Project Brainstorming Result

Project Name

한국어 기반 인공지능 생성 텍스트 감지

17 조

202202501 조은비 202002494 박범창 202002565 좌진우

지도교수: 이종률 교수님 (서명)

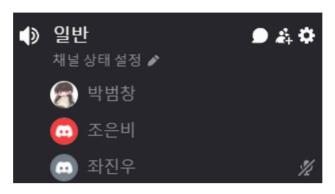
Document Revision History

REV#	DATE	AFFECIED SECTION	AUTHOR
1	2025/03/28	아이다어 발산 및 수렴 작성	좌진우
2	2025/03/28	시각화하기 작성	박범창

Contents

1.	(문제 해결을 위한) 아이디어 발산	4
2.	아이디어 수렴	5
3.	시각화하기	6

1. (문제 해결을 위한) 아이디어 발산





2. 아이디어 수렴

핵심 개념	정의 및 하위 개념 목록
한국어 자연어	1. 허위 정보를 포함한 뉴스 기사를 자동으로 판별하는 기술
처리	2. 한국어 기반 모델 개발
	3. 형태소 분석, 문장 구조 파악 등 포함
AI 생성	1. AI가 생성한 텍스트를 판별하는 기술
텍스트 탐지	2. 자연어 처리(NLP)와 머신러닝을 활용 또는 fine-tuning 사용
탐지 모델	1. 탐지 모델의 정확도, 재현율, 정밀도 등의 지표 활용
성능 평가	2. 다른 기존 모델과의 성능 비교 분석
가짜 뉴스	1. 가짜 뉴스의 판별 기준 정의
탐지	2. 각 도메인 별 신뢰도 평가 기준 설정
탐지 회피	1. 동의어 및 문장 구조 변환, 개인적인 경험 추가 등의 탐지 회피 기법
기법 대응	에 대응 방안 마련
데이터셋	1. 탐지 모델을 위한 대규모 한국어 데이터셋 확보
구축	2. AI Hub, 팩트체크 서비스 등의 자료 확보

3. 시각화하기

한국어 기반 모델 형태소, 문장 구조 파악 허위 정보 포함 기사 자동 판별 도메인 별 신뢰도 기준 설정 가짜 뉴스 판별 기준 정의 대규모 한국어 데이터셋 확보 자연어 처리 데이터셋 가짜 뉴스 구축 탐지 한국어 기반 가짜 뉴스 탐지 탐지 모델 AI 생성 성능 평가 텍스트 탐지 탐지 회피 정확도, 정밀도, 재현율 지표 AI 생성 텍스트 탐지 기법 대용 다른 모델과의 성능 비교 분석 NLP, 머신러닝, fine-tuning 노이즈 삽입 문장 구조 변형 패러프레이징 인간화 탐지 회피 기법