT를 이용한 감정분석이 기존 감정분석과 비교해 어떤 이점을 제공하는가?
  - 두 접근 방식의 결과를 어떻게 해석하고 비교할 것인가?

2. 데이터 수집 및 전처리

- **Bible 데이터**:
  - 이미 구조화된 Bible 텍스트 데이터(예: 성경 구절별 데이터)를 사용.
  - 각 구절을 분석할 수 있도록 전처리(예: 정제, 구절별 ID 추가, 필요한 메타데이터 생성).
- **기존 감정분석 방법론에 필요한 데이터**:
  - 사전 기반 감정분석(감정 사전 생성 및 매핑) 또는 전통적인 머신러닝 모델에 적합한 학습 데이터를 준비.
  - 예: 긍정/부정/중립 라벨링된 텍스트 데이터.

3. 분석 방법론 비교

- **GPT 기반 감정분석**:
  - GPT를 활용해 각 구절에 대한 감정 분석을 수행.
  - 예: ChatGPT 또는 GPT-4 API를 활용해 구절별 감정(긍정, 부정, 중립)을 분류하고, 세부적인 감정 범주(예: 희망, 위로, 슬픔 등)를 추출.
  - 장점: 더 세밀하고 맥락에 맞는 분석 가능.
- **기존 감정분석 방법론**:
  - 사전 기반 분석: 단어의 감정 점수를 더하거나, 텍스트를 분류하는 방식.
  - 머신러닝 모델: 전통적인 감정분류 모델(SVM, Random Forest, Naive Bayes 등) 사용.(현재는 X)

4. 결과 비교 및 해석

- **비교 포인트**:
  - **정확성**: 어떤 방법론이 더 높은 정확도를 보이는가?
  - **맥락 이해 능력**: 특정한 구절이 표현하는 감정을 정확히 파악했는가?
  - **확장성**: 더 다양한 감정 카테고리를 적용할 수 있는가?
  - **실행 가능성**: 어떤 방법론이 더 효율적이고 사용이 용이한가?
- **비교 결과 해석**:
  - 데이터를 시각화하여 감정 분석 결과를 비교.
  - 예: 긍정/부정/중립 분포 그래프, 주요 감정 범주의 차트.

5. 결론 및 적용

- **결론 도출**:
  - Bible 데이터 감정분석에 있어 GPT의 강점과 약점은 무엇인가?
  - 기존 방법론이 특정 상황에서 더 적합한지, 아니면 GPT가 더 우수한지 평가.
- **적용 가능성**:
  - GPT 기반 분석이 Bible 데이터를 활용한 새로운 서비스(예: 신앙 상담, 감정 기반 추천 시스템)로 확장될 수 있는지 검토. → 성경 구절 추천 서비스까지로 연결

- **Introduction** (15pts)
  - Background (5 pts) → Is the motivation of this project sufficiently explained?
  - Problem statement (5 pts) → Is the project problem specific and clear enough?
  - Purpose & goal (5 pts) → Are the purpose and goal clearly stated?
- **Data preparation** (30 pts)
  - Data searching (10 pts) → Are they found the appropriate data set for the project?
  - Data description (10 pts) → Is the data description sufficient to understand the data?
  - Data preprocessing (10 pts) → Are they conducted all necessary data preprocessing task?
- **Data analysis** (30 pts)
  - Analytic method (10 pts) → Do they select a proper method? Do they fully understand data analysis method?
  - Visualization (10 pts) → Are the visualized works easy to understand? Do they help you understand the project?
  - Interpretation (10 pts) → Do they provide correct interpretation?
- **Conclusion** (10 pts)
  - Concluding remark → Do you think the project give you the new intuition?

1. Introduction

1. **Background (5점)**
   - 감정분석은 텍스트 데이터를 이해하는 데 중요한 도구로, 특히 성경과 같은 종교적/역사적 문헌의 감정 패턴 분석은 신학적, 심리학적 연구에 유용함. 최근 GPT와 같은 생성형 AI는 기존 감정분석 방법론보다 더 높은 맥락 이해 능력을 제공하며, 이는 성경 구절 분석에 혁신적 도구가 될 수 있음.
2. **Problem statement (5점)**
   - 기존 감정분석 방법론(사전 기반, 머신러닝 기반)은 성경 텍스트와 같은 문맥적으로 풍부한 문서를 충분히 이해하지 못하는 경우가 많기 때문에 GPT를 활용하면 이런 한계를 극복할 수 있을지에 대한 검증이 필요함.
3. **Purpose & goal (5점)**
   - 프로젝트의 목표는 GPT와 기존 감정분석 방법론을 비교하여, 성경 구절 감정분석에서 두 접근 방식의 장단점을 평가하고, 더 나은 분석 프레임워크를 제안하는 것.

2. Data Preparation

1. **Data searching (10점)**

- 프로젝트는 66권의 성경 텍스트 데이터를 활용하며, 이는 구절별로 구조화되어 있음. 데이터는 온라인에서 공개된 성경 데이터 세트를 통해 수집함 → NIV BIBLE DATA(ENG)

2. **Data description (10점)**
    - 데이터는 약 31,000개의 구절로 구성되어 있으며, 각 구절은 책, 장, 절 정보와 함께 텍스트 내용으로 이루어져 있음. 프로젝트에서는 각 구절의 감정을 '긍정', '부정', '중립'으로 분류하며 추가로 세부 감정 카테고리(예: 희망, 위로, 슬픔 등)를 도출함
3. **Data preprocessing (10점)**
    - 텍스트 데이터를 모델에 입력하기 전 불필요한 특수문자를 제거하고, 각 구절에 ID를 부여하여 분석 및 비교 작업에 용이하도록 전처리함. 추가로, 감정분석 결과를 통합하기 위해 데이터프레임으로 완성함.

3. Data Analysis

1. **Analytic method (10점)**
    - **설명**:
        - GPT 기반 분석에서는 GPT API를 활용해 각 구절의 감정을 자동으로 분류하고 추가적인 분석을 함
        - 기존 감정분석 방법론에서는 사전 기반 방법을 사용하여 동일한 데이터를 분석함
        - 두 방법론의 결과를 비교해 성경의 문맥을 무엇이 더 잘 반영하는지 평가함.
2. **Visualization (10점)**
    - **설명**:
        - 결과를 각 구절별 긍정/부정의 정도로 시각화하여, 긍정/부정/중립 감정 분포와 주요 감정을 확인함
        - 각 방법론의 결과를 비교하기 위해 감정 카테고리별 정확도 및 감정 분포 그래프를 제공함
3. **Interpretation (10점)**
    - **설명**:
        - GPT는 맥락적으로 풍부한 구절에서 더 높은 정확도를 보였으며, 단순히 단어의 뜻을 분석하는 것이 아닌 상황 맥락적 정보를 풍부하게 이용하였음

    - GPT는 희망, 위로, 슬픔 등 세부 감정 분류에서도 더 나은 결과를 제공함

4. Conclusion

1. **Concluding remark (10점)**

- **설명**: 이 프로젝트는 성경 데이터 감정분석에서 GPT가 기존 방법론에 비해 더 맥락적이고 세부적인 분석을 제공한다는 점을 확인하였다. 이는 성경뿐만 아니라, 다른 문맥적으로 풍부한 텍스트 데이터를 분석하는 데에도 유용할 수 있다. 기존의 감정분석은 필요한 데이터만을 가지고 사전을 만들거나 혹은 머신러닝을 활용해야 하므로 비정형 데이터를 분석하는 것은 어려우나, GPT를 통한 분석은 추가적인 학습이 거의 필요하지 않기 때문에 더욱 유용하다.