17조
조은비, 박범창, 좌진우

한국어 기반 인공지능 생성 텍스트 탐지

11주차 – 테스트 설계서

Contents 목차

‘한국어 기반 인공지능 생성 텍스트 탐지’ 프로젝트 11주차
테스트 설계서 과제 발표 자료

INTRO 연구 질문 / 가설

연구 질문

RQ1.
한국어 기반 AI 생성 텍스트의 생성 모델 분류 시스템은 기존의 단순 AI/비AI 이진 분류 방식보다 더 높은 분류 정확도를 달성하는가?

RQ2.
AI 생성 텍스트 분류 시스템은 입력된 텍스트의 길이나 주제 유형에 따라 분류 정확도의 차이를 보이는가?

가설

H1.
다중 클래스 분류 모델을 통해 생성 주제별 AI 텍스트를 분류하는 방식은 기존 이전 분류 모델보다 분류 정확도 및 활용성이 유의미하게 향상될 것이다.

H2.
본 분류 시스템은 짧은 텍스트, 주제별 텍스트 등 다양한 조건에서도 안정적으로 높은 분류 성능을 유지할 것이다.

TEST PLAN 목적

목적

GPT-3.5, GPT-4, Claude 2.1, Gemini 등 다양한 대형 언어 모델(LLM)로 생성된 한국어 텍스트를 대상으로, 각 생성 모델을 구분하는 분류기의 성능을 평가한다.

모델 간 문체 및 응답 특성의 차이를 기반으로 분류가 가능한지를 검증하며, 프롬프트 유형이나 텍스트 길이 변화에 따른 모델의 민감도도 함께 측정한다.

TEST PLAN 테스트 상세

독립 / 종속 변수 정의

변수 유형	항목	설명
독립 변수	생성 모델 종류	GPT-3.5, GPT-4, Claude 2.1, Gemini 1.5
독립 변수	입력 텍스트 유형	뉴스 요약형, 설명형, 감정표현형, 주관적 서술형 등
종속 변수	분류 정확도	예측된 생성 모델과 실제 라벨이 일치하는 비율
종속 변수	F1-score	Precision 과 Recall 의 조화 평균
종속 변수	혼동 행렬	클래스별 분류 성능 오차 분석

TEST PLAN 테스트 상세

실험 대상 / 환경

- 데이터 : 동일 프롬프트에 대해 각 모델(GPT-3.5, GPT-4, Claude, Gemini)에서 수집한 라벨링된 텍스트 데이터셋 (총 5,000개 이상)

- 실험 환경: Python 3.x, Google Colab Pro, PyTorch, Huggingface Transformers

- 모델 구성: KoBERT, KoELECTRA 기반 다중 클래스 분류기

- 사용 도구: OpenAI API, Gemini API, Claude API (수작업 포함)

TEST PLAN 테스트 관리

실험 절차 요약

1. 프롬프트 템플릿 50 ~ 100개 수집
2. 각 프롬프트를 LLM들에 입력하여 출력 수집
3. 라벨링 및 전처리 (길이 제한, 중복 제거 등)
4. KoBERT / KoELECTRA 기반 분류 모델 학습 (80/20 split)
5. 모델별 분류 정확도, F1-score 측정
6. 주제 유형별 성능 비교, 혼동 행렬 분석

TEST PLAN 테스트 관리

측정 지표 및 도구

정량 지표 : Accuracy, Precision, Recall, F1-score, Confusion Matrix

정성 지표 : 오분류 사례, 문체 패턴 차이 분석

도구 : scikit-learn, pandas, matplotlib, seaborn

TEST CASE 테스트 케이스

ID	대상 (모델/조건)	실험 조건	테스트 데이터	평가 지표	예상 결과
TC-1	Baseline 모델 (BERT)	사전학습만 수행, 분류 미학습	GPT-4 텍스트 1000 개	Accuracy, F1-score	Accuracy 25%
TC-2	KoBERT	기본 학습	전체 모델 클래스 균등 분포	Accuracy, F1-scorc	Accuracy 60%
TC-3	KoELECTRA	기본 학습	동일 데이터셋	Accuracy, F1-score	Accuracy 65%
TC-4	KoELECTRA + Prompt Embedding	프롬프트 정보 포함 학습	동일 데이터셋	Accuracy, Confusion Matrix	Accuracy 70%
TC-5	문체 비교 실험	감정형 vs 설명형 프롬프트 그룹 분리	모델별 유형별 데이터	Precision, Recall	GPT-4 > Claude 예상
TC-6	소량 학습 테스트	학습 데이터 20% 축소	전체 동일 조건	Accuracy	Accuracy 감소 예상

TEST CASE 검증 기준

지표	설명
Accuracy	전체 예측 중 정답으로 맞힌 비율
Precision	각 생성 모델(class)에 대해 모델이 예측한 것 중 정답 비율
Recall	실제 정답 데이터 중 모델이 맞힌 비율
F1-score	Precision 과 Recall 의 조화 평균
Confusion Matrix	각 생성 모델 간 오분류 현황을 시각화한 행렬

Thank You