17조
조은비, 박범창, 좌진우

한국어 기반 인공지능 생성 텍스트 탐지

목차 INDEX

step 1.

연구 개요 및 1주차 피드백

연구 주제 및 목표

한국어 기반 AI 생성 텍스트 탐지

연구 배경

최근 생성형 AI 기술이 발전하면서
AI가 생성한 텍스트와 인간이 작성한 텍스트를
구분하기 어려워지고 있음.

AI 생성 텍스트는 뉴스, 블로그, SNS 등에 활용되며
가짜 뉴스 제작에도 악용될 가능성이 있음.

기존의 AI 탐지 모델은 영어권 위주로 연구되어
한국어 환경에서의 탐지 성능이 미흡함.

1

연구 방향

AI가 어떤 데이터를 탐지할 것인지 명확한 정의 필요
이미지, 뉴스, 댓글 등 다양한 유형의 AI 생성 데이터 논문 리뷰

2

AI 생성 여부를 판단하는 기준

AI 생성 텍스트를 탐지하는 기준이 모호함
AI 생성 텍스트의 특징을 분석하고 성능 평가 지표 설정 필요

3

기존 AI 탐지기와의 차별성

AI 탐지 대상 및 방식에서 차별성 필요
가짜뉴스의 기준이 다양하므로 이에 대한 추가 논의 진행

step 2.

기존 연구 및 한계점 분석

기존 연구 리뷰 (Survey Paper - Limitations Focus) - 1

A Million-Scale Benchmark for Detecting AI-Generated Image (Zhu Mingjian 외, 2023)

GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang,
Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, Yunhe Wang
Huawei Noah's Ark Lab
{zhumingjian, yunhe.wang}@huawei.com

GAN 및 Diffusion Model 기반 AI 이미지 탐지 연구

→ Cross-Generator 탐지 성능 저하 (98.5% → 54.9%)

step 2.

기존 연구 및 한계점 분석

기존 연구 리뷰 (Survey Paper - Limitations Focus) - 2

학생 작문 에세이와 AI 생성 텍스트의 구분을 위한 KoELECTRA 기반 탐지 모델 적용 (박소현, 2024)

학생 작문 에세이와 AI 생성 텍스트의 구분을 위한 KoELECTRA 기반 탐지 모델 적용 방법

Applying a KoELECTRA-Based Detection Model to Distinguish Between Student-Written Essays and AI-Generated Texts

저널정보

한국정보통신학회

한국정보통신학회 종합학술대회 논문집 | 학술대회자료

한국정보통신학회 2024년도 추계종합학술대회 논문집 제28권 제2호

2024.10 | 495 - 497 (3page)

저자정보

박소현 (고려대학교)

KoELECTRA 기반 학생 에세이 탐지 모델 개발

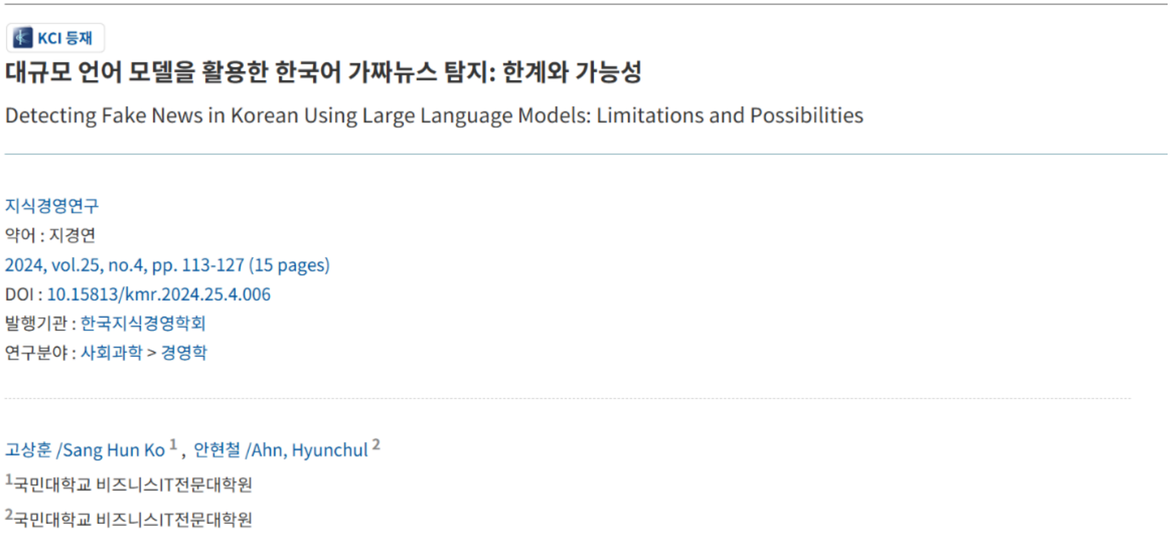
→ 특정 한정된 도메인(학생 에세이)에만 높은 성능

step 2.

기존 연구 및 한계점 분석

기존 연구 리뷰 (Survey Paper - Limitations Focus) - 3

대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성 (고상훈, 안현철, 2024)



KCI 등재

대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성

Detecting Fake News in Korean Using Large Language Models: Limitations and Possibilities

지식경영연구
약어 : 지경연
2024, vol.25, no.4, pp. 113-127 (15 pages)
DOI : 10.15813/kmr.2024.25.4.006
발행기관 : 한국지식경영학회
연구분야 : 사회과학 > 경영학

고상훈 /Sang Hun Ko [1], 안현철 /Ahn, Hyunchul [2]
[1]국민대학교 비즈니스IT전문대학원
[2]국민대학교 비즈니스IT전문대학원

LLM 기반 가짜뉴스 탐지 연구

→ 감정적이고 단순한 콘텐츠(SNS 등)에는 성능 저하

step 2.

기존 연구 및 한계점 분석

기존 연구 리뷰 (Survey Paper - Limitations Focus) - 4

인공지능 생성 텍스트 탐지 기술의 한국어 적용 (박현주 외, 2024)

인공지능 생성 텍스트 탐지 기술의 한국어 적용

A Study of AI Generated Text Detection in Korean

저널정보

대한전자공학회
대한전자공학회 학술대회 | 학술대회자료
2024년도 대한전자공학회 하계학술대회 논문집
2024.06 | 2,761 - 2,765 (5page)

저자정보

박현주 (중앙대학교)
김병준 (중앙대학교)
김부근 (중앙대학교)

DetectGPT, RADAR의 한국어 적용 실험

→ 영어권 모델을 한국어에 적용 시 성능 저하 (AUROC ≤ 0.5)

step 2.

기존 연구 및 한계점 분석

기존 연구 리뷰 (Survey Paper - Limitations Focus) - 5

딥러닝 기반 인공지능 생성 뉴스 탐지 (장예훈, 2024)

KoBART, KoELECTRA 모델 기반 탐지 연구

→ 뉴스 제목과 같은 짧은 문장에서 탐지 성능 저하 (AUROC 0.55~0.62)

step 2.

기존 연구 및 한계점 분석

기존 연구 리뷰 (Survey Paper - Limitations Focus) - 6

Have LLMs Reopened the Pandora's Box of AI-Generated Fake News? (Xinyu Wang et al., 2025)

The Reopening of Pandora's Box: Analyzing the Role of LLMs in the Evolving Battle Against AI-Generated Fake News

Xinyu Wang[1], Wenbo Zhang[1], Sai Koneru[1], Hangzhi Guo[1],
Bonam Mingole[1], S. Shyam Sundar[2], Sarah Rajtmajer[1], Amulya Yadav[1]
[1]College of Information Sciences and Technology, The Pennsylvania State University
[2]College of Communications, The Pennsylvania State University
{xzw5184, wjz5120, sdk96, hangz}@psu.edu
{bjm6940, sss12, smr48, amulya}@psu.edu

LLM과 인간의 가짜 뉴스 판별 성능 비교 연구

→ 인간보다 뛰어나지만 실제 환경과 차이가 있으며, 가짜 뉴스 탐지율 60%로 실사용 어려움

step 2.

기존 연구 및 한계점 분석

기존 연구 리뷰 (Survey Paper - Limitations Focus) - 7

MAGE: Machine-generated Text Detection in the Wild (Yafu Li et al., 2024)

MAGE: Machine-generated Text Detection in the Wild

Yafu Li[♠♣*], Qintong Li, Leyang Cui[♡†], Wei Bi[♡], Zhilin Wang[◇]
Longyue Wang[♡], Linyi Yang[♣], Shuming Shi[♡], Yue Zhang[♣†]
[♠] Zhejiang University [♣] Westlake University
[♦] The University of Hong Kong [◇] Jilin University [♡] Tencent AI lab
yafuly@gmail.com qtli@connect.hku.hk
nealcly.nlp@gmail.com linzwcs@gmail.com
{victoriabi,vinnylywang,shumingshi}@tencent.com
{yanglinyi,zhangyue}@westlake.edu.cn

다양한 LLM과 도메인에서 생성된 텍스트 탐지

→ 새로운 도메인 및 패러프레이징 공격에 취약 (99% → 68.4%)

step 2.

기존 연구 및 한계점 분석

기존 연구의 한계점

Cross-Generator 탐지 성능 저하

특정 생성기에서 학습한 탐지 모델이 다른 생성기에서 탐지 성능 저하
(예: GPT-3.5 학습 → Claude 3에서 탐지 성능 54.9%까지 하락)

한국어 AI 탐지 모델 성능 저하

DetectGPT, RADAR 등 영어 기반 탐지 모델이
한국어에서 탐지 성능 저조(AUROC 0.5 이하)

최신 AI 생성 모델 탐지 어려움

Claude 3, GPT-4 등의 최신 생성 모델이
점점 더 인간과 유사한 텍스트를 생성

step 3.

연구 목표 및 차별성

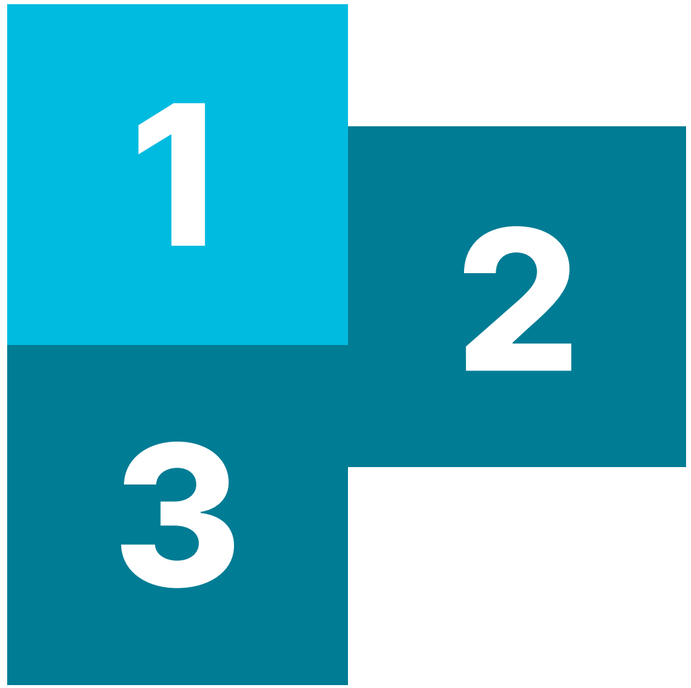
연구 목표 (Research Objectives)

한국어 기반 AI 탐지 모델 개발

- 최신 한국어 AI 모델(KULLM, KLUE-BERT 등) 적용
- 한국어 특화 데이터셋 구축 (뉴스, SNS, 커뮤니티 등)

실제 환경에서의 적용 가능성 검토

- AI 탐지 기술의 한국어 적용 확대
- 정보 신뢰도 향상에 기여



Cross-Generator 성능 개선

- 다양한 AI 생성기(GPT-4, Claude 3, KoGPT 등)에 대해 탐지 성능을 유지할 수 있도록 모델 설계

step 4.

연구 방법 및 모델 설계

연구 방법론

연구 방법	주요 내용
데이터셋 구축 및 전처리	AI 생성 텍스트(GPT-4, Claude 3, KoGPT 등) & 실제 뉴스 데이터(AI Hub, 네이버 뉴스 등) 수집
탐지 모델 설계	KoBART, KoELECTRA, KLUE-BERT 기반 모델 개발 AI 생성 텍스트 vs. 인간 작성 텍스트 차이 학습 단일 모델 vs 앙상블 모델 비교
성능 평가 (Evaluation Metrics)	Precision, Recall, F1-score, AUROC 활용

step 5.

실험 결과 및 성능 분석

실험 환경 및 설정

- AI 생성 뉴스 및 실제 뉴스 데이터셋 활용
- KoBART, KoELECTRA, DetectGPT 모델 비교 실험

Baseline 모델 비교

KoBART, KoELECTRA, KLUE-BERT 개별 성능 평가

Cross-Generator

다양한 AI 생성기에 대한 Cross-Generator 탐지 성능 분석

실제 환경에서의 탐지 가능성 평가

뉴스 기사, SNS, 블로그 등
다양한 도메인에서 모델 성능 검증

앙상블 모델 효과 분석

Soft Voting, Stacking 기법 적용 결과 분석
뉴스 제목과 본문 탐지 성능 차이 비교

step 6.

결론 및 향후 연구 방향

결론

한국어 기반 AI 생성 텍스트 탐지는 기존 영어권 모델을 그대로 적용할 경우 성능이 저하됨
KoBART, KoELECTRA, KLUE-BERT 기반 탐지 모델이 효과적이나, 뉴스 제목과 같은 짧은 문장에서 성능 저하 문제 존재
앙상블 모델을 적용하면 개별 모델 대비 성능 향상 가능

향후 연구 방향

✔ 최신 AI 모델(Gemini, GPT-4o 등) 탐지 가능 여부 분석
✔ 실시간 AI 탐지 시스템 개발 검토
✔ SNS, 블로그, 온라인 커뮤니티 데이터를 활용하여 다양한 온라인 플랫폼에서 적용 가능성 평가
✔ 패러프레이징 및 문장 변형 기법에 강한 탐지 모델 개발

Thank You