공공데이터를 활용한 KoBERT 파인튜닝과 한국어 키워드 분석 및 대시보드 시각화

팀명: 토마토주스

202002164 박준혁(불어불문학과)

201913528 이차현(분자생물학과)

202055588 임성표(정보컴퓨터공학부)

지도교수: 조준수

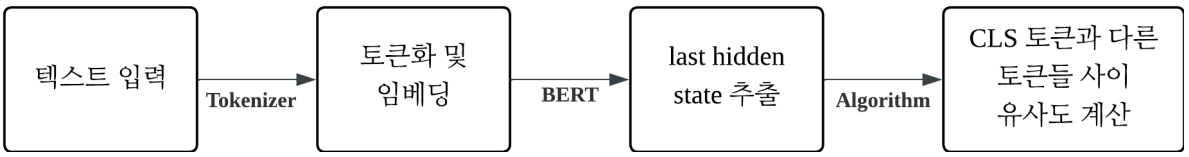
목차

1. 과제 배경 및 목적

1.1 과제 배경

자연어 처리 기술의 발전으로 키워드 추출은 텍스트 마이닝 분야에서 점점 더 중요한 역할을 차지하고 있다. 키워드 추출 기법은 방대한 텍스트를 효율적으로 분석할 수 있게 하여, 예를 들어 리뷰 분석과 같은 분야에서 분석 시간을 단축시키는 데 유용하게 활용되고 있다.

그러나 현재 사용되는 오픈소스 기반의 키워드 추출 도구들은 여러 한계를 안고 있다. 대표적인 예로, KeyBERT는 문장의 CLS 토큰과 각 토큰 벡터 간의 유사도를 기반으로 키워드를 선정하는 방식인데, 이로 인해 입력 문장이 길어질수록 성능이 저하되고 연산 시간이 급격히 증가하는 문제가 발생한다.



[그림1] 기존 KeyBERT 라이브러리 동작방식

특히 한국어처럼 형태소 단위로 의미가 구성되는 교착어의 경우, 단어의 의미가 어미 변화나 접사에 따라 달라지기 때문에 키워드를 정밀하게 추출하는 것이 더욱 어렵다. 이러한 언어적 특수성은 단순히 토큰 수준에서 유사도를 계산하는 방식으로는 충분히 반영되기 어렵다.

KeyBERT는 다양한 사전학습 언어 모델을 선택할 수 있기 때문에, 한국어를 이해하는 모델을 사용할 경우 한국어 키워드 추출도 가능하다. 그러나 이들 모델은 키워드 추출이라는 특정 목적에 맞춰 파인튜닝되어 있지 않다는 한계를 지닌다. 즉, 문장의 전반적인 의미를 파악하는 데에는 어느 정도 활용할 수 있지만, 키워드로 적합한 단어를 선별하고 추출하는 작업에는 최적화되어 있지 않아 실제 응용에서는 정확도나 일관성 면에서 한계를 보일 수 있다. 이러한 문제는 특히 어순이 자유롭고 의미가 형태소 단위로 정교하게 구성되는 한국어의 특성과 맞물려 더욱 두드러진다.

1.2 과제 목표

이러한 문제의식에서 출발한 본 프로젝트는, 한국어의 언어적 특수성과 기존 키워드 추출 모델의 한계를 극복하기 위해, 한국어에 특화된 사전학습 언어모델인 KoBERT를 기반으로 한 키워드 추출 시스템을 설계하고자 한다. 구체적으로는, KoBERT 모델을 BIO 태그(Begin-Inside-Outside) 기반의 토큰 분류 방식으로 파인튜닝(fine-tuning)하여, 문장 내에서 키워드에 해당하는 단어를 보다 정밀하게 식별할 수 있도록 한다. 이때 단순한 토큰 분류를 넘어, 조건부 무작위장(Conditional Random Field, CRF) 레이어를 추가함으로써 태깅의 연속성과 문맥적 일관성을 강화하고, 키워드 경계 구분에 있어 불안정한 예측을 최소화한다.

나아가 본 프로젝트는 단순한 키워드 추출 모델 개발에 그치지 않고, 최종 사용자가 보다 직관적이고 효율적으로 텍스트 분석 결과를 활용할 수 있도록 하는 대시보드형 웹 서비스 구현을 병행한다. 사용자는 본 시스템에 분석을 원하는 텍스트를 입력하거나, 이전에 분석한 텍스트를 불러와 확인할 수 있으며, 시스템은 입력된 텍스트로부터 추출된 키워드를 기반으로 다양한 시각적 분석 결과를 대시보드 상에 제공한다. 예컨대, 키워드 간 의미적 유사도를 시각화하기 위해 SVD(Singular Value Decomposition)와 같은 차원 축소 기법을 통해 각 키워드의 임베딩 위치를 2차원 공간에 시각적으로 배치하거나, 키워드 주변의 문맥 단어들을 attention 기반으로 가중치 시각화하여 사용자가 특정 키워드가 선택된 근거를 직관적으로 파악할 수 있도록 한다. 이러한 시각화 기능은 단순히 결과를 보여주는 것을 넘어, 시스템의 작동 방식에 대한 투명성을 제공하고, 궁극적으로는 사용자의 분석적 사고와 신뢰도를 높이는 데 기여한다.

2. 요구사항 분석

2.1 기능적 요구사항

(1) 한국어 텍스트 입력 및 전처리
사용자가 웹 대시보드에서 한국어 텍스트를 입력하면, 해당 텍스트에 대해 문장 분리, 토큰화, 특수 문자 제거 등의 전처리를 수행하며, 모델 입력 형식에 맞게 변환한다. 이를 통해 일관된 키워드 추출 결과를 보장하고 모델 성능 저하를 방지한다.

(2) 키워드 추출을 위한 KoBERT 기반 모델 구성
한국어에 특화된 사전학습 언어모델인 KoBERT를 활용하여 입력 텍스트에 대한 토큰 분류 작업을 수행하고, 각 토큰에 대해 BIO 태그를 예측한다. 이를 기반으로 키워드 후보를 식별하며, 기존 KeyBERT의 CLS 기반 방식보다 문맥을 더욱 정밀하게 반영한다.

(3) CRF 레이어를 통한 연속 키워드 안정화
모델 출력에 Conditional Random Field(CRF) 레이어를 결합하여, 의미적으로 연속된 키워드가 단절되지 않고 하나의 키워드로 인식되도록 한다. 이를 통해 어순이 자유로운 한국어에서의 키워드 단위 추출 오류를 줄인다.

(4) BIO 태그 기반 후처리 및 키워드 정제
모델의 BIO 태그 예측 결과를 기반으로, 중복 제거, 정렬, 불필요한 단어 제거 등의 후처리 작업을 수행한다. 이 과정을 통해 사용자에게 제공할 키워드 리스트의 품질을 향상시킨다.

(5) 시각적 텍스트 분석 대시보드 구현
사용자가 웹 인터페이스를 통해 분석 결과를 직관적으로 이해할 수 있도록, 추출된 키워드를 시각적으로 표시한다. 주요 키워드 하이라이팅, 키워드 빈도 그래프, 키워드 간 관계 시각화 기능 등을 포함한 대시보드를 구현한다.

(6) 데이터셋 기반 KoBERT 파인튜닝 기능
AI Hub 등에서 제공하는 공개 한국어 데이터셋을 활용하여 KoBERT 모델을 키워드 추출 목적에 맞게 fine-tuning할 수 있어야 한다. 학습 데이터는 BIO 태그 포맷으로 구성되며, 모델 구조에 맞는 학습 로직을 포함한다.

(7) 모듈화된 구조를 통한 유연한 확장성 확보
모델 구조, 데이터 전처리, 시각화 모듈 등을 독립적으로 설계하여, 다른 언어모델로의 교체, 기능 추가 및 성능 개선 작업이 용이하도록 한다. 또한, 향후 멀티모달 확장이나 다양한 NLP 태스크로의 응용을 고려한 구조를 갖춘다.

2.2 비기능적 요구사항

(1) 성능

키워드 추출 시스템은 입력된 텍스트에 대해 실시간으로 결과를 제공할 수 있어야 한다. 시스템의 응답 시간은 5초 이내로 제한하며, 사용자의 입력 텍스트 길이에 관계없이 일관된 성능을 유지해야 한다. 단, 사용자의 최대 입력 텍스트 길이는 모델 최대 입력 토큰인 512 토큰이다. 대시보드 상에서 키워드 분석 결과가 빠르게 로드되고 시각화되어야 하며, 대시보드의 사용자 경험을 방해하지 않도록 최적화된 성능을 보장해야 한다.

(2) 확장성

모델 파인튜닝이나 데이터 전처리 모듈이 유연하게 확장 가능해야 하며, 새로운 학습 데이터가 추가되더라도 기존 시스템에 영향을 최소화하도록 설계되어야 한다.

(3) 사용성

웹 대시보드는 직관적이고 사용자 친화적인 인터페이스를 제공해야 한다. 사용자는 복잡한 설정 없이 간단히 텍스트를 입력하고 키워드를 추출할 수 있어야 한다. 다양한 데이터 시각화 기능을 제공하여 사용자가 텍스트 분석 결과를 직관적으로 이해할 수 있도록 해야 하며, 시각화 옵션을 사용자 맞춤형으로 제공할 수 있어야 한다.

(4) 유지보수성

시스템은 향후 버그 수정, 기능 개선, 업데이트가 용이하도록 유지보수성 높은 코드로 작성되어야 한다. 코드 및 시스템 문서화가 잘 되어 있어 개발자들이 시스템을 유지보수하거나 확장할 때 용이해야 한다. 모델과 시스템 업데이트 시, 테스트 및 검증 과정이 명확히 정의되어 있어야 하며, 기존 기능에 대한 영향을 최소화하도록 해야 한다.

3. 개발 환경 및 사용 기술

3.1 개발 환경

(1) 개발 도구: VScode,
(2) 사용 언어: Python, TypeScript, JavaScript, SQL

3.2 사용 기술

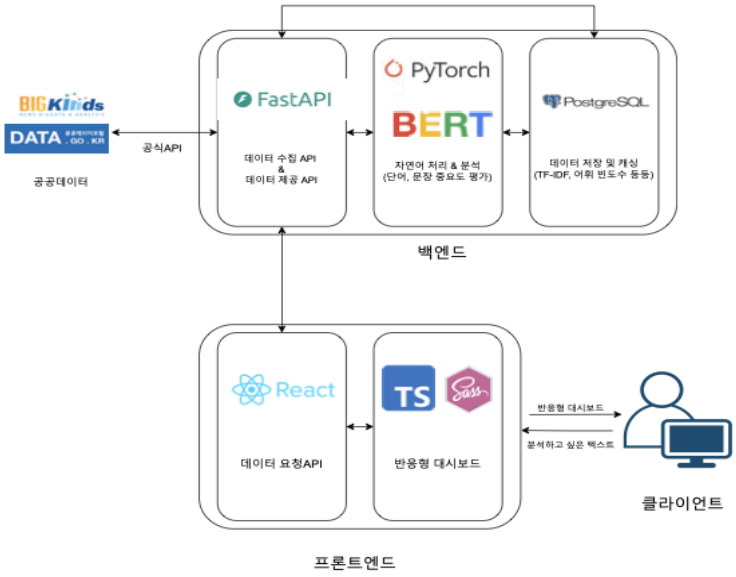
(1) 프레임워크: FastAPI, React
(2) 데이터베이스: PostgreSQL
(3) 중요 라이브러리: PyTorch, TorchCRF, Transformers, Chart.js
(4) 클라우드: AWS RDS, EC2
(5) 배포: Docker

3.3 사용 모델

KoBERT는 본 프로젝트의 핵심 엔진으로서, 한국어 텍스트의 특성을 깊이 이해하고 정밀한 키워드 추출을 수행하는 데 최적화된 사전 학습 언어 모델이다. BERT 모델 구조를 기반으로 설계된 KoBERT는 대규모 한국어 말뭉치를 학습하며, 어미나 접사와 같은 형태소 변화에 민감하게 반응하는 한국어의 교착어적 특징을 효과적으로 포착한다. 이는 단순히 토큰 단위의 유사도에 의존하는 기존 방식의 한계를 극복하고, 문맥 속에서 단어의 의미를 정확하게 파악하여 키워드 추출 성능을 극대화할 수 있는 기반이 된다. KoBERT의 양방향 문맥 이해 능력은 한국어의 복잡하고 유연한 문장 구조 내에서 키워드를 정확하게 식별하는 데 중요한 역할을 한다. 또한, Transformer 아키텍처를 채택하여 긴 텍스트에서도 효율적인 정보 처리가 가능하며, 텍스트 분류, 개체명 인식 등 다양한 자연어 처리 task에 적용될 수 있는 뛰어난 확장성을 제공한다. 특히, 널리 활용되는 Hugging Face Transformers 라이브러리와의 높은 호환성은 본 프로젝트에서 KoBERT를 손쉽게 통합하고, 다양한 관련 도구 및 기술과의 연동을 용이하게 한다. 이러한 장점들은 KoBERT를 한국어 키워드 추출 시스템 구축에 가장 적합한 선택으로 만들어 준다.

4. 시스템 설계

본 시스템은 사용자가 입력한 한국어 텍스트로부터 핵심 키워드를 추출하고, 이를 다양한 방식으로 시각화하여 제공하는 것을 목표로 한다. 전체적인 시스템 구조는 크게 프론트엔드, 백엔드, 그리고 자연어 처리 부분으로 나누어 설명할 수 있다.



[그림2] 시스템 구조도

4.1 프론트엔드

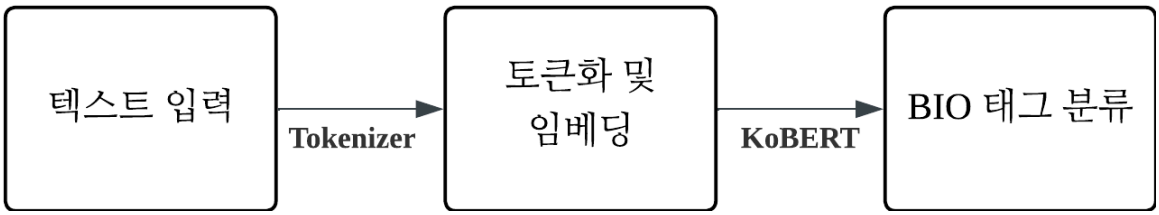
프론트엔드는 사용자와 직접 상호작용하는 부분을 담당한다. React를 주요 기술로 활용하여 사용자 인터페이스를 구축하고 사용자 경험을 향상시키는 데 초점을 맞춘다. 사용자는 웹 브라우저를 통해 시스템에 접근하여 분석을 원하는 텍스트를 입력할 수 있으며, 분석 결과는 시각화된 형태로 제공되어 직관적인 이해를 돕는다. TypeScript는 React 컴포넌트 개발 시 코드의 안정성과 유지보수성을 높이는 역할을 수행한다. 추출된 키워드의 빈도나 키워드 간의 관계는 Chart.js를 통해 다양한 그래프 형태로 시각화되어 사용자에게 제공된다. 프론트엔드는 백엔드가 제공하는 데이터 시각화 API를 통해 분석 결과를 받아와 화면에 효과적으로 표현한다.

4.2 백엔드

백엔드는 프론트엔드로부터의 요청을 처리하고, 핵심적인 데이터 처리 및 저장 기능을 수행한다. FastAPI 프레임워크를 사용하여 API 서버를 구축하며, 프론트엔드에 분석 결과를 제공하는 역할을 한다. PostgreSQL 데이터베이스는 추출된 키워드, 원문 텍스트, 분석 결과 등의 데이터를 효율적으로 저장하고 관리하는 데 사용된다. 백엔드는 자연어 처리 결과를 바탕으로 데이터를 저장하고, 프론트엔드의 요청에 따라 필요한 데이터를 조회하여 응답한다.

4.3 자연어 처리

자연어 처리 부분은 입력된 한국어 텍스트에서 의미 있는 키워드를 추출하는 핵심적인 역할을 수행한다. 한국어 특화 사전 학습 모델인 KoBERT는 PyTorch 환경에서 구현 및 실행된다. KoBERT 모델은 입력 텍스트에 대한 토큰 분류(BIO 태깅)와 CRF 레이어 적용을 통해 정밀한 키워드 추출 작업을 진행한다.



[그림3] 새로운 키워드 추출 방식

백엔드에서 FastAPI를 통해 텍스트 분석 요청을 받으면, KoBERT 모델을 활용하여 키워드를 추출하고, 필요에 따라 단어나 문장의 중요도를 평가한다. 추출된 키워드는 백엔드를 통해 PostgreSQL 데이터베이스에 저장되고, 프론트엔드에 시각화 형태로 제공된다. 키워드 간의 의미적 관계를 시각화하기 위해 SVD(특이값 분해)와 같은 차원 축소 기법이 활용될 수 있으며, 특정 키워드가 선택된 근거를 시각적으로 보여주기 위해 Attention 기반 가중치 시각화가 적용될 수 있다.

이처럼 프론트엔드, 백엔드, 그리고 자연어 처리 각 부분은 명확한 역할을 분담하여 시스템의 효율적인 작동을 가능하게 한다. 한국어 텍스트 분석에 특화된 KoBERT 모델을 중심으로 정확하고 효율적인 키워드 추출 기능을 제공하며, 사용자 친화적인 인터페이스와 다양한 시각화 기능을 통해 텍스트 분석 경험을 향상시키는 것을 목표로 한다. 또한, 각 기능별 모듈화 설계를 통해 향후 기능 확장 및 유지보수 용이성을 확보하고자 한다.

5. 현실적 제약 사항 분석 결과 및 대책

5.1 현실적 제약 사항

언어 모델이 학습하는 데이터의 확보 및 활용은 모델의 성능과 직결되는 중요한 문제인 동시에, 법적인 측면에서 신중한 고려가 요구되는 부분이다.

최근 인공지능 모델 학습에 사용되는 데이터셋의 저작권 문제가 사회적으로 중요한 이슈로 부각되고 있다. 방대한 양의 텍스트 데이터를 학습하여 언어의 패턴과 의미를 이해하는 언어 모델의 특성상, 학습 데이터에 저작권이 있는 콘텐츠가 포함될 경우 법적인 분쟁의 소지가 발생할 수 있다. 따라서 본 프로젝트에서는 이러한 저작권 문제를 사전에 방지하고, 안전하고 합법적인 범위 내에서 모델을 학습시키는 것을 최우선 과제로 설정한다.

이러한 배경 하에, 초기 데이터 확보 방안으로 고려했던 빅카인즈(BIG Kinds) 데이터셋은 다양한 언론사의 기사를 포함하고 있어, 명확한 저작권 문제를 회피하기 어렵다는 판단을 내렸다. 따라서 본 프로젝트에서는 저작권으로부터 자유롭거나, 명확한 사용 허가를 받은 데이터만을 학습에 활용하는 것으로 계획을 수정한다.

5.2 대책

이에 대한 해결책으로, AI Hub에서 정식으로 인가받아 제공받은 데이터셋만을 학습 데이터로 활용할 예정이다. AI Hub는 정부 차원에서 구축하고 관리하는 데이터 플랫폼으로, 저작권 문제가 없는 양질의 한국어 텍스트 데이터를 제공받을 수 있다는 장점을 가진다. 그러나 AI Hub에서 제공하는 데이터셋 역시 데이터의 성격 및 제공 정책에 따라 모델이 특정 도메인에 특화되거나, 다양한 장르의 텍스트를 포괄적으로 학습하는 데 제약이 발생할 수 있다는 점을 인지하고 있다. 즉, 학습 데이터의 범위와 특성에 따라 모델이 뛰어난 성능을 발휘할 수 있는 특정 도메인이 제한될 수 있다는 현실적인 한계에 직면하게 된다. 한편, AI Hub의 데이터 제공 정책상, 본 프로젝트에서 학습에 사용된 데이터셋의 상세 내용은 외부에 공개할 수 없음을 명확히 밝힌다.

6. 개발 일정 및 역할분담

6.1 개발 일정

	데이터 확보 및 정제	모델 개발	모델 기능 확장 연구	모델 기능 확장 구현	데이터베이스 설계	백엔드 API 개발	프론트엔드 UI/UX 개발	프론트엔드 API 개발	서비스 통합	서비스 배포
1주차										
2주차										
3주차										
4주차										
5주차										
6주차										
7주차										
8주차										
9주차										
10주차										
11주차										
12주차										

6.2 역할분담

박준혁	자연어 처리 담당으로서 AI 허브에서 필요한 데이터를 수집하고, 수집된 데이터의 토큰 길이에 따른 균등 분포를 고려하여 전처리 작업을 진행한다. BERT 모델을 기반으로 텍스트 분류 성능을 향상시키기 위해 선형 레이어를 추가하고, 개체명 인식 등 BIO 태깅 task를 위해 CRF 레이어를 추가하여 인공지능 모델을 학습한다. 학습된 모델의 성능을 다양한 지표를 사용하여 객관적으로 평가하고, 더 나아가 기능 확장을 위해 관련 분야의 논문을 심층적으로 연구하고 이를 실제 구현에 적용한다.
이차현	프론트엔드 개발을 담당하여 사용자 인터페이스(UI) 및 사용자 경험(UX) 향상을 위한 API를 구현하고, 데이터 시각화를 위한 대시보드를 개발한다. 특히, 텍스트 데이터 분석 결과를 효과적으로 시각화하기 위해 워드클라우드 및 다양한 형태의 플롯 차트를 구현하여 사용자가 입력한 텍스트의 의미를 직관적으로 파악할 수 있도록 지원한다.
임성표	백엔드 개발을 담당하여 데이터 저장 및 관리를 위한 데이터베이스를 설계 및 구현하고, 프론트엔드와의 효율적인 데이터 교환을 위한 API를 개발한다. 또한, 워드클라우드 생성을 위해 텍스트 데이터에서 명사, 형용사, 동사 등 품사별로 형태소를 추출하는 기능을 구현하여 데이터 분석의 기초를 마련한다.