17조
조은비, 박범창, 좌진우

한국어 기반 인공지능 생성 텍스트 탐지

7-8주차 - 시퀀스 다이어그램

Contents 목차

17조 ‘한국어 기반 인공지능 생성 텍스트 탐지’ 프로젝트
8주차 유스케이스 다이어그램 과제 발표자료

Introduction 연구 배경

“왜 한국어 기반 가짜뉴스 탐지가 중요한가?”

- 가짜뉴스 확산 피해 급증
- 대부분 연구는 영어 중심
- 영어권 모델 → 한국어 적용 시 정확도 급감 (90% → 50~60%)

Introduction 연구 목적

한국어 기반 AI 생성 가짜뉴스 탐지 방법

- 가짜뉴스로 인한 피해 최소화
- 기사 신뢰도 향상

→ 가짜뉴스로 인한 피해 감소 및 건강한 정보 사회에 기여

Introduction 연구 질문 및 가설

연구 질문

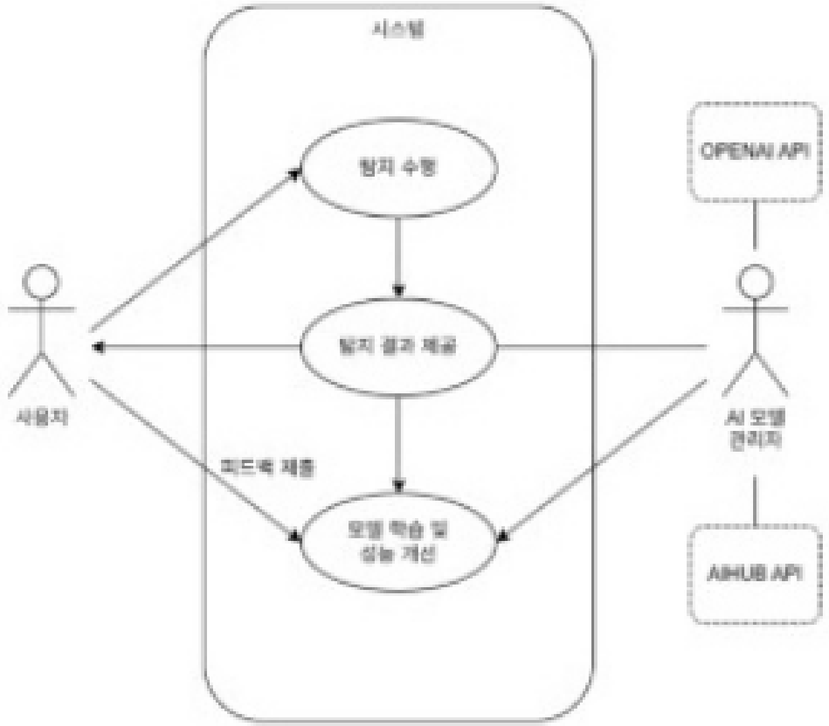
RQ1.
연구한 모델이 기존 연구모델에 비해 더 높은 정확도로 가짜뉴스를 판별하는가?

가설

H1.
AI 생성 기사를 만들기 위한 기초 기사 데이터는 AI로 생성된 기사가 아니다.

Usecase Diagram 소프트웨어의 사용 사례 Diagram

- Actor: 사용자, 관리자
- 기능: 뉴스 접근 → 탐지 수행 → 결과 확인 → 피드백



Usecase Diagram

소프트웨어 활용 사례

주요 Actor	AI 모델 관리자: 탐지 모델을 학습 및 배포하며 성능을 지속적으로 개선 일반 사용자: 뉴스 소비자. 탐지된 결과를 통해 뉴스의 신뢰도를 확인함
주요 기능 구성 요소	- 탐지 모델: KoBART / KoELECTRA 등 - AL 뉴스 생성기 (LLM 모델 기반) - 탐지 결과 시각화
입/출력 데이터	입력 데이터(결과): 실제 뉴스 기사 제목 및 내용 : AI를 통해 생성된 AI 뉴스 텍스트 출력 데이터(결과): AI 생성 가짜 뉴스 여부 (O, X, 모름) : 탐지 확률, 정밀도, 재현율 등의 결과 지표
데이터 Flow	① [뉴스 콘텐츠 접근] : 사용자가 뉴스에 접근 ② [탐지 수행] : 해당 뉴스를 탐지 ③ [탐지 결과 제공] : 화면에 다음과 같은 형태로 결과 제공 - ✔: 가짜 뉴스 아님 - ×: 가짜 뉴스 - ? : 모름 ④ [사용자 피드백] : 사용자가 해당 뉴스에 대한 의견 제출 가능
외부 시스템 연계	OPENAI API : GPT 모델을 사용하여 AI 생성 뉴스 텍스트를 생성 AIHUB API: 뉴스 기사 데이터셋 제공 그 외 다른 모델 사용 가능

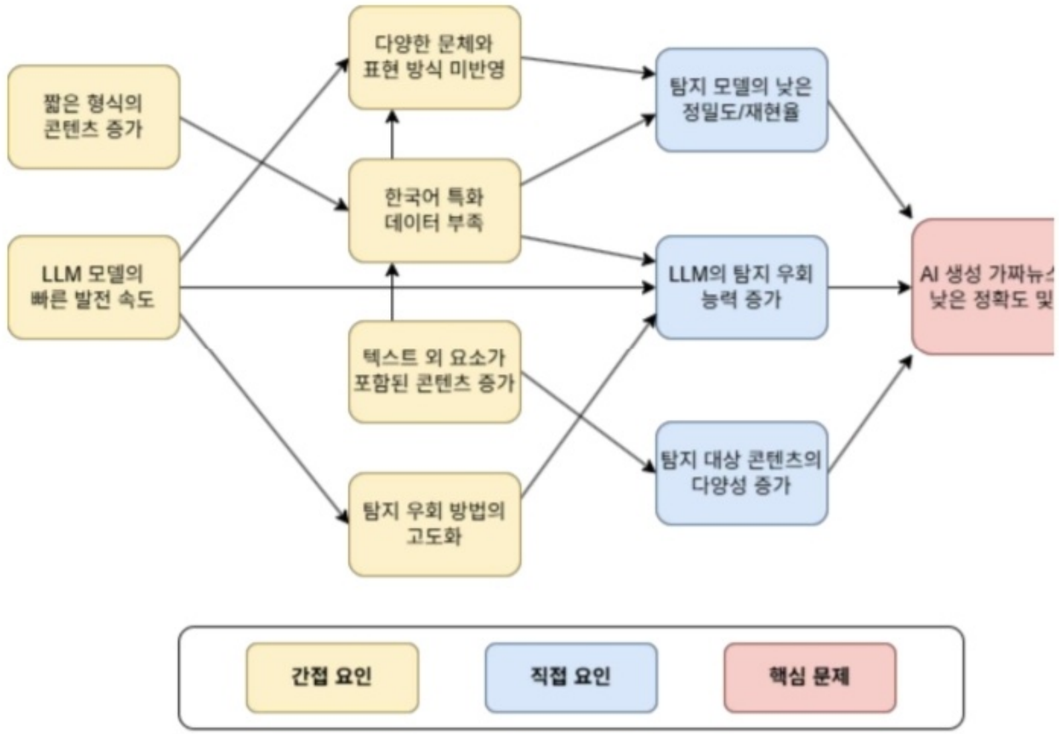
- 구성 요소: KoBART / KoELECTRA, GPT 뉴스 생성기, 시각화 모듈

- 입력: 실제 뉴스, AI 생성 뉴스
- 출력: O, X, ? (탐지 결과 및 확률 등)

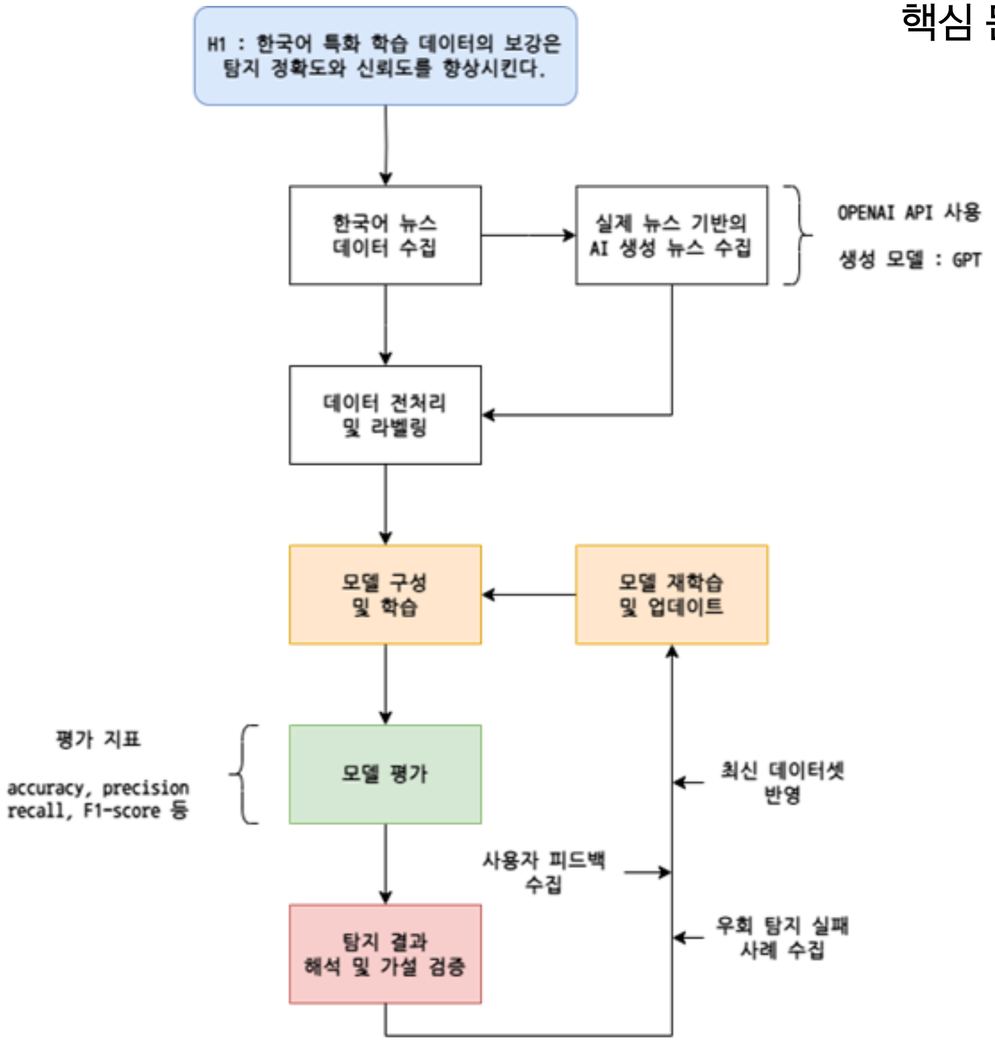
- 데이터 흐름: 뉴스 접근 → 탐지 → 결과 제공 → 피드백

Usecase Diagram 문제 해결에 대한 사용 사례 Diagram

- **탐지 정확도 향상 → 정보 신뢰성 확보**



Sequence Diagram



해결 방법에 대한 알고리즘 순서도

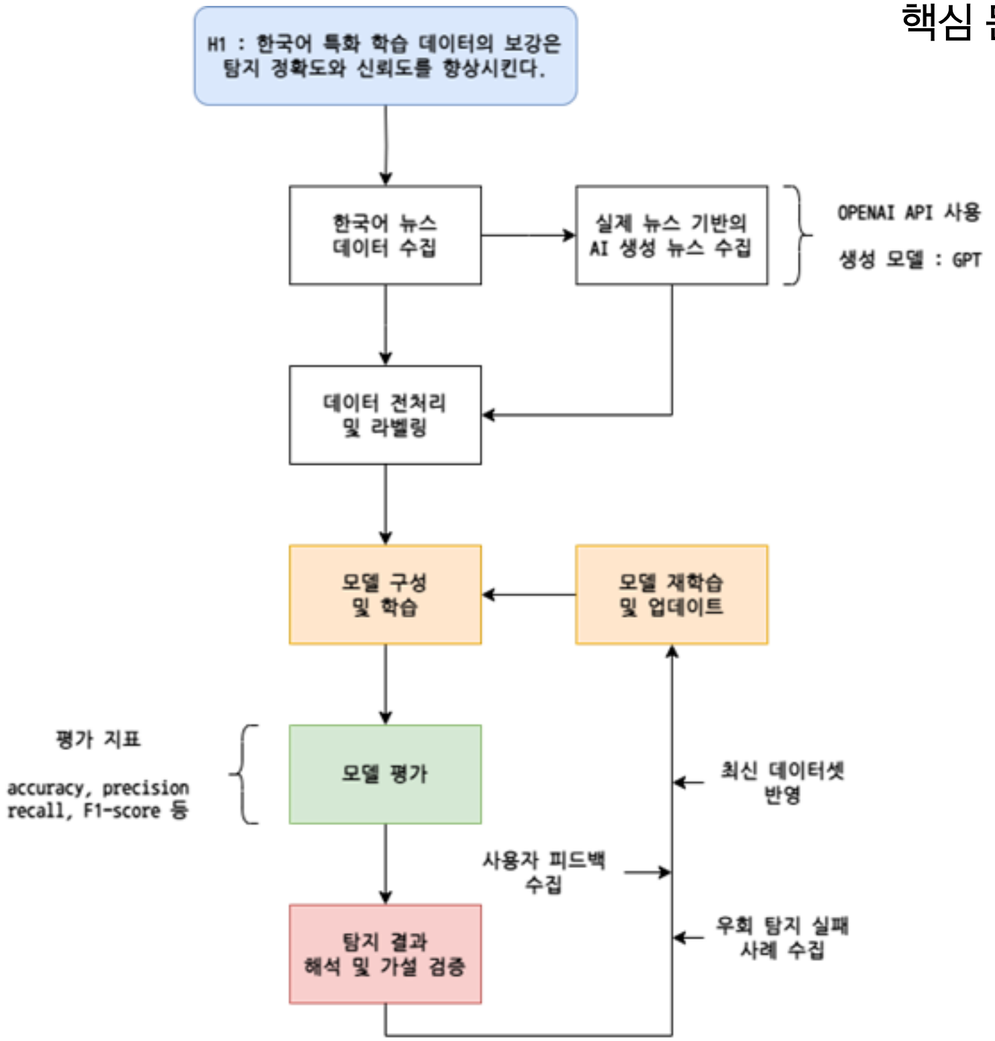
핵심 문제 정의: AI 생성 가짜뉴스 탐지의 낮은 정확도와 신뢰도

1. 데이터 수집 및 AI 사용 뉴스 생성
   - AI Hub 뉴스 데이터
   - AI 뉴스 생성은 OPENAI API 사용
   - 생성 모델 : GPT

2. 데이터 전처리 및 라벨링
   - 진짜 / 가짜 뉴스 라벨링
   - 내용 생성이 안되거나 빈약한 데이터 삭제

3. 모델 구성 및 학습
   - KoBART, KoELECTRA 모델을 사용

Sequence Diagram



해결 방법에 대한 알고리즘 순서도

핵심 문제 정의: AI 생성 가짜뉴스 탐지의 낮은 정확도와 신뢰도

4. 모델 평가
   - 정확도, 정밀도, 재현율, F1-score 등의 지표 활용
   - KoELECTRA, KoBART 등의 모델마다 각 지표 측정 및 평가

5. 탐지 결과 해석 및 가설 검증
   - 탐지 결과를 해석하고, 해당 결과가 가설을 뒷받침하는지 검증
   - 사용자 피드백, 우회 탐지 실패 사례를 수집하고, 최신 데이터셋을 반영하여 발전하는 LLM의 우회 능력에 대응할 수 있도록 함

6. 모델 재학습 및 업데이트
   - 빠르게 발전하는 인공지능에 대항할 수 있도록 지속적인 학습과 피드백 적용

Thank You