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

문제정의서(연구개발계획서)

Project Name	한국어 기반 인공지능 생성 텍스트 탐지

17 조

202202501 조은비
202002494 박범창
202002565 좌진우

지도교수: 이종률 교수님 (서명)

Document Revision History

REV#	DATE	AFFECTED SECTION	AUTHOR
1	2025/04/03	4, 5 번 문항 작성	좌진우
2	2025/04/03	1, 2 번 문항 작성	박범창
3	2025/04/04	3번 문항 작성 및 문서 정리	조은비

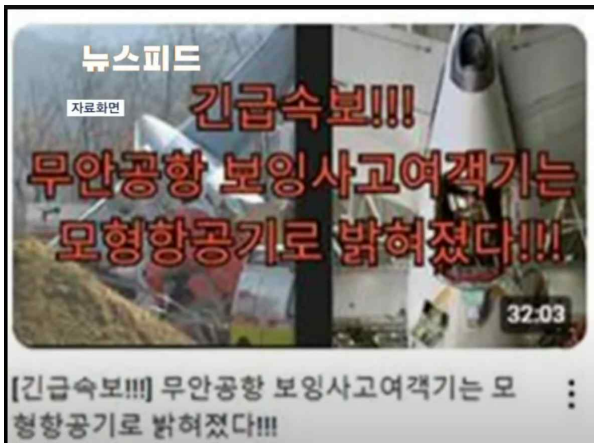
Table of Contents

1. 연구 개발의 필요성

한국어기반 AI 가짜뉴스 탐지가 필요한 이유

2024년 미국 대선 후보자 유세시기에 트럼프의 지지층을 늘리기위해 흑인 유권자들이 트럼프를 지지한다는 내용의 기사들이 쏟아짐. 이는 단순 기사 뿐만이 아니라 AI생성 이미지에 스토리텔링을 넣는 식으로 사람들을 속여넘기는 등 치밀한 면모를 보여줌.

또한 2025년에 있었던 무안공항 여객기 참사 사건이 발생했을 때, 이 사건은 그래픽으로 조작된 사건이며 여객기는 모형이고 유가족 또한 모두들 배우라고 주장한 60대 유튜버도 있었음. 현재 그는 기소되어있는 상태.



이렇게 가짜뉴스들이 다양하게 판을 치고있는 만큼 AI를 통해 가짜뉴스를 판별하고 신뢰도있는 정보를 수집하며, 시민들이 가짜 뉴스에 흔들리지 않도록 도울 수 있는 도구가 필요하다고 생각해 프로젝트를 시작하게 됨.

국내외 연구개발 현황

국내외 연구 개발 현황은 문제점개요서에서 다루었듯 기존의 연구는 영어 기반에 치중되어있으며, 한국어 기반 연구는 미비한 현황임. 추가적으로 짧은 글(500자 미만)에 대해서는 탐지를 잘 수행하지 못하며 그나마 있는 한국어 기반 연구도 제한된 주제(에세이)에서 연구한 결과라는 부분임. 게다가 영어 기반 모델들도 약 60%의 가짜뉴스 탐지율로 실생활에서 뉴스기사의 신뢰도를 측정하기에는 무리가 있는 결과물임.

가짜뉴스 관련 법령

형법

- 307조(명예훼손) : 공연히 허위사실을 적시해 사람의 명예를 훼손한 경우
- 309조(출판물 등에 의한 명예훼손) : 출판물로 허위사실을 적시한 경우
- 314조(업무방해죄) : 허위사실을 유포해 타인의 업무를 방해한 경우

정보통신망법 70조 : 사람을 비방할 목적으로 정보통신망을 통해 허위사실을 유포한 경우
공직선거법 250조(허위사실 공표죄) : 선거에서 후보자 등에 대한 허위사실을 공표한 경우
국가보안법 4조 : 사회질서의 혼란을 조성할 허위사실을 날조하거나 유포한 경우

2. 연구 개발의 목표 및 내용

프로젝트 연구목표

목표 : 한국어 가짜뉴스 탐지 AI모델 개발
연구개발 내용 : 현재 나와있는 다양한 AI모델들을 사용, 파인튜닝, 프롬프트 설정을 통해 성능 비교 분석 진행 후 어떤 방식이 한국어 가짜뉴스 탐지에 적합한지 선별. 이후 선별된 모델을 경량화 해 스마트폰에 집어넣어 인터넷이 연결되지 않은 off-line 상태에서도 동작할 수 있도록 하는 것이 궁극적인 목표.

프로젝트를 위한 핵심문제

- 가짜뉴스 판별기준 : 처음 나온 기사의 경우 비교할 정보 부족, 기사들 끼리 미러링(똑같은 내용 복사 후 올리기, 이후 개별적으로 수정). 일부 내용은 진실이지만 일부 내용은 거짓일 경우 처리문제

- 도메인 별 신뢰도 측정기준 : 가짜뉴스 판별을 위해 기사가 나온 출처의 신뢰도 점수를 함께 사용할 계획인데, 이 과정에서 각 신문사 별 신뢰도를 어떤 방식으로 측정할지 결정해야함.

- 모델 및 데이터셋 구축 : 뉴스 기사들의 Fact check를 수행할 기관이나 수행된 Data가 있어야하데

SNU 팩트체크가 2024년부로 서비스 중단. 모델 또한 KoELECTRA, KoBERT, LLM 등 다양한 모델 중 어떤 모델이 가장 성능이 좋을지 예측하고 선정해야함. Fine-tuning의 경우 돈과 시간이 들어가기때문에 선정기준을 설정해야 함.

- 모델의 성능평가기준 : 모델의 성능평가 지표로 정확도, 재현율, 정밀도 등 일반적으로 사용되는 수치를 사용할 수 있지만 이런 수치들이 뒤섞여있을 때 어떤 모델이 더 우수하다고 할 수 있는지 결정할 기준을 설정해두어야 모델을 선별하고 기능을 개선해 나갈 수 있음.

- 탐지 회피기법 대응방안 마련 : 최근 생성형 AI들은 사람처럼 보이게 하기위해 패러프레이징, 어휘 다양화, 오타삽입 등 다양한 '인간화' 전략을 구사하고 있음. 이에 따른 대처방안을 마련해야할 필요성 존재.

사용자에게 제공하고 싶은 가치

사회를 혼란스럽게 만드는 가짜뉴스들의 범람 앞에서 믿을 수 있는 도구를 가지고 가짜 뉴스들에 휘둘리지 않고 믿음직스러운 기사들을 보며 어둠 속 하나의 등불같은 길을 알려주는 가치를 제공하고 싶음

브레인스토밍 시각화

3. 이해당사자 인터뷰/ 설문 인사이트

한국어 기반 인공지능 생성 텍스트 탐지 기술의 필요성과 현실적인 문제점을 파악하기 위해 총 7명을 대상으로 인터뷰 및 설문조사를 진행하였다. 인터뷰는 AI 및 자연어처리(NLP) 분야의 전문가를 대상으로 진행하였으며, 설문조사는 최근 ChatGPT 및 SNS 기반 콘텐츠를 접한 일반 사용자 5명을 대상으로 실시하였다. 이를 통해 실제 사용자 및 전문가의 관점에서 공통적으로 나타나는 문제점과 요구사항을 도출하였다.

문제점	문제점 파악 방법	문제 상세 기술		
		이해당사자	고충/니즈	이유
한국어 기반 AI 생성 텍스트 탐지의 어려움	인터뷰 및 설문조사 (교수 1명, 관련 업계 종사자 1명, 일반 사용자 5명)	NLP 연구자, 교수, 일반 사용자 (SNS/뉴스 이용자)	AI가 생성한 글을 사람이 쓴 글과 구별하기 어려워 정확한 판별 도구 필요	영어 기반 도구는 한국어에서 탐지 정확도가 낮고, 실제 사용자들은 혼란을 겪고 있음
기존 AI 탐지 모델의 한계 (짧은 문장 탐지 실패)	AI 관련 논문 조사 및 사용자 대상 설문 응답 분석	언론 소비자 및 SNS 사용자	댓글이나 뉴스 제목 등 짧은 문장에서도 AI 탐지가 가능했으면 좋겠음	기존 탐지 모델은 긴 문장 위주로 학습되어 짧은 문장에서 탐지 성능이 매우 낮음
AI 생성 가짜 뉴스의 확산 위험	AI 뉴스 사례 분석 및 사용자 인터뷰	일반 뉴스 소비자, 커뮤니티 사용자	신뢰할 수 있는 정보와 가짜 정보를 구분하고 싶어함	가짜 뉴스가 AI로 쉽게 생성되어 실제 뉴스처럼 유포되고 있어 혼란 발생
생성형 AI 기술에 대한 사회적 불신	SNS 커뮤니티 사례 및 뉴스 댓글 분석	일반 사용자, 교육기관 관계자	AI 생성 정보의 출처나 신뢰도를 표시해주는 시스템 원함	사용자들은 정보의 진위를 판별하기 어려워 AI 콘텐츠에 대한 신뢰가 낮아지고 있음

현재 한국어 기반의 인공지능 생성 텍스트 탐지 시스템은 초기 단계에 머물러 있으며, 실제 적용에는 여러 가지 한계가 존재한다.

1. **기존 탐지 기술의 언어적 한계**
   현재 시중에 존재하는 AI 생성 텍스트 탐지 모델(DetectGPT, GPTZero, OpenAI Text Classifier 등)은 대부분 영어를 대상으로 개발되었으며, 영어 문장 구조와 어휘 특성을 기반으로 작동한다. 그러나 한국어는 교착어로서 어순이 유동적이고 조사와 어미 변화가 많기 때문에 동일한 구조를 적용했을 때 성능이 급격히 떨어진다.
2. **짧은 텍스트에서의 탐지 성능 저하**
   KoELECTRA나 KoBART 등 일부 한국어 특화 모델을 활용한 연구에서는 에세이나 기사 본문처럼 문장 길이가 긴 경우 비교적 높은 탐지 성능을 보였으나 뉴스 제목, 댓글 짧은 SNS 포스트와 같은 짧은 문장에서는 탐지 정확도가 현저히 낮다.
3. **한국어 학습 데이터의 부족**
   대부분의 탐지 모델은 영어 기반의 XSum, SQuAD 등 대규모 정형 데이터셋을 사용하여 학습되었으며 한국어에 특화된 학습 데이터는 뉴스 기사에 한정되어 있다. 실제 사용 환경에서는 커뮤니티 글, 트위터, 블로그 등 다양한 형태의 텍스트가 존재하지만, 이들을 포함한 데이터셋은 거의 없는 상태이다.
4. **사용자 경험 및 실생활 적용 한계**
   일반 사용자들은 생성형 AI가 만든 콘텐츠를 자주 접하고 있지만 이를 즉각적으로 식별할 수는 없다 AI가 만든 콘텐츠를 사용자에게 알림으로 표시하거나 위험도를 평가해주는 서비스가 없어 정보 소비 과정에서 혼란을 겪고 있다. 특히, 정치적·사회적 이슈가 걸린 콘텐츠에서는 신뢰성 판단의 기준이 모호하여 잘못된 정보가 빠르게 확산되는 문제로 이어진다.

현재 시스템은 영어 중심, 긴 문장 중심, 정형화된 환경 중심으로 설계되어 있어 한국어 환경과 실제 사용 맥락에서는 정확성·실용성·적용성 모두에서 한계가 뚜렷하다. 이를 해결하기 위한 요구사항은 아래와 같다.

- 한국어 특화 탐지 모델 구축
- 짧은 문장에 강한 모델 구조 설계
- 실사용 환경을 반영한 데이터셋 수집 및 학습
- 탐지 결과를 사용자에게 직관적으로 제공하는 UI/UX 설계

4. 기대 효과 및 향후 확장 가능성

본 연구를 통해 한국어 기반 인공지능 생성 텍스트 탐지 및 가짜 뉴스 판별 기술을 개발함으로써 다음과 같은 효과를 기대할 수 있다.

⑴ 사용자 관점

- 인공지능이 생성한 허위 정보로부터 보호받을 수 있으며, 신뢰성 있는 정보 획득 가능
- 가짜 뉴스 탐지 시스템을 통해 올바른 정보 소비 환경 조성

⑵ 사회적 관점

- 허위 정보 확산 방지를 통해 사회적 혼란 및 피해 최소화
- 언론 및 공공기관에서 신뢰할 수 있는 정보 제공을 위한 필터링 도구로 활용 가능
- 교육 및 공공 정책 수립에 있어 객관적인 정보 제공

⑶ 산업적 관점

- 언론사, SNS 플랫폼, 검색 엔진 등에서 탐지 시스템을 적용하여 콘텐츠 신뢰성 향상
- AI 기반의 탐지 기술을 활용한 신규 비즈니스 모델 개발 가능

향후 확장 가능성

⑴ AI 탐지 모델의 정확도 향상 및 성능 개선

- 딥러닝 및 최신 AI 탐지 기법을 적용하여 성능을 지속적으로 향상
- 다양한 AI 모델(GPT-4, Claude 3 등)이 생성하는 텍스트 탐지 가능 여부 분석

⑵ 한국어 데이터셋 확대 및 모델 최적화

- 다량의 한국어 데이터셋을 확보하여 모델 학습을 고도화

⑶ 다양한 플랫폼 적용 가능성 검토

- SNS, 블로그, 뉴스 기사 등 실생활 데이터를 활용하여 탐지 성능 평가
- 각 플랫폼의 특성에 맞춘 탐지 기법 연구

⑷ 탐지 회피 기법 대응 모델 개발

- 텍스트 변형, 의미 왜곡 등의 탐지 회피 기법을 분석하여 대응 모델 설계
- 인간이 개입한 허위 정보 생산 방식까지 탐지할 수 있는 기술 연구

5. 연구 개발의 추진전략 및 방법

본 연구의 목표는 한국어 기반 인공지능 생성 텍스트 탐지 모델을 개발하고, 가짜 뉴스 및 허위 정보 판별 기술을 고도화하는 것이다. 이를 위해 다음과 같은 전략을 추진한다.

⑴ 사전 조사 및 인공지능 관련 기초 학습

- 기존 AI 탐지 모델 및 탐지 기술에 대한 문헌 조사
- 자연어 처리 개념 및 기술 이해

⑵ 데이터셋 구축 및 탐지 모델 구현

- 한국어 가짜 뉴스 및 AI 생성 텍스트 데이터셋 수집 및 정제
- SNS, 블로그, 뉴스 기사 등의 데이터를 확보하여 실생활 적용 가능성 분석
- KoBART, KoELECTRA, KLUE-BERT 기반 모델 개발 및 학습
- AI 생성 텍스트 vs 인간 작성 텍스트 차이 학습

⑶ 가짜 뉴스 탐지 기법 연구

- 기존 탐지 기법 분석 및 성능 비교

⑷ 모델 성능 평가 및 개선

- 실생활에서 발생하는 가짜 뉴스 및 허위 정보 데이터에 대한 성능 테스트
- 지속적인 피드백을 통해 모델 고도화

목표

⑴ 정량적 목표

- 연구 논문 1편 이상 작성 및 제출
- 파일럿 테스트에서 85% 이상의 탐지 정확도 달성

⑵ 정성적 목표

- 가짜 뉴스 탐지 및 AI 생성 텍스트 탐지 기술의 사회적 기여도 분석

6. AI 도구 활용 정보

사용 도구	*GPT-4*
사용 목적	*인터뷰 내용 요약 정리*
프롬프트	● *해당 내용을 바탕으로 인터뷰 내용을 요약해 줘*
반영 위치	1. *이해당사자 인터뷰/ 설문 인사이트 (p.6)*
수작업 수정	*있음(표 형식으로 수정, 문장 정리)*

7. 참고문헌(Reference)

AI가 생성한 '가짜 이미지'로 흑인 유권자 공략하는 트럼프 지지자들
https://www.bbc.com/korean/articles/c29wl2kp7nyo

"무안공항 참사 여객기는 모형"…가짜뉴스 주의보 [뉴스피드]
https://www.youtube.com/watch?v=Pbr03fMxI7g

SNU팩트체크 7년 만에 중단... "한국 언론자유 퇴보"
https://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0003055180

ZhuMingjian 외 9명(2023) “A Million-ScaleBenchmark forDetecting AIGenerated”, 『SCOPUS』,

박소현. (2024). 학생 작문 에세이와 AI 생성 텍스트의 구분을 위한 KoELECTRA 기반 탐지 모델 적용 방법. 한국정보통신학회 종합학술대회 논문집, 28(2), 495-497.

고상훈, 안현철. (2024). 대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성 지식경영연구, 25(4), 113-127.

박현주, 김병준, 김부근. (2024-06-26). 인공지능 생성 텍스트 탐지 기술의 한국어 적용. 대한전자공학회 학술대회, 제주.

Chang, Y.-H. (2024) “A Study on Deep Learning-Based Detection of AI-Generated News,”Annual Conference of KIPS. 한국정보처리학회, pp. 698-700. doi: 10.3745/PKIPS.Y2024M10A.698.

Wang, X., et al. (2025). Have LLMs reopened the Pandora’s box of AI-generated fake news?Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL).

Li, Y., et al. (2024). MAGE: Machine-generated text detection in the wild.Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).