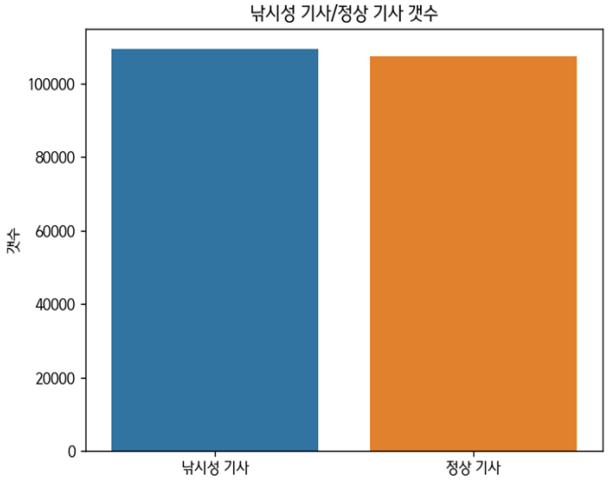
낚시성 기사 유무/유형 판독, 기사 요약 및 빈출 단어 프로그램

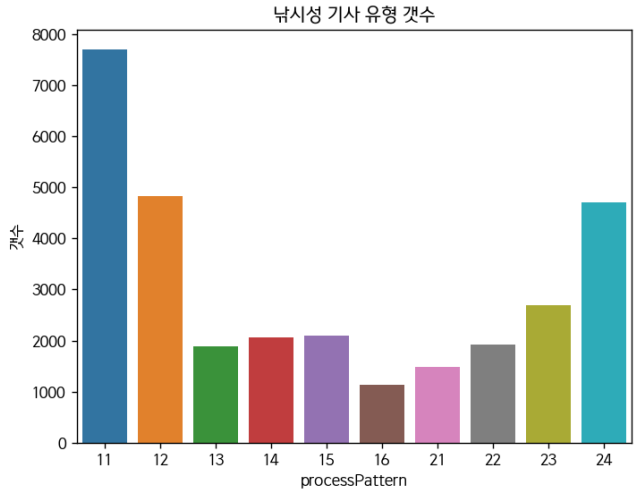
▌담당 EDA, 전처리, LLM 배포

▌프로젝트 파이프라인 설계

URL — Crawling (title, text) → Text-to-Text Large Language Model Fine Tuning, 기사 요약, 빈출 단어 빈도 수 Module — LangChain → Gradio를 통해 낚시성 기사 유무, 유형 기사요약 및 빈출단어 빈도 수 출력

▌EDA 데이터 탐색을 통해 필요한 feature 선택 및 종속변수 데이터 분포도 확인





1. 종속변수 중 하나인 낚시성 기사 유무에 대한 분포도는 거의 균등하게 이루어져 있는 것을 확인
2. 낚시성 기사 모든 유형의 분포는 불균형하게 되어 있어 학습 데이터에 대한 결정 필요

팀 프로젝트 경험

기간 23.10.16-23.11.10 (4주)

데이터 요약

INPUT 기사 제목(String), 기사 본문(String)

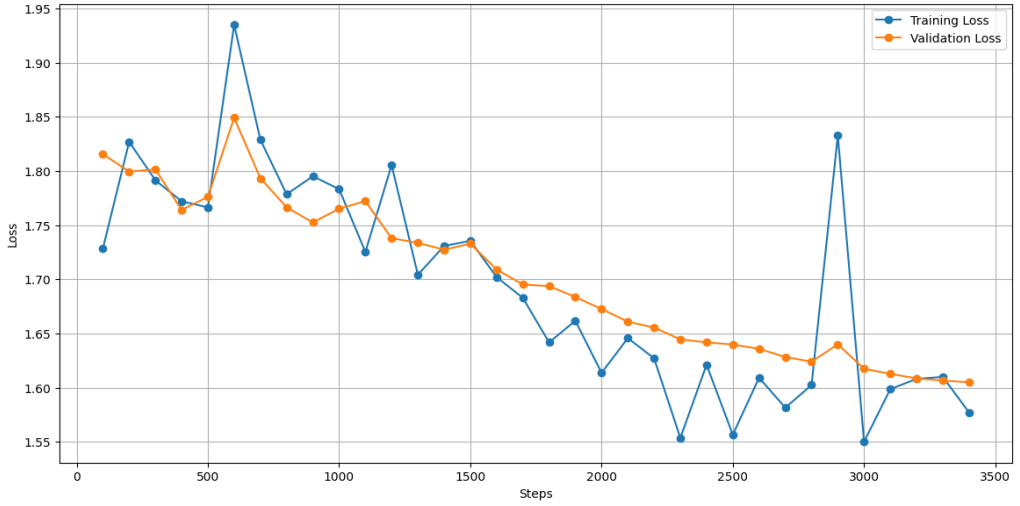
OUTPUT 낚시성 기사 유무(0,1, Integer), 낚시성 기사 유형(Integer)

데이터 전처리

- 기사부제목이 Null인 Feature 삭제 (자원 문제로 모든 데이터가 학습되지 않아 데이터 축소)
- 낚시성 기사 유형 코드 이름 변경 (Integer ▶ String, 학습 및 배포 시 사용자가 쉽게 인식할 수 있도록 변경)
- 데이터 불균형 해결을 위한 샘플링 (낚시성 기사 유형별로 각 1,000개씩 추출하여 균등한 학습데이터 완성)
- 기사 내용 text 특수 문자 제거 (학습시 성능이 좀 더 향상되기 위해 처리)

최종학습모델을 위해

최종학습모델 : Polyglot Ko 1.3B



1. LLaMa2, Polyglot 학습
   ▶ Polyglot 구축
2. 모델 사이즈 조정
   ▶ 12.8B, 5.8B, 1.3B 중 1.3B으로 구축
3. max_steps, learning_rate 조정
4. Prompt 수정

팀 프로젝트 경험

기간 23.10.16-23.11.10 (4주)

배포

Hugging Face → LangChain → gradio

학습모델 업로드 → LLM 모델과 인터페이스 연결 → UI 구축

결과