SK네트웍스 Family AI과정 5기

모델링 및 평가 수집된 데이터 및 전처리 문서

☐ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 수집된 데이터 및 전처리 문서
- 제출 일자 : 2025.02.19
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN05-final-3Team
- 작성 팀원 : 최영민

1. 개요

1.1 데이터 설명

본 문서는 **FinPilot Project**에서 **LLM Application**의 '요약 및 확장 프로세스' 간 활용할 **sLLM(smaller Large Language Model)** 모델의 파인튜닝을 위한 데이터 수집 및 전처리 과정을 정리한다. 데이터의 출처, 수집 방법, 전처리 단계, 품질 관리 및 저장 방식에 대한 내용을 포함한다.

본 문서는 **FinPilot Project**에서 **LLM Application**의 '요약 및 확장 프로세스'에서 효과적으로 활용될 **sLLM(smaller Large Language Model)** 모델의 파인튜닝을 위해 필요한 데이터 수집 및 전처리 과정을 상세하게 정리한 문서이다.

이 문서에서는 데이터의 출처, 수집 방법, 전처리 단계, 품질 관리 절차, 그리고 데이터 저장 방식 등 다양한 측면을 다루며, 모델의 학습 데이터가 신뢰성과 정확성을 갖추도록 하는 데 초점을 맞춘다. 특히, 금융 및 증권 도메인에 특화된 모델을 구축하는 과정에서, 기업 분석, 시장 동향과 같은 전문적인 정보가 포함된 데이터를 수집하고 가공하는 방법론을 다룬다.

1.2 데이터 수집 목적

FinPilot은 비즈니스 과정에서 활용할 수 있는 문서 작성 어시스턴트로, 금융 및 증권 도메인에서의 정교한 언어 이해와 문서 처리 능력이 요구된다. 이에 따라, 언어 모델델 기반의 애플리케이션이 금융 및 증권 분야의 특수 용어와 표현을 정확하게 인식하고, 이를 문서 요약 및 확장 프로세스에서 일관되게 유지할 수 있도록 설계하는 것이 필수적이다. 특히, 기업 분석과 관련된 내용을 처리할 때, 원문에서 전달하고자 하는 핵심 정보를

정확하게 추출하고 변형 없이 유지하는 것이 중요하다. 이러한 요구를 충족하기 위해, 금융 및 증권 도메인에 특화된 데이터셋을 신중하게 선정하고 수집하는 방안을 계획하였다. 이를 통해, 기업 재무제표 분석, 시장 동향 예측, 투자 전략 수립과 같은 전문적인 내용을 효과적으로 이해하고 반영할 수 있도록 모델을 최적화할 예정이다.

2. 데이터 수집

2.1 데이터 출처

- ‘JobKorea’ 기업 분석 보고서 크롤링 데이터
    - 55건의 기업 분석 보고서에서 발췌한 본문 224건

2.2 데이터 유형

- 텍스트 데이터

3. 데이터 전처리

3.1 전처리 개요

- 데이터 정제
- 요약문 / 확장문 생성
- EDA
- 파인튜닝 형식변환
- 학습, 평가 데이터 분할

3.2 데이터 정제

- 불필요한 공백 및 특수 문자(‘\n’, ‘|’, 등) 제거
- 정제 결과 csv 파일 저장, 인코딩 변환 (UTF-8 통일)
- 데이터 정제 결과

3.3 요약문 / 확장문 생성

- 개요
  - OpenAI API를 활용하여 원본 텍스트에 대한 요약문 및 확장문 데이터셋을 생성하였습니다. 이 데이터셋은 주어진 원본 텍스트를 보다 간결하게 요약하거나, 상세하게 확장하는 것을 목적으로 합니다.
- 데이터셋 생성 과정
  - LLM 설정
    - LLM으로 OpenAI의 `gpt-4o` 모델을 선택하고, API를 통해 호출할 수 있도록 설정합니다. `temperature=0`을 사용하여 결과의 일관성을 유지합니다.

```
llm = ChatOpenAI(
    model="gpt-4o",
    api_key=OPENAI_API_KEY,
    temperature=0
)
```

  - 요약문 생성
    - 요약문을 생성하기 위해, LLM에게 요약을 요청하는 메시지를 구성합니다. `SystemMessage`를 사용하여 모델에게 역할을

지정하고, **HumanMessage**로 원본 텍스트를 전달합니다.

```python
def summarize_text(llm, text):
    messages = [
        SystemMessage(content="너는 전문적인 요약가야. 주어진 텍스트를 간결하게 요약해 줘.
        HumanMessage(content=text)
    ]
    response = llm(messages)
    return response.content
```

- **SystemMessage**: 모델에게 요약 전문가로 동작하도록 지시.
- **HumanMessage**: 요약할 원본 텍스트를 입력.
- LLM이 응답한 결과를 **response.content**로 반환하여 요약문을 저장.

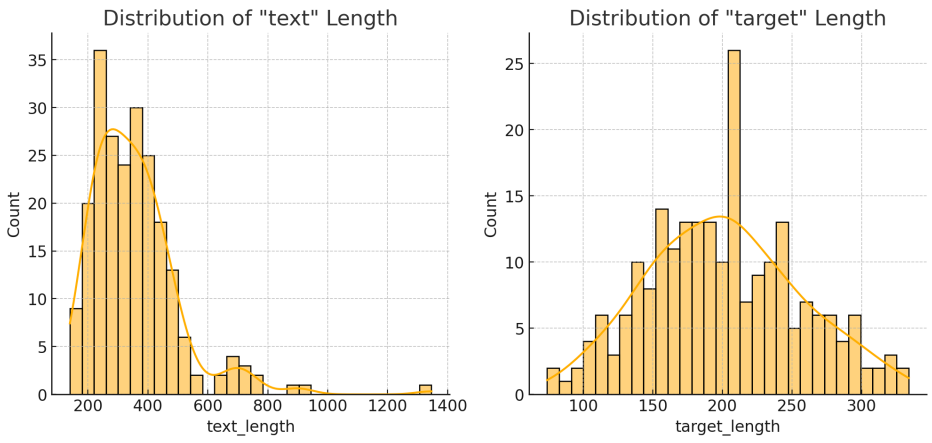
○ 확장문 생성

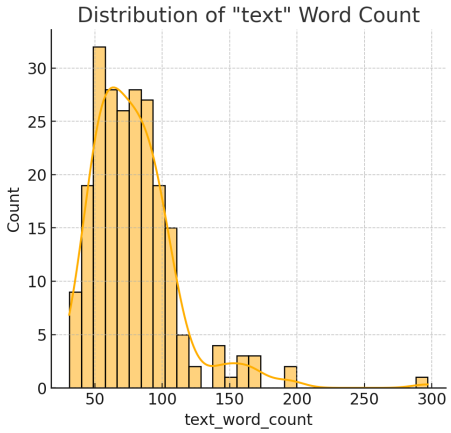
- 확장문을 생성하는 과정도 요약문 생성과 유사하며, 모델에게 문장을 확장하도록 지시하는 메시지를 보냅니다.

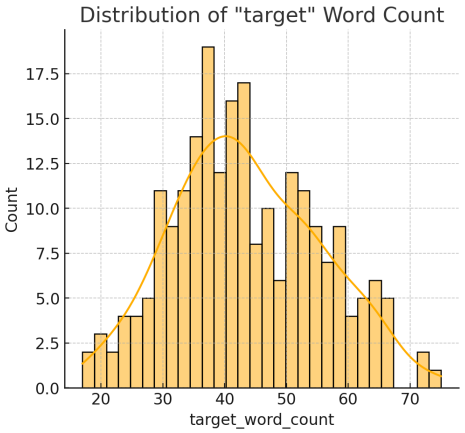
```python
def expansion_text(llm, text):
    messages = [
        SystemMessage(content="너는 전문적인 문장 분량 확장자야. 주어진 텍스트의 분량을 사실적으로 늘려줘."),
        HumanMessage(content=text)
    ]

    response = llm(messages)

    return response.content
```

- **SystemMessage**: 모델이 문장을 확장할 수 있도록 지시.
- **HumanMessage**: 확장할 원본 텍스트를 입력.
- 결과를 **response.content**로 반환하여 확장문을 저장.

3.4 EDA (요약문)

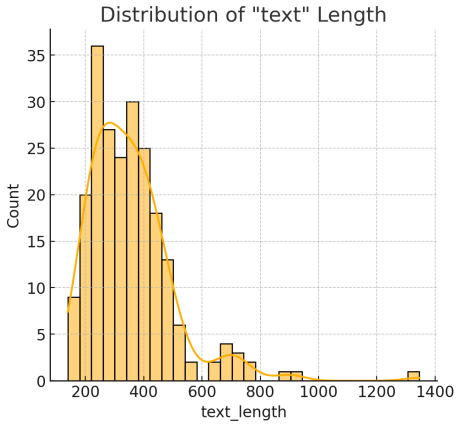


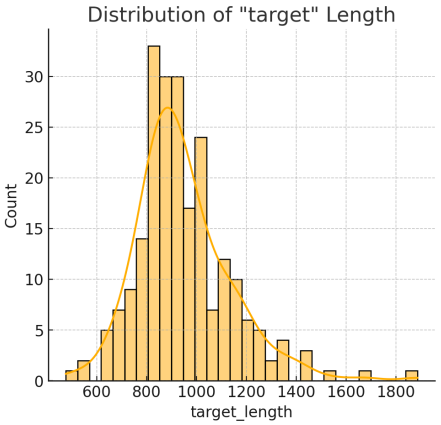


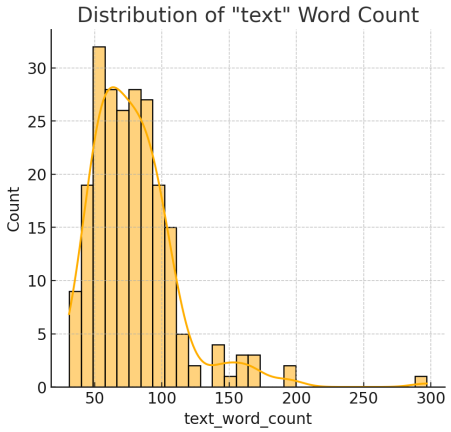


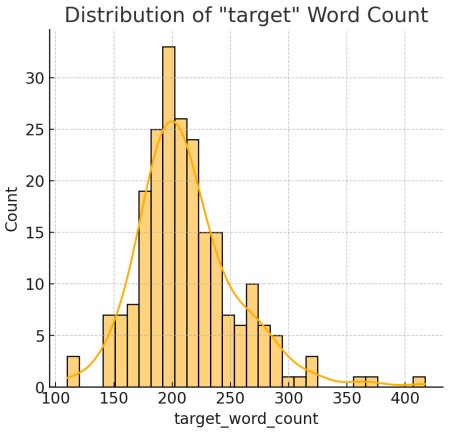
- 원본 텍스트 분석
  - 총 데이터 수 : 224
  - 평균 길이 : 약 355.6자
  - 최소 길이 : 140자
  - 최대 길이 : 1,346자
  - 텍스트 길이는 대부분 250~420자 범위에 분포함
  - 히스토그램 분석 결과, 긴 텍스트 보다 중간 길이(300~400)의 텍스트가 많음
  - 단어 수 분석
    - 평균 단어 수 : 78.9개
    - 최소 단어 수 : 31개
    - 최대 단어 수 : 297개
    - 대부분의 단어 수는 60~90개 범위에 분포함
- 생성한 요약문 분석
  - 총 데이터 수 : 224
  - 평균 길이 : 약 200.4자
  - 최소 길이 : 74자
  - 최대 길이 : 334자
  - 요약문 길이는 대부분 160~240자 범위에 분포함
  - 단어 수 분석
    - 평균 단어 수 : 43.5개
    - 최소 단어 수 : 17개
    - 최대 단어 수 : 75개
    - 대부분의 단어 수는 35~50개 범위에 분포함

3.5 EDA (확장문)









- 원본 텍스트
  - 총 데이터 수 : 224
  - 평균 길이 : 약 355.6자
  - 최소 길이 : 140자
  - 최대 길이 : 1,346자
  - 텍스트 길이는 대부분 250~420자 범위에 분포함
  - 히스토그램 분석 결과, 긴 텍스트 보다 중간 길이(300~400)의 텍스트가 많음
  - 단어 수 분석
    - 평균 단어 수 : 78.9개
    - 최소 단어 수 : 31개
    - 최대 단어 수 : 297개
    - 대부분의 단어 수는 60~90개 범위에 분포함
- 확장문
  - 총 데이터 수 : 224
  - 평균 길이 : 약 948자
  - 최소 길이 : 476자

- 최대 길이 : 1,887자
- 텍스트 길이는 대부분 840~1028자 범위에 분포함
- 단어 수 분석
  - 평균 단어 수 : 212.8개
  - 최소 단어 수 : 110개
  - 최대 단어 수 : 417개
  - 대부분의 단어 수는 180~230개 범위에 분포함
  - 원본 텍스트보다 약 2~3배 긴 문장으로 구성됨.

3.6 파인튜닝 형식 변환

- 하나의 문장으로 '지시문'과 '모델 응답'을 구분하여 명시함.
  - 아래의 형태와 같이 파인튜닝 형식으로 변환
    - "### Instruction : {지시문} \n\n ### Response : {target}<eos>"

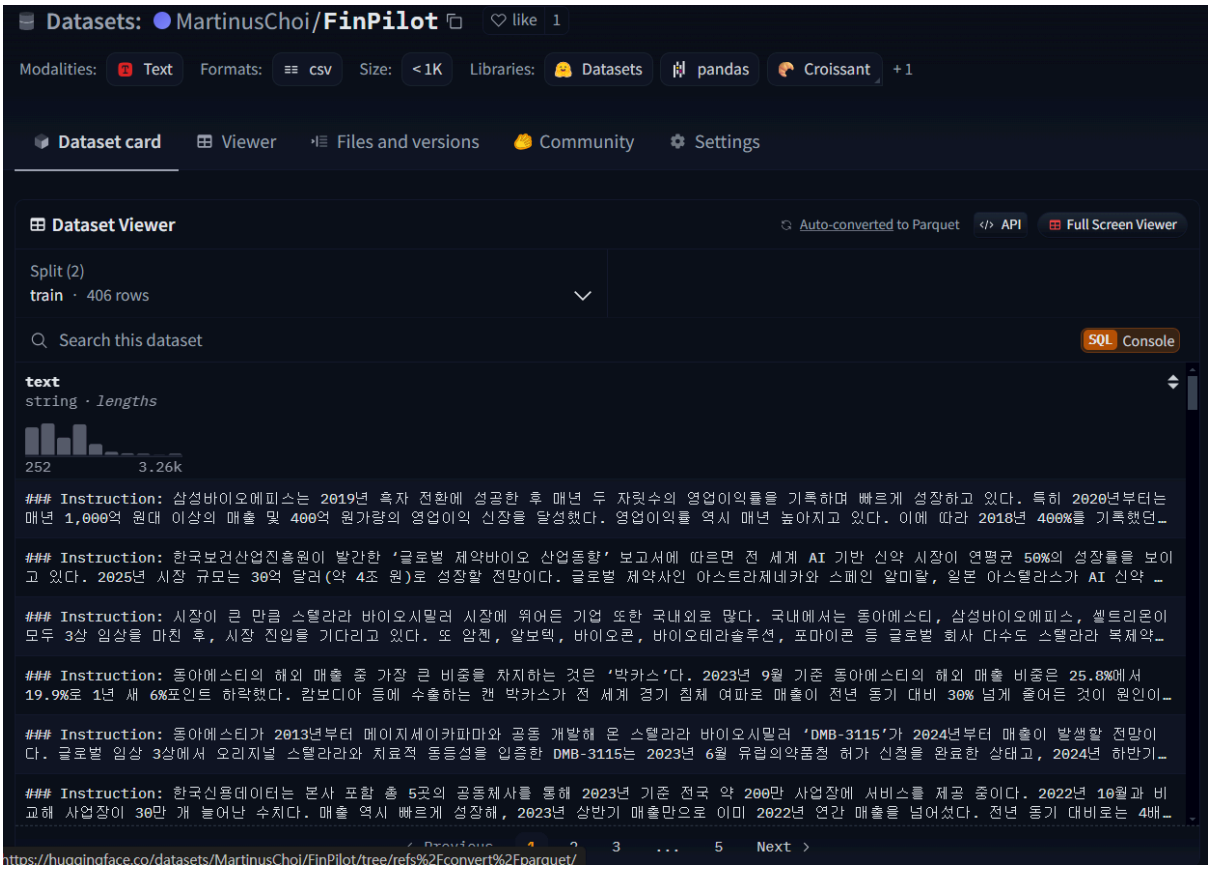
```
def generate_prompt(texts, targets):
    output_texts = []
    for text, target in zip(texts["text"], targets["target"]):
        prompt = f"### Instruction: {text}\n\n### Response: {target}<eos>"
        output_texts.append(prompt)
    return output_texts
```

- 변환된 파인튜닝 프롬프트를 DataFrame으로 병합하여 저장함

text
Instruction: 삼성바이오에피스는 2019년 혹자 전환에 성공한 후 ...
Instruction: 한국보건산업진흥원이 발간한 '글로벌 제약바이오 산업동...
Instruction: 시장이 큰 만큼 스텔라라 바이오시밀러 시장에 뛰어든 ...
Instruction: 동아에스티의 해외 매출 중 가장 큰 비중을 차지하는 ...
Instruction: 동아에스티가 2013년부터 메이지세이카파마와 공동 개...
...
Instruction: 컬러강판 시장의 경쟁이 격화되고 있다. 철강에 디자인...
Instruction: 그동안 축적된 기술력을 토대로 최근 급격하게 성장하고...
Instruction: 전자 제품의 판매는 사회문화적 변화를 비롯해 미세먼지...
Instruction: 우아한형제들은 2010년 6월 배달 앱 '배달의민족'...
Instruction: 경북 구미 불산 유출사고, 가습기 살균제 피해와 기후...

3.7 학습, 평가 데이터 분할

- 요약 프롬프트 데이터, 확장 프롬프트 데이터를 포함하여 도합 448개의 sample을 404개의 train_dataset과 44개의 test_dataset으로 분할하였음
- 처리가 완료된 데이터셋은 [HuggingFace Dataset Hub](#)에 저장하여 팀원 모두 쉽게 사용할 수 있도록 관리하였음



3.8 데이터 검증

- 원본 텍스트 데이터에 대해 적절한 요약문과 확장문 데이터셋이 구성되었는 지 Case Test
- 원본 텍스트

미-중 갈등에 따른 공급망 재편, 러시아-우크라이나 전쟁 등 지정학적 리스크와 고금리로 인한 소비 부진 등 복합적인 위기로 산업계가 어려움을 겪고 있다. 한국무역협회는 2023년 수출이 부진할 것으로 전망했고, 디스플레이업계도 글로벌 경기 침체에 따라 수출 전망에 빨간불이 켜졌다. 세계적으로 불황이 장기화되면서 TV, 태블릿PC 등 전자제품의 제고가 쌓이면서 OLED 등 패널 수요가 급감한 것이다. 특히 LED 분야에는 값싼 제품을 공급하는 중국 업체들의 난립으로 국내 업계 성장이 크게 둔화됐다.

- 요약문

미-중 갈등, 러시아-우크라이나 전쟁, 고금리로 인한 소비 부진 등으로 산업계가 어려움을 겪고 있다. 한국무역협회는 2023년 수출 부진을 전망했으며, 글로벌 경기 침체로 디스플레이 업계의 수출 전망도 악화되었다. 전자제품 재고가 쌓이면서 OLED 패널 수요가 급감했고, LED 분야에서는 중국 업체들의 저가 제품 공급으로 국내 업계 성장이 둔화되었다.

- 확장문

한국무역협회는 이러한 상황을 반영하여 2023년 수출이 부진할 것으로 전망하고 있다. 특히, 디스플레이 업계는 글로벌 경기 침체의 직격탄을 맞고 있다. 세계 경제의 불황이 장기화되면서 TV, 태블릿PC 등 전자제품의 재고가 쌓이고 있으며, 이는 OLED 등 패널 수요의 급감을 초래하고 있다. 디스플레이 업계는 기술 혁신을 통해 새로운 수요를 창출하려고 노력하고 있지만, 글로벌 경기 침체라는 거대한 파도 앞에서는 역부족인 상황이다.

특히 LED 분야에서는 값싼 제품을 공급하는 중국 업체들의 난립이 국내 업계의 성장을 크게 둔화시키고 있다. 중국 업체들은 저렴한 가격을 무기로 시장 점유율을 확대하고 있으며, 이는 국내 업체들에게 큰 위협으로 작용하고 있다. 국내 업체들은 품질과 기술력으로 차별화를 시도하고 있지만, 가격 경쟁에서 밀리면서 어려움을 겪고 있다. 이러한 상황은 국내 LED 업계의 장기적인 성장 가능성을 위협하고 있으며, 이에 대한 대책 마련이 시급한 상황이다.

결국, 이러한 복합적인 위기 상황 속에서 한국 산업계는 새로운 돌파구를 모색해야 하는 과제를 안고 있다. 기술 혁신과 새로운 시장 개척을 통해 위기를 극복하고, 지속 가능한 성장을 이루기 위한 노력이 절실히 요구되고 있다.

- 결론 : 원본 텍스트에 대한 적절한 요약문 및 확장문 데이터셋이 구성된 것을 확인 할 수 있었다.

4. 향후 사용 계획

- 향후 sLLM 파인튜닝을 진행할 예정이며, 이를 위해 Hugging Face Dataset Hub에 저장된 데이터셋을 로드하여 모델을 학습할 계획입니다. 초기 데이터셋을 기반으로 한 1차 학습 이후에도, 지속적인 성능 향상과 모델의 적응성을 높이기 위해 추가 데이터 수집 및 정제 과정을 거칠 예정입니다.
- 또한, 데이터셋의 품질을 유지하고 최신 데이터를 반영하기 위해 주기적인 데이터셋 업데이트 및 버전 관리를 수행할 계획입니다. 이를 통해 모델이 최신 트렌드를 반영할 수 있도록 개선하고, 특정 도메인 또는 응용 분야에서의 성능을 극대화할 수 있도록 할 것입니다.