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

문제점 개요서

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	2025/03/20	Survey Paper, Limitations and Research Gaps 추가	박범창
2	2025/03/20	Survey Paper, Limitations and Research Gaps 추가	조은비
3	2025/03/20	Survey Paper, Limitations and Research Gaps 추가	좌진우

Table of Contents

1. Survey Paper - Limitations Focus

번호	연구 제목(저자)	저널/컨퍼런스 (연도)	주요 내용 요약	한계점
1	A Million-Scale Benchmark for Detecting AI-Generated Image(Zhu Mingjian 외 9명)	SCOPUS(2023)	확산모델(Diffusion Model)과 적대 신경망(GAN)을 사용한 대규모 생성형 AI이미지 데이터 셋 구축 및 다양한 성능평가 기존 탐지모델(CNNSpot, Spec)등은 확산 모델로 생성된 이미지에 성능이 좋지않음.	동일한 생성기로 생성, 훈련한 경우 높은 정확도를 보여주었으나(98.5%) cross-generator 평가를 진행할경우 크게 저하(54.9%)
2	학생 작문 에세이와 AI 생성 텍스트의 구분을 위한 KoELECTRA 기반 탐지 모델 적용(박소현)	한국정보통신학회(2024)	학생 essay를 기반으로 KoELECTRA 모델을 적용해 10-Fold 교차검증을 통한 생성형 AI탐지 평가. KoELECTRA모델은 한국어 자연어 처리(NLP)에 특화된 모델. 정확도, 정밀도, 재현률, F1점수 모두 95% 이상으로 성공적인 실험	학습데이터인 학생 essay는 주제가 한정되어있어 특정 데이터 셋에만 적용할 수 있는 연구결과라는 점이 한계점.
3	대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성(고상훈, 안현철)	한국지식경영학회, KCI(2024)	한국어 환경에서 LLM이 가짜뉴스 탐지에 적합한지, 어떤 요약방식이 탐지에 효과적이며 최적의 방법은 무엇인지 가짜뉴스의 조작방식(정보왜곡, 사실-허위 결합, 감정적 접근) 기존 가짜뉴스 탐지방법(뉴스 컨텐츠 분석, 배경정보 활용) 가짜뉴스를 추출, 생성형으로 나눠 어느 방법이 탐지에 적합한지 연구 -> 생성형이 탐지에 더 적합한 모습 확인 하지만, 이유기반 프롬프트 설정시 기존 뉴스 탐지(59.8%)보다 우수한 성능(62%) 확인	학습 데이터셋이 전문 기자가 작성한 데이터라 단순하고 감정적인 콘텐츠(ex 트위터, 커뮤니티 토막글 등)에는 적용 불가. 이유기반 프롬프트 방법을 사용해도 정확도는 62%정도라 실사용 하기에는 무리
4	인공지능 생성 텍스트 탐지 기술의 한국어 적	대한전자공학회 (2024)	영어권 AI 생성 텍스트(AIGT) 탐지 모델(DetectGPT, RADAR)의 한국어 적용 가능성을 분석. 다양한 한국어 도메인(뉴스, 과학	기존 영어권 AIGT 탐지 모델을 한국어에 직접 적용한 경우 탐지 성능이 0.5 AUROC 이하로 감

	용 (박현주, 김병준, 김부근)		등)에 대한 AIGT 탐지 실험 수행.	소. 한국어의 언어적 특성을 고려하지 않은 모델 구조로 인해 성능 저하 발생. 최신 AI 생성 모델(Claude 3 등)이 인간과 유사한 텍스트를 생성하여 탐지가 더욱 어려워짐
5	딥러닝 기반 인공지능 생성 뉴스 탐지 (장예훈)	ACK 2024 학술 발표대회 (2024)	AI가 작성한 뉴스가 증가함에 따라 가짜뉴스 탐지 필요성이 대두됨. KoBART, KoELECTRA, 그리고 두 모델을 결합한 앙상블 모델을 사용하여 AI 생성 뉴스 탐지 모델을 개발. 실험 결과 KoBART 모델이 Accuracy 0.9995로 가장 높은 성능을 기록.	기사 제목과 같은 짧은 문장에서는 탐지 성능이 현저히 떨어짐. KoELECTRA, KoBART 모델 모두 제목 분류에서 낮은 AUROC(0.55~0.62)와 비정상적인 Precision 및 Recall 결과를 보임. AI 뉴스 생성에 GPT-3.5-turbo를 사용하였으며, 최신 한국어 특화 모델(KULLM, KLUE-BERT 등) 적용이 필요함.
6	Have LLMs Reopened the Pandora's Box of AI-Generated Fake News? (Xinyu Wang et al.)	NAACL (2025)	인간이 LLM을 활용하여 가짜 뉴스를 생성하고, 별도의 참가자들이 해당 뉴스가 가짜인지 판별하도록 실험을 설계함. 실제 뉴스 탐지의 경우 LLM(GPT-4o)이 인간보다 68% 더 정확하게 실제 뉴스를 판별했지만, 가짜 뉴스의 경우 LLM과 인간 모두 약 60% 정도의 정확도로 비슷한 정도를 보임 (사람은 미리 가짜 뉴스의 존재 사실을 알았음)	- 데이터셋의 한계: 뉴스 데이터가 참가자가 직접 LLM을 사용하여 제작한 것이므로, 실제 온라인에서 발견되는 가짜 뉴스와 다를 가능성이 있음. 실제 뉴스 데이터는 주류 언론에서(신뢰도가 높은) 수집되었지만, 현실에서는 신뢰도가 낮은 출처의 기사도 많아 탐지 정확도가 달라질 수 있음 - LLM 탐지 성능의 한계: 현재 LLM의 탐지 성능이 가짜 뉴스에 대해 60% 수준이기 때문에, 현실적인 탐지 도구로 사용하기 어려움. 또한, 가짜 뉴스 제작자가 LLM 탐지를 회피하는 기술을 발전시킬 경우 탐지율이 더욱 낮아질 수 있음

7	MAGE: Machine-generated Text Detection in the Wild (Yafu Li et al.)	ACL (2024)	AI 생성 텍스트 탐지를 위해 다양한 도메인과 27개의 LLM에서 생성된 텍스트를 포함한 MAGE 테스트베드를 구축함. 특정 도메인에서는 높은 성능(최대 99%)을 보였으나, 새로운 도메인이나 LLM에서는 성능이 감소(최저 68.4%). 다양한 공격에 대해 탐지 모델의 성능이 크게 저하되어, 간단한 변형만으로도 탐지 회피가 가능함. 완벽한 AI 생성 텍스트 탐지는 아직 어렵고, 탐지 성능을 높이기 위해 하이브리드 모델 및 새로운 기법이 필요함.	LLM이 생성한 텍스트는 연구팀이 사전 정의한 프롬프트를 기반으로 생성됨 → 실제 AI가 자율적으로 생성하는 텍스트보다 구조적일 가능성이 높음. 모든 탐지 실험이 정형화된 데이터셋에서 수행됨 → 실제 환경에서는 사용자가 직접 생성하거나 편집한 텍스트가 많기 때문에 탐지 성능이 다를 수 있음.

2. Limitations and Research Gaps

번호	기존 연구	한계점	연구 필요성	본 연구의 기여
1	A Million-Scale Benchmark for Detecting AI-Generated Image(Zhu Mingjian 외 9명, 2023)	동일한 생성기로 생성, 훈련한 경우 높은 정확도를 보여주었으나(98.5%) cross-generator 평가를 진행할경우 크게 저하(54.9%)	cross-generator 평가에서도 높은 성능을 낼 수 있도록 다양한 generator 학습 및 평가 필요.	학습 이미지 개수 증가 -> 탐지 성능 향상확인 얼굴 및 예술이미지에 대한 높은 정확도 저해상도(112x112), 가우시안 블러처리, JPEG 형식등에서 우수한 성능 확인
2	학생 작문 에세이와 AI 생성 텍스트의 구분을 위한 KoELECTRA 기반 탐지 모델 적용(박소현, 2024)	학습데이터인 학생 essay는 주제가 한정되어있어 특정 데이터 셋에만 적용할 수 있는 연구 결과라는 점이 한계점.	다양한 주제에 대한 학습데이터가 필요함. 이를 통한 재 학습을 실시한다면 광범위한 주제에서의 생성형 AI 텍스트 탐지를 실현할 수 있을 것.	제한된 주제에서의 KoELECTRA 모델의 우수한 성능 입증 기존에 주로 사용되던 BERT모델보다 KoELECTRA 모델의 우수한 부분(인간 - ai간 미세한 차이 식별)
3	대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성(고상훈, 안현철, 2024)	학습 데이터셋이 전문 기자가 작성한 데이터라 단순하고 감정적인 콘텐츠(ex트위터, 커뮤니티 토막글 등)에는 적용 불가. 이유기반 프롬프트 방법을 사용해도 정확도는 62%정도라 실사용 하기에는 무리	데이터셋에 다양한 데이터(단순하고 짧은 토막글) 등 포함해 학습이 필요함 정확도를 향상시키기위한 다양한 방법 연구할 필요성 있음.	LLM 프롬프트 기반 가짜뉴스 탐지 성능평가 다양한 방법(요약, 추출, 프롬프트 설정)을 통한 성능 향상법 탐구
4	인공지능 생성 텍스트 탐지 기술의 한국어 적용 (박현주, 김병준,	기존 영어 기반 AI 생성 텍스트(AIGT) 탐지 모델(DetectGPT, RADAR)은 주로 영어와 같은 굴절어에서만 실험되었으	한국어의 언어적 특성을 반영한 AIGT 탐지 모델 개발이 필요함. 기존 연구에서 사용한 토큰화 및 데이터 전처리 방	한국어 전용 데이터셋을 구축하고, AIGT 생성 및 탐지 모델을 학습하여 한국어 환경에서의 탐지 성능 분석 고려.

	김부근, 2024)	며, 한국어와 같은 교착어에서는 성능이 저하됨. 기존 연구에서 사용한 데이터셋(XSum, SQuAD 등)은 영어 기반이며, 한국어에 적용 시 성능 저하. Claude 3가 생성한 AI 텍스트가 인간과 유사하여 DetectGPT 및 RADAR 기반 탐지 모델의 성능이 더욱 저하됨.	법이 한국어에 최적화되지 않음. 한국어에 특화된 데이터셋(한국어 뉴스, KorQuAD, 과학 데이터셋 등)을 사용하여 실험이 필요함. 최신 AI 생성 모델(Claude 3, GPT-4 등)의 탐지를 위한 새로운 접근 방식이 필요함.	Claude 3, GPT-4 등 최신 생성 모델에 대한 탐지 성능을 분석하고, 탐지 성능을 향상시키기 위한 새로운 접근법을 제안함.
5	딥러닝 기반 인공지능 생성 뉴스 탐지 (장예훈, 2024)	KoBART, KoELECTRA 모델 기반 탐지 모델이 AI 생성 뉴스 내용을 정확하게 탐지했지만, 뉴스 제목과 같은 짧은 문장에서는 탐지 성능이 크게 저하됨. 특히 KoBART 모델의 Precision과 Recall이 0.0으로 계산되는 등 정상적인 성능을 보이지 않음. AI 뉴스 생성에 사용된 GPT-3.5-turbo는 최신 한국어 특화 모델이 아니므로, 탐지 성능의 일반화에 한계가 있음.	짧은 문장(뉴스 제목)에 대한 탐지 성능을 향상시키기 위한 새로운 접근법이 필요함. 또한, 최신 한국어 특화 AI 모델(KULLM, KLUE-BERT 등)을 활용하여 AI 뉴스 생성 및 탐지 성능을 개선할 필요가 있음.	기사 제목 탐지 성능 향상을 위해 기존 KoBART, KoELECTRA 모델의 fine-tuning을 최적화하고, 추가적인 데이터 증강 기법을 적용할 예정. 한국어에 특화된 KULLM, KLUE-BERT 등의 모델을 활용하여 탐지 성능을 비교하고, 새로운 접근법을 제안함. 단일 모델 대비 성능이 높은 앙상블 모델(soft voting, stacking 등)을 최적화하여 탐지 성능을 극대화함.
6	Have LLMs Reopened the Pandora's Box of AI-Generated Fake News? (Xinyu Wang et al. 2025)	- 데이터셋의 한계: 뉴스 데이터가 참가자가 직접 LLM을 사용하여 제작한 것이므로, 실제 온라인에서 발견되는 가짜 뉴스와 다를 가능성이 있음 - LLM 탐지 성능의 한계: 현재 LLM의 탐지 성능이 가짜 뉴스에	AI가 생성한 가짜 뉴스를 탐지하는 실험을 수행했지만, 실제 환경에서의 적용 가능성이 제한적이다. 실험이 특정 도메인(주류 언론 뉴스)에 국한되었으며, 다양한 뉴스 출처나 언어적 변형(패러프레이징, 번역 등)에 대한 탐지 성능은 충분히 검증되지 않았다. 또한	본 연구는 한국어 기반 AI 생성 가짜 뉴스 및 허위 정보 탐지 모델을 개발하여 기존 연구의 한계를 해결할 수 있다. 다양한 한국어 뉴스 및 온라인 플랫폼(SNS, 블로그 등) 등에서 데이터셋을 구축하여 실제 환경에서 발생하는 가짜 뉴스 탐지 성능을 검증할 수 있다.

		대해 60% 수준이기 때문에, 현실적인 탐지 도구로 사용하기 어려움. 또한, 가짜 뉴스 제작자가 LLM 탐지를 회피하는 기술을 발전시킬 경우 탐지율이 더욱 낮아질 수 있음	한국어 환경에서 AI 생성 가짜 뉴스를 탐지하는 연구는 부족하여, 한국어 특성을 반영한 탐지 모델이 필요하다.	
7	MAGE: Machine-generated Text Detection in the Wild (Yafu Li et al. 2024)	LLM이 생성한 텍스트는 연구팀이 사전 정의한 프롬프트를 기반으로 생성됨 → 실제 AI가 자율적으로 생성하는 텍스트보다 구조적일 가능성이 높음. 모든 탐지 실험이 정형화된 데이터셋에서 수행됨 → 실제 환경에서는 사용자가 직접 생성하거나 편집한 텍스트가 많기 때문에 탐지 성능이 다를 수 있음.	이 연구는 AI 생성 텍스트 탐지를 위해 다양한 도메인과 LLM에서 생성된 텍스트를 분석했지만, 주로 영어 데이터에 초점을 맞추었으며, 한국어 데이터에 대한 연구는 이루어지지 않았다. 또한, 기존 탐지 모델들은 특정 도메인과 LLM에서 훈련되었을 때 높은 성능을 보였지만, 새로운 도메인과 LLM에서 생성된 텍스트 탐지는 여전히 어려운 문제로 남아 있다. 특히, 패러프레이징 공격에 취약하여 탐지 성능이 크게 저하되는 문제가 존재한다.	한국어 문법과 어휘적 특성을 반영하여 AI 모델을 개발해 볼 수 있다. 실제 뉴스, SNS, 커뮤니티 게시글 등에서 수집한 데이터를 활용한 한국어 AI 탐지 데이터셋을 구축하여 실제 환경에서의 활용 가능성을 높여볼 수 있을 것이다.