

한국어 기반 인공지능 생성 텍스트 감지

17조

202202501 조은비

202002494 박범창

202002565 좌진우



목차



아이디어 발산



아이디어 수렴 1. 데이터셋 구축

다양한 분야의 자료 수집

뉴스, 블로그, SNS 등
다양한 분야의 한국어
데이터를 수집

LLM을 활용한 확장

LLM 모델을 활용한
기존 데이터로부터의
추가 데이터 생성

자료 수집 방안

AI Hub, 팩트체크
서비스 등의 믿을 수 있는
데이터 확보

아이디어 수렴 2. 한국어 자연어 처리

문장 구조 분석

형태소 분석, 문장 구조
파악 등의 기술이 필요한가?

기존의 미흡한 모델

지금까지의 모델들은
영어 모델에 초점을 둬
한국어 기반 모델은 부족

허위 정보 판별

단순 AI 생성 텍스트 탐지에
그치지 않고, 가짜 뉴스,
허위 정보 등에 활용 가능한
모델 구축

아이디어 수렴 3. 탐지 모델 성능 평가

탐지 모델의 성능 평가

탐지 모델의 정확도, 재현율, 정밀도 등의 지표 활용

기존 모델과의 차이점

다른 기존 모델과의 성능 비교 분석

아이디어 수렴 4. 탐지 회피 기법 대응

다양한 탐지 회피 기법

패러프레이징 기법, 언어 재배치, 문장 구조 변환, 인간화 등의 수많은 탐지 회피 기법들에 대한 대처 방안 모색의 필요성

LLM의 발전

AI가 발전함에 따라 대응 방안도 함께 발전해야 함

아이디어 수렴 5. 가짜 뉴스 탐지

- 가짜 뉴스의 판별 기준은?

- 각 도메인 별 신뢰도의 기준은?

아이디어 수렴 6. AI 생성 텍스트 탐지

모델 구축의 방법

Fine-tuning, 머신 러닝, NLP, RAG 등의 다양한 기술
이용 가능

한국어 기반으로의 적용

다양한 데이터셋 구축과 함께 AI 생성 한국어 텍스트의
올바른 탐지

시각화하기

