## 한국어 기사 판별용 모델

| 모델 | 기반 | 특징 | 장점 | 단점 |
|---|---|---|---|---|
| KoBERT | SKT, BERT | 한국어 전용 학습 | 국내 뉴스 기사, 챗봇에 강함 | RoBERTa보다 성능 살짝 낮음 |
| KoELECTRA | Kakao Brain | 빠르고 정확함 | 경량화 + 성능 균형 | 토크나이저 세팅이 조금 복잡 |
| KLUE-BERT | KLUE 벤치마크 | 다양한 한국어 NLP 태스크 학습 | 고성능, 한국어 문맥에 강함 | 모델 사이즈 큼 |
| KoBigBird / KoLongformer | 긴 문서 특화 | 뉴스 기사 전체 입력 가능 | 긴문장 입력 가능 | 무거움, GPU 필요 |
| KcBERT / KcELECTRA | 위키 + 뉴스 + 일상말투 | 온라인 뉴스, 댓글에 강함 | 한국어 뉴스 다양성 대응 | 비속어 포함 가능성 |

## 모델+토크나이저

| 목적 | 모델 | 토크나이저 |
|---|---|---|
| 일반 기사 문맥 분류 | KoBERT / KLUE-BERT | BertTokenizer |
| 빠른 실시간 판별 | KoELECTRA | ElectraTokenizer |
| 뉴스 전체 기사 분석 | KoBigBird | BigBirdTokenizer |
| 뉴스 + 커뮤니티형 기사 | KcELECTRA | ElectraTokenizer |

## 한국어 데이터셋

| 이름 | 특징 |
|---|---|
| AI Hub 뉴스 데이터 | 정치/경제/사회 카테고리 뉴스 다수 |
| KLUE | 한국어 언어 이해 벤치마크 (문장 분류 포함) |
| CROWN dataset (news+community) | 뉴스+커뮤니티 대응 가짜뉴스 |

## 실현 가능성 VS 프로젝트 목표

| | 실현 가능성 있는 모델 | 목표에 부합하는 모델 |
|---|---|---|
| 모델 | KoBERT / KoELECTRA | KLUE-RoBERTa / KoDeBERTa |
| 문맥 이해력 | 중하(기본적인 문맥 처리) | 중상~상(뉘앙스·관계 추론까지 가능) |
| 난이도 | 쉬움 (BERT 기반, 코드 많음) | 중상 ~ 어려움 (구조 복잡, 학습 느림) |
| 학습 속도 | 빠름 (Colab/Streamlit OK) | 느림 (GPU 필요, 서버 부하↑) |
| 메모리 사용량 | 낮음 (8~12GB GPU 가능) | 높음 (최소 12~16GB GPU 추천) |
| 데이터 전처리 | 비교적 간단 (뉴스 본문 토큰화) | 토큰화는 유사하지만 fine-tuning 정교함 필요 |
| 성능 기대치 | ★★★☆☆ (F1: 80~85%) | ★★★★☆ ~ ★★★★★ (F1: 87~92%) |
| 리소스/예제 | 많음 (GitHub/HuggingFace) | 적당히 있음 (논문 수준, 발표도 가능) |
| 장점 | 실용성 강조, 서비스화 용이 | 기술적 차별성 강조, 문맥 기반 분석 어필 |
| 현실성 | 시간과 달성 가능에 맞춤 | 목표에 부합, 도전적 & 고급 |

# 생각해둔 모델

1. KoELECTRA (조금 어려움)
한 줄 정리: 빠르고 효율적인 모델을 사용하여 실시간 추론에 강점을 둔 모델로, 문맥 기반 판단을 잘 처리할 수 있음
토크나이저: ElectraTokenizer
장점:
-효율성이 뛰어나서 BERT보다 빠르게 학습하고 추론할 수 있음. 실시간 뉴스 판별에 적합
-ELECTRA의 Generator-Discriminator 구조가 문맥을 파악하고 더 정확한 판단을 내릴 수 있게 도와줌
-Hugging Face에서 모델을 손쉽게 로드하고 학습시킬 수 있음
-RoBERTa보다 성능은 다소 떨어질 수 있지만, 속도와 효율성을 고려할 때 매우 실용적
단점:
-RoBERTa보다는 문맥 이해 능력이 조금 떨어질 수 있음 (특히 복잡한 논리 관계나 긴 문장에서 성능이 조금 낮을 수 있음)
-RoBERTa보다는 적지만 여전히 GPU 8~12GB 이상 필요

2. KLUE-RoBERTa (어려움)
토크나이저: AutoTokenizer
한 줄 정리: 문맥 이해 능력이 뛰어나고, 정확한 가짜뉴스 판별에 유리함(문맥과 논리적 관계 분석에 탁월함)
장점:
-BERT보다 개선된 마스킹 방식으로 문맥 파악 능력이 뛰어남. 뉴스 기사나 긴 문장에서 더 나은 결과를 도출
-최신 한국어 모델 중 하나로, 발표 시 기술적 차별화 가능
단점:
-BERT보다 큰 모델이라 학습에 시간이 오래 걸릴 수 있음(GPU 리소스 필요)
-12GB 이상의 GPU 메모리가 필요할 수 있음. 추론 시에도 메모리 소비 많음
-실시간 추론은 느릴 수 있음

3. KoBERT(쉬움)
토크나이저: BertTokenizer
한 줄 정리: 간단한 구현과 빠른 테스트가 중요하다면 적합한 모델. 학습 속도가 빠르고 기본적인 문맥 이해가 가능
장점:
-비교적 빠른 학습과 적은 메모리 사용. Colab에서도 실행 가능
-기본적인 문맥을 파악할 수 있으며, 한국어 뉴스 텍스트에서는 좋은 성능을 보임
-BERT 기반이므로 구현하기 쉬움. 실습 코드나 예제 자료가 풍부해서 학습에 용이
-적당한 성능을 보여주며, 학습과 테스트가 비교적 빨라 빠른 프로토타이핑이 가능
단점:
-RoBERTa나 KoELECTRA에 비해 문맥 이해 능력은 다소 떨어짐. 긴 문장에서의 흐름 파악이 약할 수 있음