P R O J E C T D O C U M E N T

Test Result Document

Project Name	한국어 기반 인공지능 생성 텍스트 탐지

17 조

202202501 조은비
202002494 박범창
202002565 좌진우

지도교수: 이종률 교수님 (서명)

Table of Contents

1. Introduction

1.1. Objective

본 연구의 목표는 일반적으로 구현되어있는 GPT Killer, 생성형 AI판독기와는 다르게 생성형 AI들의 특징을 분석해 LLM 각각의 특징을 뽑아내고 이를 기반으로하여 어떤 AI가 한국어 텍스트를 만들었는지 판별하기 위함임.

이번 연구에서는 KoELECTRA와 RoBERTA 모델에 대해 성능 측정을 진행함. KoBERT 모델 또한 함께 사용해보았으나, 주목할만한 성능이 나오지않아 해당 모델을 제외하고 실험을 진행함. 따라서, 원래 KoELECTRA 모델에 대해서만 진행하려 했으나, 추가적인 선행연구 조사에서 RoBERTA를 사용해 좋은 성능을 낸 결과를 확인해 추가해서 비교분석을 진행함.

실험 진행은 다음과같은 방식으로 진행하였음. 먼저 AIHub에서 제공해주는 '뉴스 기사 기계독해 데이터'를 활용하도록함. 전체 데이터는 양이 많아 1000자~1200자에 해당하는 기사 500개를 선정한 후, 이를 각각의 LLM에 넣은 후 250~300자 정도로 요약을 해달라고 Query를 날림. 이후 이렇게 얻은 기사 요약데이터에 라벨링을 수행한 후 어떤 LLM이 작성했는지 학습시키는 방법을 사용함.

모델	Precision (정밀도)	Recall (재현율)	F1-score
Claude	0.78	0.92	0.85
Copilot	0.79	0.85	0.82
Gemini	0.94	0.81	0.87
GPT-4	0.88	0.77	0.82

TestSet을 통한 정확도는 약 84%가 나왔으며, 다른 성능지표들은 위와 같음. 표를 해석해보자면, Claude는 재현율은 매우 높으나 정밀도는 낮은것으로 미루어보아, Claude를 잘 찾아내나, 다른 LLM 요약 기사를 Claude로 오해하는 경우가 많음. Gemini는 정밀도는 가장 높으나 재현율은 낮은것을 통해 분석해보면, Gemini가 요약한 일부 기사 데이터는 다른 LLM이 만든것으로 분류된 것을 알 수 있음. GPT-4의 경우 정밀도는 준수한 편이나 재현율이 매우 낮은 것을 알 수 있음. GPT-4또한 다른 LLM이 만들었다고 오해받는 일이 다수인 것으로 보임. Copilot의 경우 정밀도는 약간 낮으나, 재현율은 준수한 모습을 보여줌.

해당 연구에대해 결론을 내리자면, 비교적 적은 데이터셋을 가지고도 유의미한 성능지표를 도출해낸다는 것을 알 수 있음. 다만, 한계점을 이야기해보자면 실험 설계당시 목표로 했던 90% 정확도에는 도달하지 못했음. 이에대한 원인을 분석해보자면 데이터셋 부족을 원인으로 지목해 볼 수 있을 것 같음. 따라서 후속 연구에서는 데이터셋을 1000개, 2000개로 늘려가면서 성능지표를 관찰해보는 것 또한 좋을 것이라 생각됨.

2. Experiment Result Report

1. 서론

1.1 실험 개요

- 한국어기반 생성형 AI 출처 다중 분류기(4진 분류)

- 입력 데이터:
  - 전체 데이터 수: 2000개 (GPT-4, Copilot, Claude, Gemini)
    - tRAINING set : 1600개
  - VALIDATION/TEST SET : 400개

- 사용 모델: KoELECTRA, RoBERTA
- 가동 환경: Google Colab

1.2 실험 방법

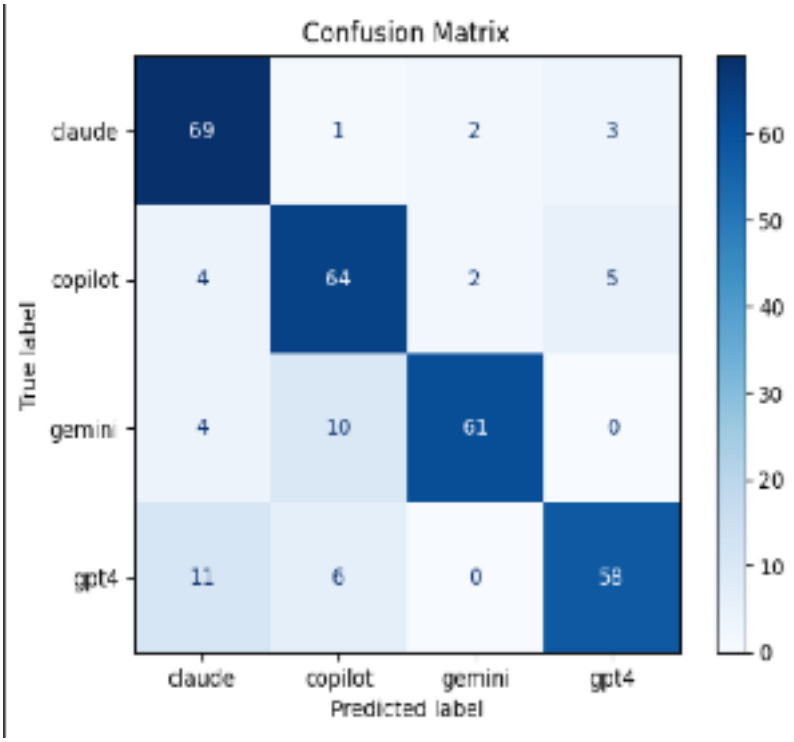
- 생성형 AI를 통해 요약한 뉴스 기사들을 Data로 활용해, 출처 AI를 맞추는 방식으로 모델을 학습시킴
- 평가기준은 정확도, 정밀도, 재현율, F1-Score, 혼동행렬 등을 활용해 측정하도록 함

2. 테스트 결과 상세

2.1 테스트 결과 개요

- 수치/성능 요약표(정확도, 성공률, 전송률 등)

드딘	Precision(정밀도)	Recall (재현율)	F1-score
Claude	0.78	0.92	0.85
Copilot	0.79	0.85	0.82
Gemini	0.91	0.81	0.87
GPT-4	0.88	0.77	0.82



2.2 테스트 결과 상세 분석

- model을 돌리기위해 Google Colab을 사용했는데, 무료버전을 사용해서(RoBERTa-base) 제약이 많았음(RoBERTa-large 모델을 사용했으면 정확도가 더 높았을 것)
- 데이터 수가 적어 정확도가 예상한 것 보다 낮음. 증강기법 사용하면 모델분류가 어려워 질 수 있었고, 수작업으로 데이터를 만들기엔 시간이 촉박했음. 이때문에 특히 Copilot 탐지가 어려움
- 특정 모델간의 차이는 찾기 힘들었는데, Claude와 GPT-4 생성 모델의 혼동률이 높게나옴

2.3 실험 결과의 한계와 위협 요인

- 학습 데이터셋이 부족해 설계단계에서 목표로 했던 90%에는 도달하지 못함
- 4개의 LLM만 학습을 진행해, 다른 모델로 생성된 데이터에 대해서는 오분류

3. 결론

[핵심발견]

- Gemini와 Claude는 서로 다른 특징이 드러나며, 분류가 상대적으로 쉬움
- GPT-4와 Copilot은 비슷한 지표를 보이는 것으로보아 유사한 요약패턴을 사용할 확률 존재

[후속연구방향]

- 요약 텍스트 간의 특징 분석
- 데이터셋을 확장한 후 성능지표 추이 관찰