

updated 2024.10.31

a journey from social science to machine learning practitioner

Portfolio

Data Scientist
Shin Jaesol

ysys143@naver.com

010-7196-8188

for more information, visit my blog socml.tistory.com and github <https://github.com/ysys143>

About Me

🌱 저는 '성장에 미친 사람', 신재술입니다. 저는 사회학 석사 과정을 거쳐 다양한 사회적 문제를 데이터로 분석해 오면서, 데이터 분석을 넘어 인공지능을 통한 실질적인 문제 해결에 도전하게 되었습니다. 출발이 조금 늦었지만, 그만큼 성장 곡선은 누구보다 가파릅니다. 절편값은 작을지 몰라도, 제 기울기와 가속도는 대단히 큼니다. 그리고 그 배경에는 아이러니 하지만 얼핏 종잡을 수 없는 것처럼 보이는 제 인생궤적이 있습니다.

나무를 태우면 수분과 휘발성 물질이 먼저 사라지고 탄소만 남게 됩니다. 숯은 연소에 불필요한 물질들이 날아가고 탄소만 남은 상태이므로 더 불에 잘타는 아이러니를 지니고 있습니다. 지금 저는 시행착오를 통해 불필요한 요소들은 날아가고, 더욱 강렬하게 집중할 수 있는 상태가 되었습니다. 연소 후 남은 숯처럼 더 순수하고 강한 열정과 학습 능력으로 빠르게 성장해 왔습니다.

📖 NLP, 프롬프팅, RAG, 감성분석, 소셜리스닝

✉️ ysys143@naver.com

🎓 리츠메이칸대학 사회학연구과 현대사회전공 석사과정 졸업

[국가자격]

빅데이터 분석기사(한국데이터산업진흥원)

SQL개발자(한국데이터산업진흥원)



My Stack

python, pandas, numpy, scikit-learn, git, duckDB, R, MySQL, Keras, PyTorch, PyQt, gradio, streamlit, konlpy, huggingface, langchain, openai, gemini, gensim, faiss, llama-index, BeautifulSoup, Selenium, pyppeteer, PyMuPDF, pdfplumber, pytesseract, tabula-py, reportlab

My Strength

저의 강점은 빠른 학습과 적응력, 그리고 실질적인 문제 해결을 위한 몰입입니다. 비전공자로 시작했음에도 단기간 내에 ML/AI 개발 역량을 쌓았고, 실제 보험가입예측, 한국은행 RAG 챗봇, 유튜브 댓글 분석기 등 모든 프로젝트에서 목표 및 아이디어를 제시하고, 개발과정에서도 주도적인 역할을 해왔습니다. 저는 다양한 도메인에서 쌓은 경험을 바탕으로 문제를 파악하고 최적의 AI 솔루션을 제안하고자 합니다.

My Goal

데이터와 머신러닝을 통해 최고의 의사결정을 만들어낼 수 있는 사람,
나아가 최고의 프로덕트로 완성시키는 사람.

Trajectory

2024.9-12 가짜연구소 All-you-need-is-RAG 러너
2024.9-12 신촌문화정치연구그룹 '비판적 AI읽기 세미나' 참여
2024.9-10 NIPA-Google ML 부트캠프 협력 실무 프로젝트(3기) 수료
2024.10 Google Cloud Transformer Models and BERT Model
2024.10 Tableau 신병훈련소(bootcamp) 25기
2024.10 데이터리안 SQL 데이터 분석 캠프 실전반 34기 수료
2024.9 데이터리안 SQL 데이터 분석 캠프 입문반 33기 수료
2024.7-10 Google Machine Learning Bootcamp 5기 수료
2024.8 Deep Learning Specialization Certification 취득

2023.12 빅데이터분석기사 취득
2023.10 SQL개발자 취득
2018.8 KOSSDA 웹데이터를 활용한 빅데이터 분석 수료

2022-2023 군산 미공군기지에서 군복무
2019-2021 리츠메이칸대학 사회학연구과 석사과정 졸업
2020 투명사회를 위한 정보공개센터 활동가
2018-2020 중앙대학교 대학원 사회학과 석사과정 수료
2012-2017 한국예술종합학교 영상이론과 졸업
2012 한국과학영재학교 졸업

2024 한일차세대학술포럼 제21회 국제학술대회 발표
2019 비판사회학회 하계워크숍 발표
2010 한국산업응용수학회 포스터 발표

Project1: BOK Reports Analyzer

한국은행 보고서 기반 Q&A 챗봇

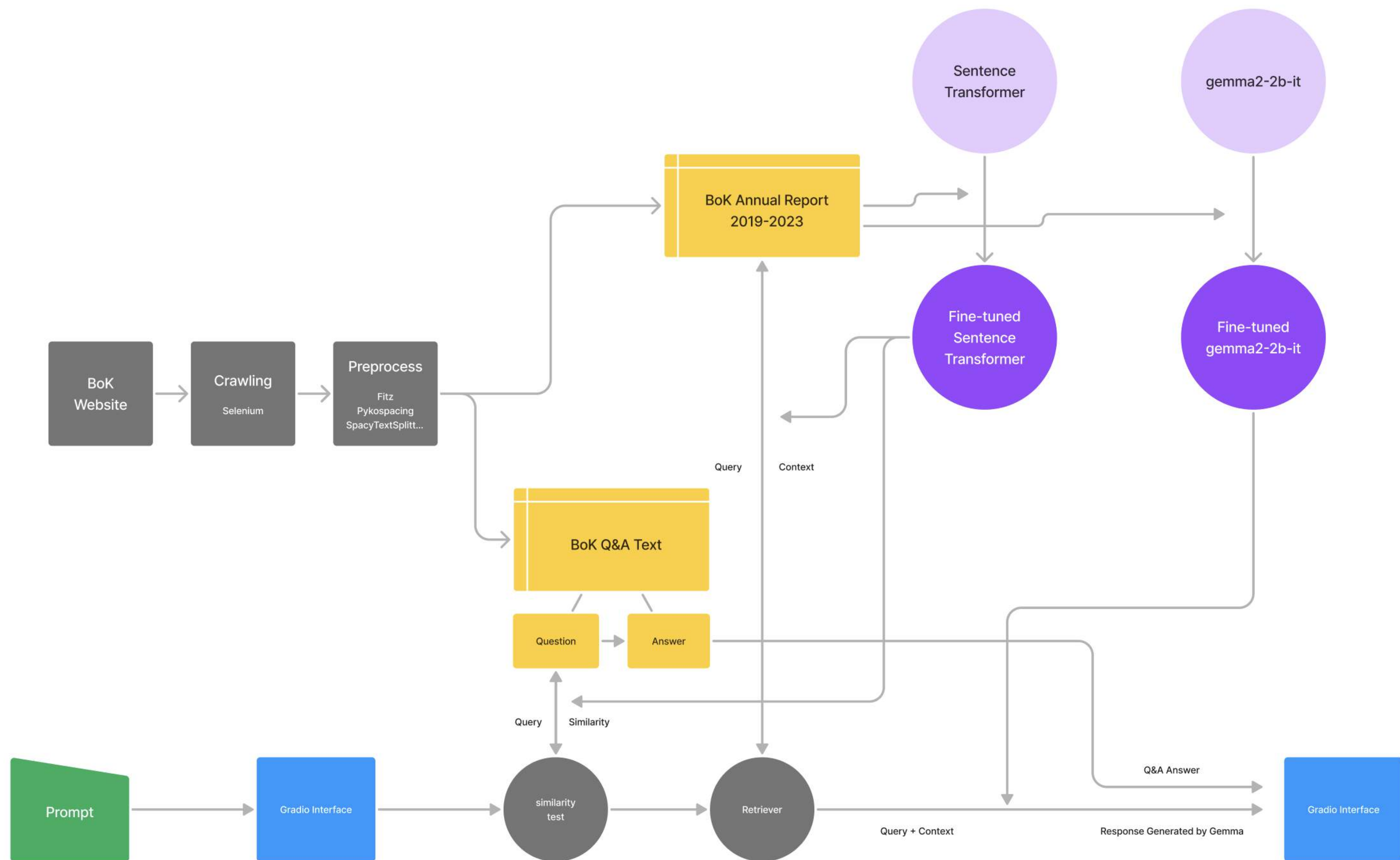
문제 정의

한국은행(BOK)의 발표와 보고서는 금융권 및 경제 전반에 중요한 영향을 미치며, 정책 결정자와 경제 참여자에게 유의미한 정보를 제공합니다. 그러나 방대한 양과 복잡성으로 인해 필요한 정보를 신속하게 찾기 어렵다는 문제가 있었습니다. 이를 해결하고자 한국은행의 연간 보고서와 Q&A 자료를 기반으로 정책 및 연구 내용을 신속하게 파악할 수 있는 Q&A 챗봇 시스템을 구축했습니다. 이 챗봇은 정책 분석 및 의사결정에 도움을 주며, 사용자에게 신뢰성 있는 답변을 제공하고자 했습니다.

해결 방법

2019-2023년 한국은행 보고서와 Q&A 자료를 기반으로 하여 데이터 수집, 전처리, 임베딩 및 LLM 모델 파인튜닝을 통해 Q&A와 RAG(Retrieval-Augmented Generation) 기반의 답변 생성 시스템을 구현했습니다. 데이터 수집은 Selenium을 활용하여 한국은행 웹사이트에서 보고서를 일괄적으로 수집하였고, PDF 텍스트 추출은 PyMuPDF를 활용했습니다. 전처리 과정에서는 gpt-4o-mini 모델을 few-shot prompt 방식으로 활용하여 표 형식 데이터를 제거하고, 필요 없는 형식 요소를 정제했습니다. 데이터는 Langchain의 SpacyTextSplitter를 통해 분할하여 일관성을 유지했습니다.

Q&A 자료에 대해서는 질문문장의 임베딩을 통해 Q&A 챗봇을 구현했습니다. 사용자의 질문에 대한 임베딩을 생성하고, 유사도가 높은 경우에는 Q&A 자료에서 명확한 답변을 직접 제공하고, 그렇지 않은 경우에는 유사도가 높은 문장을 컨텍스트로 설정해 파인튜닝한 Gemma 모델이 답변을 생성하도록 했습니다. 임베딩 모델로는 Huggingface의 Sentence Transformer를 사용했으며, 문장 분할과 정보 검색에서 답변 품질이 미흡하다는 점을 개선하기 위해 retriever(임베딩 모델)와 llm(generation) 모두를 파인튜닝했습니다. 최종 인터페이스는 Gradio를 통해 사용자가 쉽게 접근할 수 있도록 설계했습니다.



BOK Reports Analyzer의 system architecture

성과

BOK Reports Analyzer는 한국은행 자료에서 필요한 정보를 신속히 찾을 수 있는 기반을 마련하여 정책 담당자와 경제 참여자들이 원하는 정보를 빠르게 제공할 수 있도록 했습니다. 다만, 표와 이미지 형식의 데이터를 적절히 활용하지 못했고, 지식 그래프를 비롯해 보다 정교한 RAG 기법을 시도해보지 못했습니다. 그 결과 답변 품질이 만족할 만한 수준에 미치지 못한 한계가 있었습니다. 그러나 이 프로젝트를 통해 향후 더 본격적인 RAG 프로젝트에 대한 기반을 마련했습니다.

한계와 발전 방향

현재 챗봇 시스템은 답변의 정확성 및 환각(허위 답변) 제어에 미흡한 점이 있으며, 표 형식 데이터를 효율적으로 처리하지 못한 한계가 있습니다. 앞으로는 더 강력한 RAG 시스템을 통해 환각 현상을 제어하고, 표 형식 데이터를 자동으로 파싱하는 기능을 추가할 계획입니다. 실시간 데이터 업데이트를 지원하기 위한 데이터 수집 자동화 기능을 강화하고, 다양한 질의응답 구조에 대응하는 멀티태스크 학습과 지식 그래프 구축을 통해 정보의 연관성을 강화하여 사용자에게 더욱 신뢰할 수 있는 챗봇 시스템을 제공할 것입니다.

Demo Output

Chatbot

물가안정이 중앙은행에게 중요한 목표인 이유는 무엇인가?

Q&A에 등재된 내용을 기반으로 답변을 구합니다.

물가안정목표제(Inflation targeting)는 중앙은행이 명시적인 중간목표 없이 일정기간 또는 중장기적으로 달성해야 할 물가상승률 목표치를 미리 제시하고 이에 맞추어 통화정책을 수행하는 방식입니다. 즉, 중앙은행은 통화량, 금리, 환율 등 다양한 정보변수를 활용하여 장래의 인플레이션을 예측하고 실제 물가상승률이 목표치에 수렴할 수 있도록 통화정책을 수행하며 이후 그 성과를 평가하고 시장의 기대와 반응을 반영하면서 정책방향을 수정해 나갑니다. 이 제도는 1990년 뉴질랜드에서 처음 도입되었으며 이후 캐나다, 영국, 스웨덴 등 선진국과 한국, 멕시코 등 신흥시장국, 그리고 체코, 폴란드 등 체제전환국까지 확산되어 운영되고 있습니다.

물가안정목표제는 대체로 다음과 같은 요소를 포함합니다. 첫째, 물가안정목표(target)의 구체적인 수치를 공표합니다. 둘째, 통화정책의 목표가 여러 개 있을 수 있지만 물가안정이 가장 중요한 목표여야 합니다. 일부에서는 물가안정목표제가 오직 물가안정만을 추구하는 것으로 이해하는 경우도 있으나, 일반적으로 물가안정과 여타 목표를 함께 추구하는 경우도 물가안정목표제에 포함합니다. 셋째는 책임성입니다. 이는 수치로 제시된 목표를 달성하는지 여부를 점검하고 그에 대한 일정한 책임을 부담함으로써 목표달성의 신뢰성을 높이는 것입니다. 책임성을 강화하기 위해 흔히 거론되는 요소 중의 하나가 투명성입니다. 물가안정목표를 국민들에게 명시적으로 알리고 평가를 받음으로써 책임성이 제고될 수 있습니다. 이러한 책임성과 투명성은 결국 경제주체에게 통화정책에 대한 신뢰, 즉 주어진 목표를 달성할 것이라는 기대를 심어줌으로써 통화정책의 유효성을 제고하게 됩니다.

한국은행은 1998년부터 물가안정목표제를 도입 운영하고 있으며 2016~2018년까지의 물가안정목표는 소비자물가 상승률 기준 2%로 설정하였습니다. 물가안정목표 운영 상황에 대한 정기적인 점검·설명은 국회 제출 법정보고서인『통화신용정책보고서』를 통해 연 4회 실시하고 있습니다. 정기적 설명책임 이외에 소비자물가 상승률이 6개월 연속 물가안정목표를 ±0.5%포인트 초과 또는 하회할 경우 총재의 기자회견담화 등을 통해 초과 원인, 물가 전망경로, 통화정책방향 등을 설명하고 있으며 이후에도 이러한 상황이 지속되는 경우 3개월마다 후속 설명책임을 이행하고 있습니다.

질문 입력

대화 기록 초기화

1) Q&A시스템에 의해 답변이 출력된 경우 :
성공 - 속도가 빠르고 자세한 답변이 제공된다.
한계 - 질문과 맞지 않는 답변을 제공할 때도 많음.

Chatbot

2023년 통화정책 운용에서 가장 중요한 목표는?

소비자 물가 상승률이 물가 안정 목표에 근접하도록 통화신용정책을 운영하고 있다.

질문 입력

대화 기록 초기화

2) RAG에 의해 답변이 출력된 경우
성공 - 관련 문서를 찾아와 모델이 이를 참고하여 응답하는 과정이 잘 이루어지고 있다.
한계 - 지나치게 단답식이거나 답변을 못하고 뺏어버리는 문제.

Chatbot

중앙은행과 정부 간의 갈등적 관계에 대해

이야기해보자.

중앙은행-정부 갈등, 역사와 현황 분석

1. 역사:

- 한국 전통에서의 차이: 한국 사회는 특히 근대 이후로 국가적인 경제 정책을 주도하는 정부와 중앙은행의 역할을 구분하지 않고 자유로운 경제 시장을 추구했다.
- 근본적인 갈등: 중앙은행은 금융 안전 및 통화 보존에 대한 책임을 지고 있으며, 정부는 경제 성장과 일자리 창출에 초점을 맞추는 데 중점을 두었다. 이러한 역할 수준의 차이 때문에 정부와 중앙은행 사이의 갈등이 발생하기 시작했다.
- 1987년 경제 위기를 타격: 1987년 외환 위기는 정부와 중앙은행의 조율이 부족했던 상황을 드러냈다. 이는 정부의 경쟁 우위를 위한 노력으로 인한 과잉 투자와 불안정한 환경 등의 문제점을 반영했다.
- 2008년 금융 위기에 직면: 2008년 금융 위기는 정부와 중앙은행의 협력이 중요성을 더욱 강조했다. 극심한 불평등과 복잡한 경제 상황 속에서 정부와 중앙은행의 동일한 의식과 협력이 필수적이라는 것을 보여주었다.

2. 현재 상황:

- 경제 활동 강조: 현재 중앙은행은 경제 활동 증진과 저금리 유지를 위해 노력하고 있다. 하지만 정부는 경제 성장을 위한 다양한 정책을 제시하고 있다.
- 감시 및 평가: 중앙은행은 정부의 정책 결과를 감시하고 평가하며 개입한다. 하지만 정부는 중앙은행의 행동을 객관적으로 판단하거나 적극적으로 반박하는 경우가 많다.

질문 입력

대화 기록 초기화

3) 파인튜닝된 모델
성공 - 비교적 한국은행이나 경제 관련 지식들을 잘 습득하고 있다.
한계 - 텍스트에 기반한 답변보다는 사전학습된 내용이 더 많이 반영되고 있다.

Contribution

- 프로젝트 아이디어를 제안했습니다. Gemma 모델 사용 요구사항을 바탕으로 명확한 목표와 데이터 수집의 범위를 설정할 수 있다고 판단했기 때문입니다.
 - RAG 기반 챗봇을 제안했으며, RAG가 높은 범용성과 관심을 받는 기술이라는 점에서 챗봇 구조를 RAG로 구현할 것을 제안했습니다. 데이터 소스로는 사회과학 배경을 바탕으로 정책 문서에 대한 머신러닝 기반 처리를 다루고 싶어 한국은행 보고서를 선택했습니다.
 - Selenium을 통해 한국은행 웹사이트에서 보고서를 일괄 수집하고, PyMuPDF와 OpenAI API를 통해 표 형식 데이터를 제거하고 텍스트 데이터를 정제하는 작업을 수행했습니다.
 - Huggingface의 Sentence Transformer를 사용한 임베딩 모델을 제안하고, 오픈소스를 통한 구현 요구에 따라 한국어로 파인튜닝된 모델을 사용했습니다.
 - 각 단계에서 답변의 품질이 미흡한 부분을 분석하여 retriever(임베딩 모델)와 llm(generation) 모두를 파인튜닝하자는 개선 방안을 제안했습니다.
 - 질문 임베딩을 생성하고, 유사도 수준에 따라 Q&A 답변과 RAG 기반 답변을 선택하도록 하는 로직을 설계하고 구현했습니다.
 - 사용자 친화적인 인터페이스를 제공하기 위해 Gradio를 활용해 챗봇 인터페이스를 설계하고 구축하여, 프로젝트의 최종 결과물이 사용자에게 직관적으로 전달될 수 있도록 했습니다.
 - 최종 보고서 작성에 필요한 시스템 구조도와 기타 산출물들을 작성했습니다. 또한 프로젝트가 온라인으로 진행되어 느슨해질 수 있는 점을 고려해 지속적으로 회의 일정을 잡고, 오프라인 미팅을 위한 장소 섭외 등을 추진했습니다.
-

Project2: TRENDPOP

KPOP Youtube MV 감정분석 대시보드

문제 정의

유튜브는 KPOP 산업에서 글로벌 팬덤을 연결하고 성장시키는 중요한 플랫폼입니다. 팬들은 댓글을 통해 아티스트와 소통하며 다양한 반응을 표현하지만, 유튜브가 제공하는 기본 지표로는 팬덤의 감정 흐름이나 주요 이슈를 깊이 있게 파악하기 어렵습니다. 이에 유튜브 댓글 데이터를 분석하여 KPOP 팬덤의 반응을 모니터링하고 인사이트를 제공하는 대시보드를 구축하고자 했습니다. 이 대시보드를 통해 아티스트의 평판 관리와 팬들과의 소통을 지원할 수 있습니다.

해결 방법

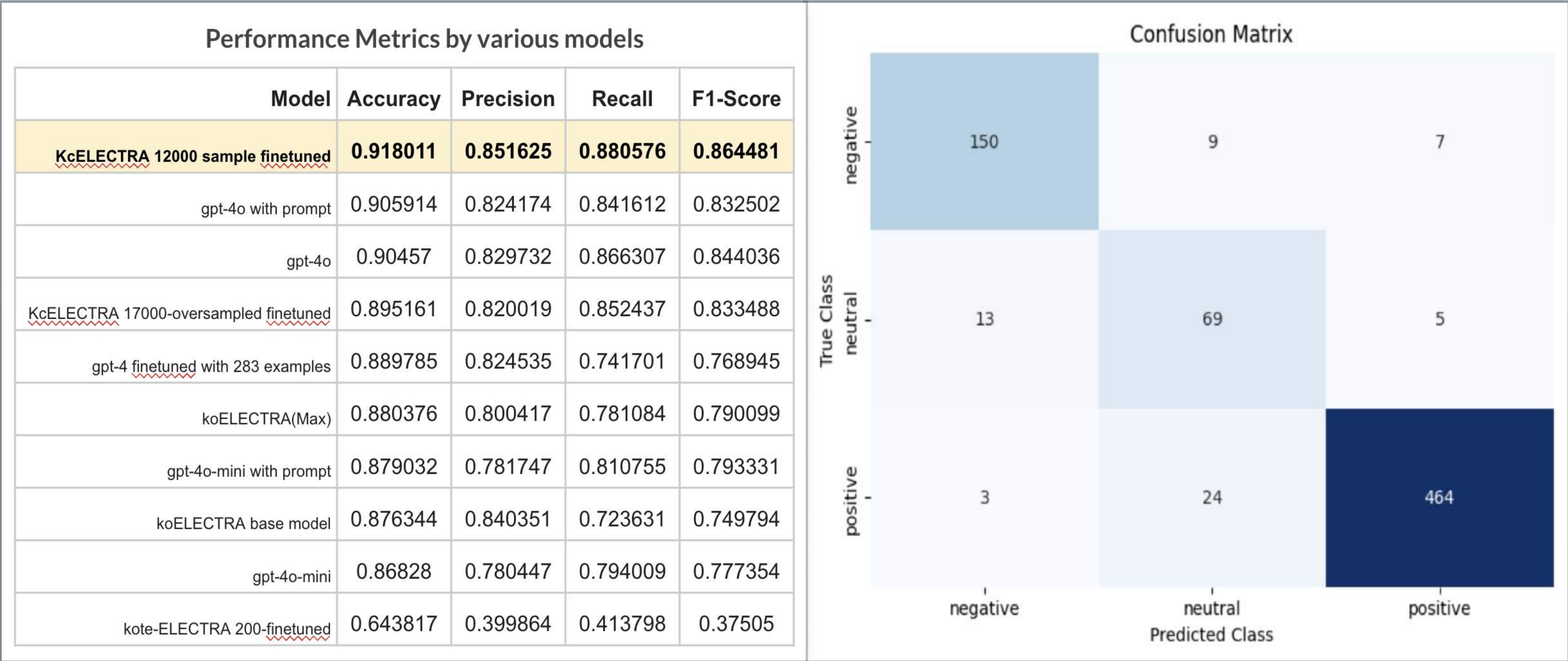
프로젝트는 4세대 KPOP 걸그룹 TOP 10의 주요 뮤직비디오 댓글을 대상으로 진행되었습니다. 데이터 수집은 Selenium과 YouTube API를 사용해 총 900만 개의 유튜브 댓글을 수집했고, 댓글의 절반 이상이 한글인 경우 이를 한국어 댓글로 간주하여 최종적으로 64만 개의 한국어 댓글을 추출했습니다. 이후 HTML 태그와 이모지 제거, 특수 문자 및 HTML 엔티티 제거, 웹 링크와 타임코드 제거 등의 전처리 과정을 거쳐 댓글을 정제했습니다. 이와 함께, 명사 추출 후 잘못 파싱된 단어를 필터링하고 불용어 및 대체어 사전을 구축하여 일관성 있는 데이터를 확보했습니다.

전처리된 데이터는 감정 분석 및 키워드 추출에 사용되었으며, Python과 PyTorch를 활용해 KcELECTRA 모델을 KOTE 데이터셋으로 파인튜닝한 모델을 적용했습니다. 이때, 직접 라벨링한 12,000개의 댓글 데이터를 학습에 활용했으며, 성능 평가는 3명의 코더가 만장일치로 분류한 733개의 댓글을 통해 진행해 신뢰도를 확보했습니다. 또한, 키워드 추출을 통해 각 그룹, 뮤직비디오, 멤버별로 중요한 키워드를 파악하고, 이를 필터링하여 특정 키워드별 팬덤 반응을 모니터링할 수 있도록 했습니다.

감정 변화가 급격히 나타난 시점에는 주요 뉴스 요약を提供하는 RAG 기반 기능을 구현하여, 팬덤 감정 변화의 원인을 파악할 수 있도록 했습니다. 구글과 네이버 뉴스 API가 과거 뉴스 구간 검색을 지원하지 않아 크롤링을 통해 뉴스 제목과 링크를 데이터베이스로 구축하고, OpenAI 임베딩 모델을 통해 관련 기사를 검색하고 요약하여 대시보드에 통합했습니다. 최종 분석 결과는 Streamlit 기반의 대시보드에서 실시간으로 확인할 수 있어, 사용자들이 팬덤의 감정 변화를 직관적으로 모니터링할 수 있도록 했습니다.

제목	연도	데이터/레이블링	접근	전처리	모델	학습/성능향상	성능평가	
			어휘기반(Lexicon), ML기반(ML), 신경망기반(NN), LLM API(Prompt)	형태소분석기, 불용어처리, 명	감정점수합산, SVM/Boostin	비지도학습/사전학습, 지도학		
2013 KOSAC - 한국어 감정 및 의견 분석 코퍼스	2013		Lexicon	형태소분석	svm, 군부정 극성, 주관적(se		10류를 교차 감정법 사용	
2019 자연어 처리 기술을 이용한 감정분석 기법에 관한 연구	2019	트위터데이터 크롤링	Lexicon	형태소분석기	비학습데이터, 기계학습 병행			
2020 기계학습을 이용한 Aspect-Based Sentiment Analysis 기반 전가차	2020		ML, NN	불필요한 텍스트 요소 제거 (U 토큰화 및 품사 태깅 표제어 추출(Lemmatization) 불용어(Stopwords) 제거 데이터 불균형 해결을 위한 오	요소 추출: TextRank, Naïve 감성분석: Logistic Regressi 임베딩 : TF-IDF, GloVe			
2024 한국어 구문분석을 활용한 의존 관계 패턴 기반의 감성사전 구축 기법	2024	소평을 후기 분석 데이터	Lexicon	KSS 문장분리기, 한 어절 의존 구문분석, 규칙 기반 접근	비지도 or 지도 X 의존 관계	F1 score		
2019 BERT 기반 한국어 감정 사전을 이용한 감정 예측기 개발	2019	온라인 댓글 및 리뷰데이터 수	NN	토큰화, 형태소분석기 - 불용어 nn사용	nn사전학습, 파인튜닝, 하이퍼	정확도, f1스코어, 정밀도, 재현율		
2023 초거대 언어 모델을 활용한 감정 분석 연구	2023	GoEmotions(레딧 댓글) K	NN, Prompt	형태소분석기, 동사 형용사 추	BERT(KcBERT, RoBERTa GPT(3.5, 4)	사전학습, Zero-shot promp	F1 score, Pearson 상관계수	
2016 한국어 감정분석 코퍼스를 활용한 양상정보 기반의 감정분석 연구	2016			표현유형 태그, seed태그주석	선형 svm,scikit-learn	확률자질 (PES, PPOSS, PS REF-TYPE-SPEECH, RE	10겹 교차검증법, precision, recall, f1, accuracy	
LLM과 앙상블 머신러닝모델을 융합한 한국인 감정분석 모델 설계	김현지	2024	AlHub 감정 대화 데이터세트 레이블 - 기쁨, 슬픔, 분노, 불	ML + NN	Konlpy.Okt 명사추출 FastText를 통해 각 '단어'를	KoBERT + RF 앙상블	FastText - 트레이닝 KoBERT - 트레이닝	Accuracy, Precision, Recall, F1-Score
2021 Building the Korean Sentiment Lexicon for Finance (KOSELF)	조수지	2021	한국어금융감성사전, 애널리스트	lexicon	토큰화, konlpy형태소분석기,		로지스틱회귀분석, 회귀분석	
리뷰 감성분석 - KoBERT, KoGPT-2, KoBART	이민아	2023	구글 플레이스토어 카카오톡/ 평점 1~2 부정, 3 요청, 4~5 ;	NN	5자 이 하 제거, 한글 자/모음 고정된 입력 크기를 유지(64 t	KoBERT, KoGPT-2, KoBA	KoBERT, KoGPT-2, KoBA 하이퍼파라미터 튜닝 - 배치사	Accuracy, Precision, Recall, F1-Score
라디오 청취자 문자 사연을 활용한 KoBERT 기반 한국어 다중 감정 분석 연구 이재아		2023	1) AlHub 감정 대화 데이터 2) 라디오 청취자 문자 사연을 NN 7가지 - '행복', '슬픔', '놀람', '기	NN	15,831문장 -> 5,531문장 비문법적인 노이즈를 가진 문? KoBERT	KoBERT	KoBERT 파인튜닝 모델1 - AlHub 데이터로 학습 모델2 - 라디오 청취자 데이터	Accuracy(라디오 청취자 test set)
[감정 분석 모델] 한국어 감정 분석 데이터셋 KOTE 논문 리뷰 / Python에서		2024	KOTE 데이터셋	ML	최소 길이는 10, 최대 길이는 : 가장 긴 텍스트의 상위 10%로 길이는 404, 평균은 57.32, 중	KoELECTRA 를 KOTE로		
뉴스 기사 제목의 감정 라벨링 기법 연구	하재룡	2022	코로나19 관련 뉴스 기사 제목 Naver 포털에서 월별로 약 2t 공정, 부정 수작업 레이블링	Lexicon, NN	관용구, 특수문자 등 제거	1) KNU Lexicon 2) KoELECTRA	모델1 - 도메인 단어 사전을 기 모델2 - KoELECTRA 파인튜 모델3 - 추가 파인튜닝(NSM 모델4 - 세가지 모델을 적절한	Accuracy, Precision, Recall, F1-Score
딥러닝을 활용한 감정 분석 과정에서 필요한 데이터 전처리 및 형태 변형	서해진, 신	2022	0) EDA 1) 소문자화(영어) 2) 토큰화 3) 축약어 4) 태그 제거 5) 이모티콘 제거 6) 외국문자 제거 7) 무의미하게 반복된 문자 제거 8) 명사추출 9) 불용어 제거 9) 너무 긴 문장 제거					

선행연구 검토를 통해 가장 성능이 좋은 kcBERT 모델을 베이스로 선택



파인튜닝된 최종 모델은 LLM을 뛰어넘는 분류 성능을 보였음

NewJeans	'Cool With You' & 'Get Up' (side B)	다니엘	양조위	노래	뉴진스	뮤비	음색
		민지	사랑해	뉴진스	해석	뮤비	민희진
		하니	뉴진스	사랑해	인간	뮤비	양조위
		해린	뉴진스	사랑해	취한	화이팅	고양이
		해인	뉴진스	목소리	사랑해	노래	표정
	'Cool With You' (side A)	다니엘	뉴진스	사랑해	양조위	요정	노래
		민지	사랑해	뉴진스	미모	민희진	노래
		하니	결혼	뉴진스	사랑해	민희진	화이팅
		해린	뉴진스	사랑해	화이팅	분명	민희진
		해인	목소리	사랑해	뉴진스	도입	노래
	'Hype Boy' Official MV (DANIELLE&HAERIN ver.)	다니엘	뮤비	사랑	표정	매력	연기
		민지	농구	뉴진스	금발	고양이	싱잉
		하니	뉴진스	고양이	하이브	데뷔	사랑해
		해린	고양이	뮤비	매력	사랑해	아기
		해인	애기	고양이	따름	원통	고삼
	'Hype Boy' Official MV (HANNI ver.)	다니엘	뮤비	노래	민희진	얼마나	절음
		민지	뮤비	뉴진스	화이팅	해린	햇갈림
		하니	버전	사랑해	뮤비	뉴진스	언니
		해린	앞머리	노래	아기	고양이	생글생글
		해인	데뷔	클럽	해린	햇갈림	서로

=== 각 토픽에서 중요한 단어 5개 ===

토픽 1: 0.201*''사랑'' + 0.052*''별로'' + 0.044*''응원'' + 0.028*''항상'' + 0.019*''의상''

토픽 2: 0.484*''미스'' + 0.034*''화이팅'' + 0.019*''활동'' + 0.015*''이제'' + 0.012*''장르''

토픽 3: 0.074*''릴리'' + 0.050*''파트'' + 0.046*''실문'' + 0.042*''규진'' + 0.039*''지우''

토픽 4: 0.052*''데뷔'' + 0.034*''그룹'' + 0.027*''하나'' + 0.017*''중간'' + 0.016*''한국''

토픽 5: 0.297*''노래'' + 0.039*''뮤비'' + 0.030*''이벤'' + 0.027*''느낌'' + 0.027*''중독''

토픽 6: 0.033*''그것'' + 0.029*''보컬'' + 0.028*''타이틀곡'' + 0.027*''수룩'' + 0.025*''파이팅''

토픽 7: 0.058*''느낌'' + 0.028*''스타일'' + 0.028*''지니'' + 0.027*''역시'' + 0.026*''권모술수''

토픽 8: 0.074*''대박'' + 0.074*''가사'' + 0.020*''본부'' + 0.019*''신인'' + 0.019*''마음''

토픽 9: 0.036*''음악'' + 0.032*''일그룹'' + 0.025*''대중'' + 0.020*''생각'' + 0.020*''대중성''

토픽 10: 0.067*''축하'' + 0.065*''원백'' + 0.060*''최고'' + 0.043*''당신'' + 0.030*''컨셉''

=== 문서별 가장 관련 있는 토픽 ===

문서 1: 토픽 5 (노래), 확률: 0.4087

문서 2: 토픽 7 (스타일), 확률: 0.3717

문서 3: 토픽 1 (사랑), 확률: 0.2831

4 외분핵을 필두로 다 평행뜨길리다로저 핵판이 얼마나 노력했을까멤버들 모두 작하고 이... 노래

5 노래는 좋은데 가사는 개짜치는 별 노래

6 해린이 외모제크 이후로 전편된 멤버 모두 작하고 예뻐 실력도 탑이네 노래

7 와 설윤이랑 지우랑 비주얼 폭발하는 노래 하니 타이틀로 내자 연믹 요즘 솔솔 물... 음악

8 이어폰 끼고 들으니까 무슨 연극같은 멤버

9 구림 보컬

10 이 노래가 워 라고위에서위 안에 들어갈줄 알았는데 노래

11 노래가 연이따구나 대형기획사에서 낸 뮤비 조회수가 증소 걸그룹보다 못하네 강 예능할... 음악

12 사랑해요 사랑

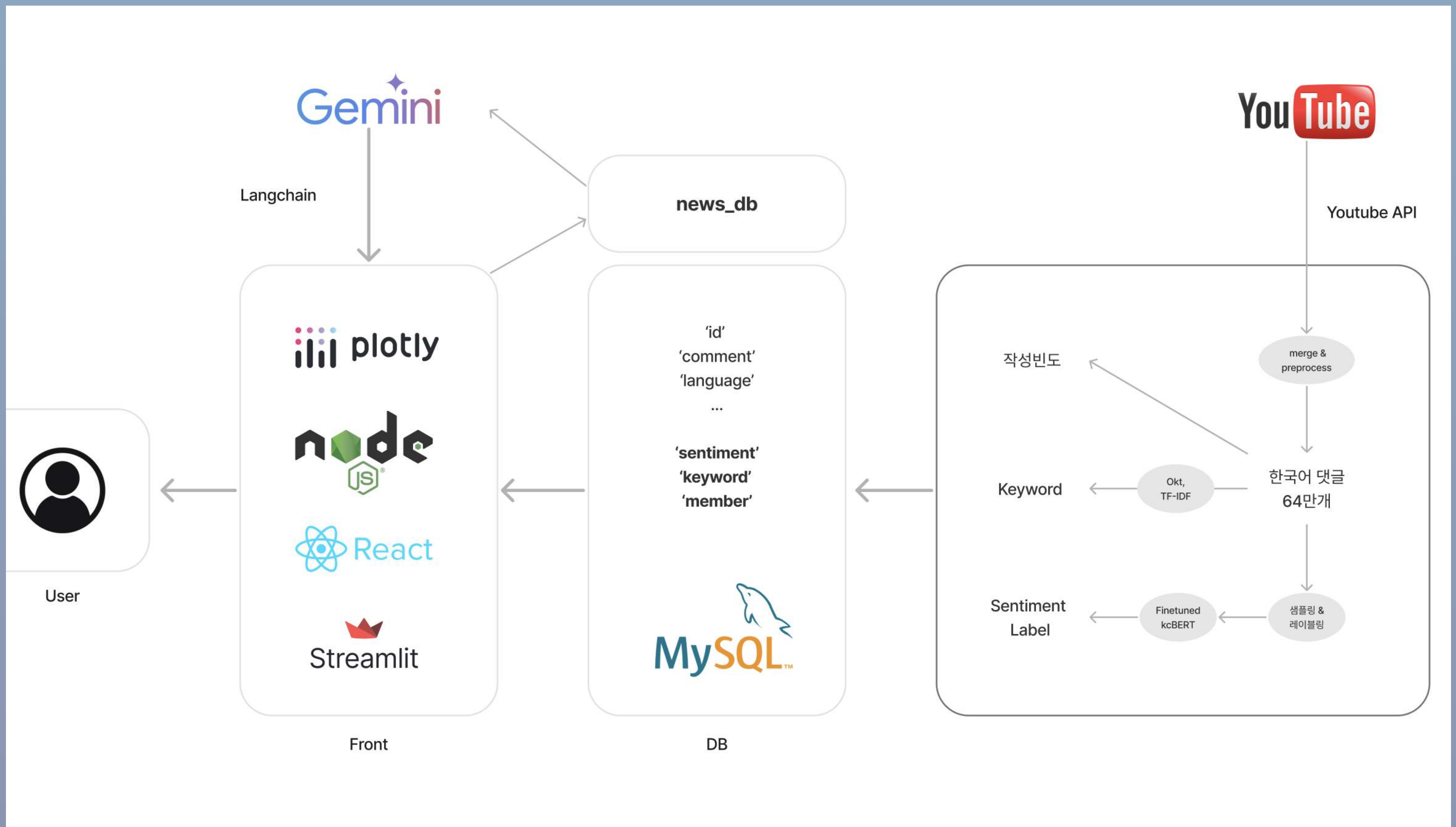
13 규진이 교정웃음 웨이러이쁘지 멤버

14 중간에 연믹스 나오는 부분좀 그만했으면 좋겠어 백급겨 노래

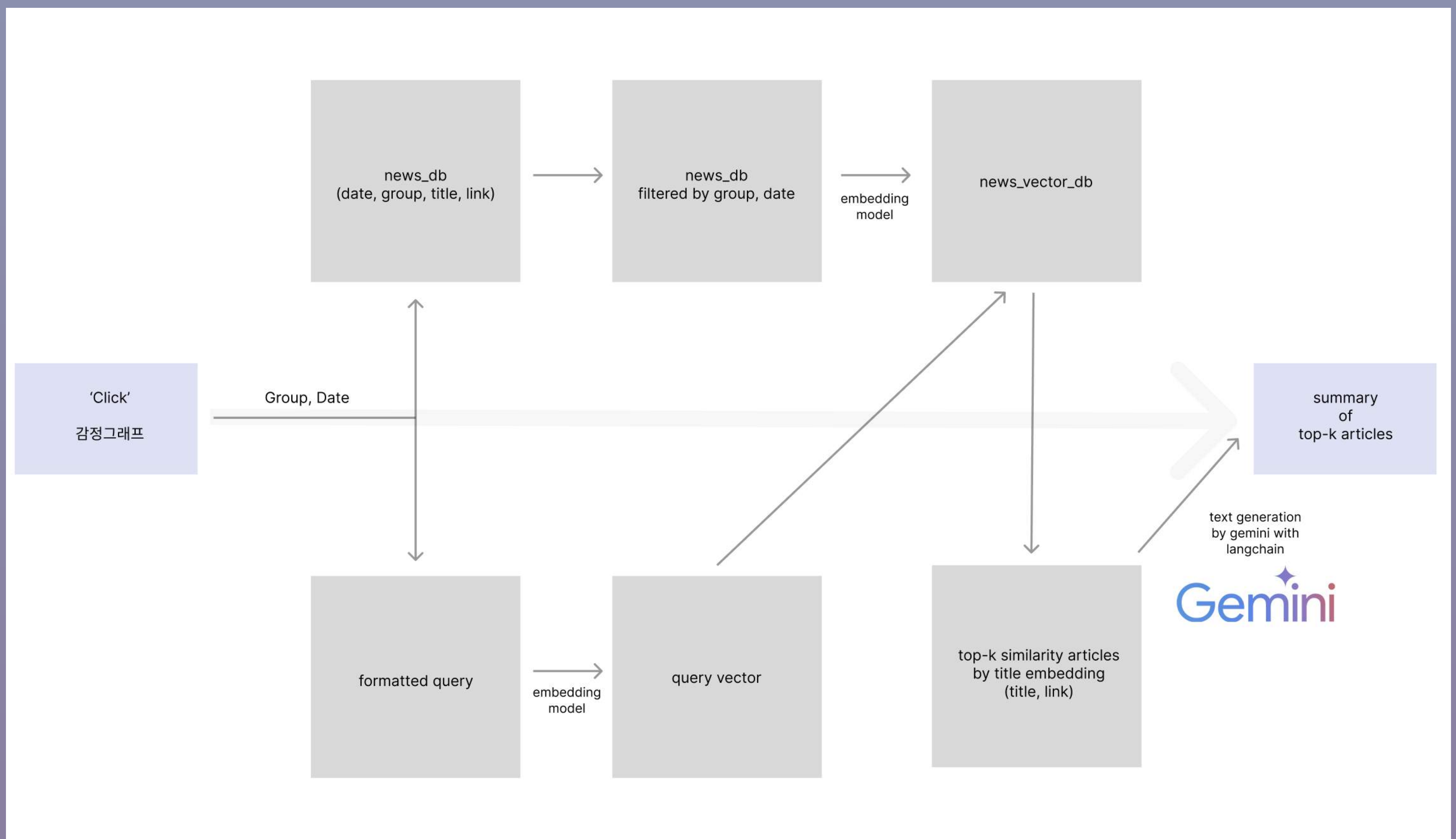
15 소녀감성 다 필요없어평생 힙합해줘 노래

16 판 아티스트 최초 빈보드 라티 모직 음악 차트 음트보로 진저 거세 어깨 나 바지매려과 가사

키워드 추출과 토픽 모델링(LDA)



Trendpop Youtube Analyzer의 전체 system architecture

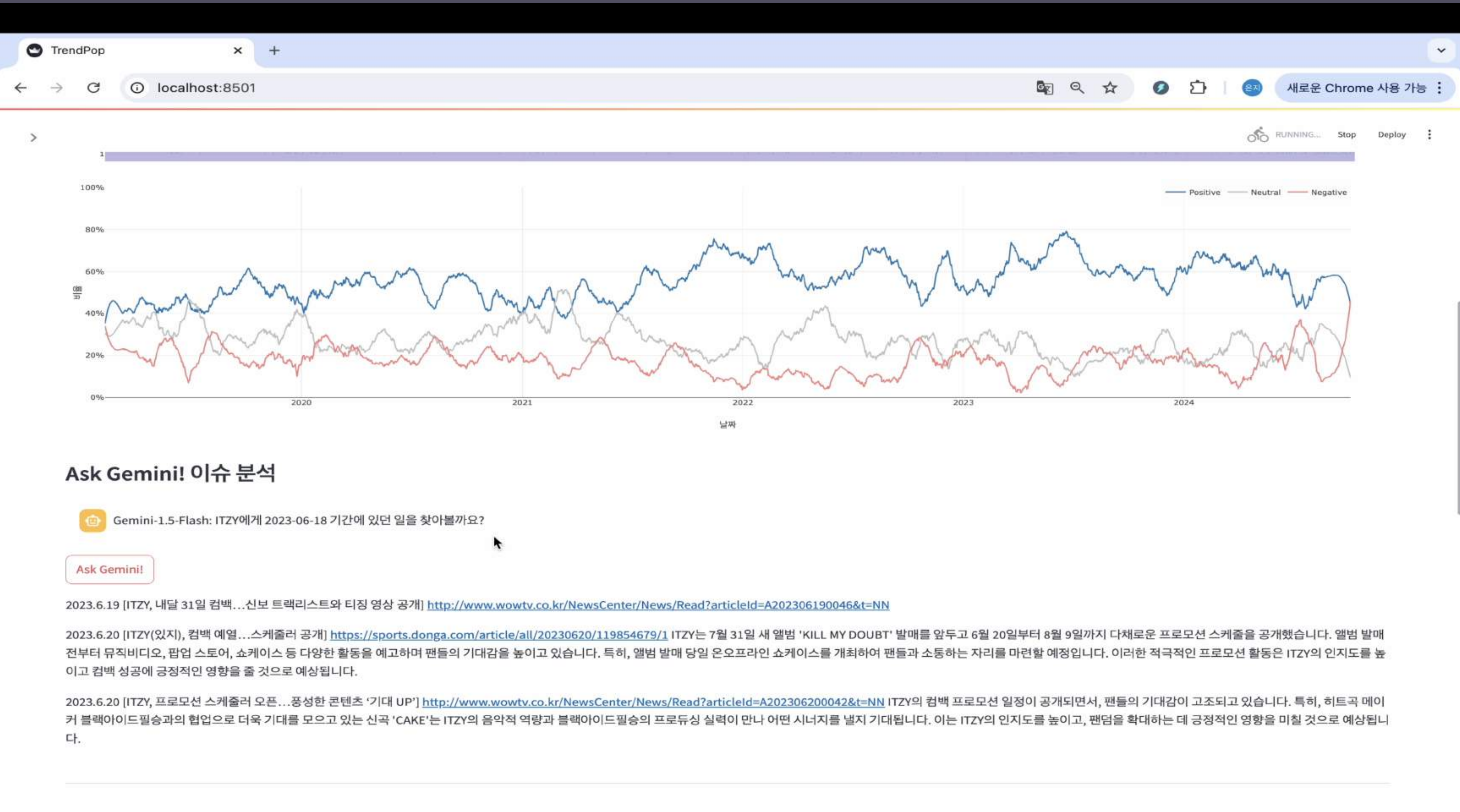


Gemini를 활용한 뉴스요약봇의 workflow diagram

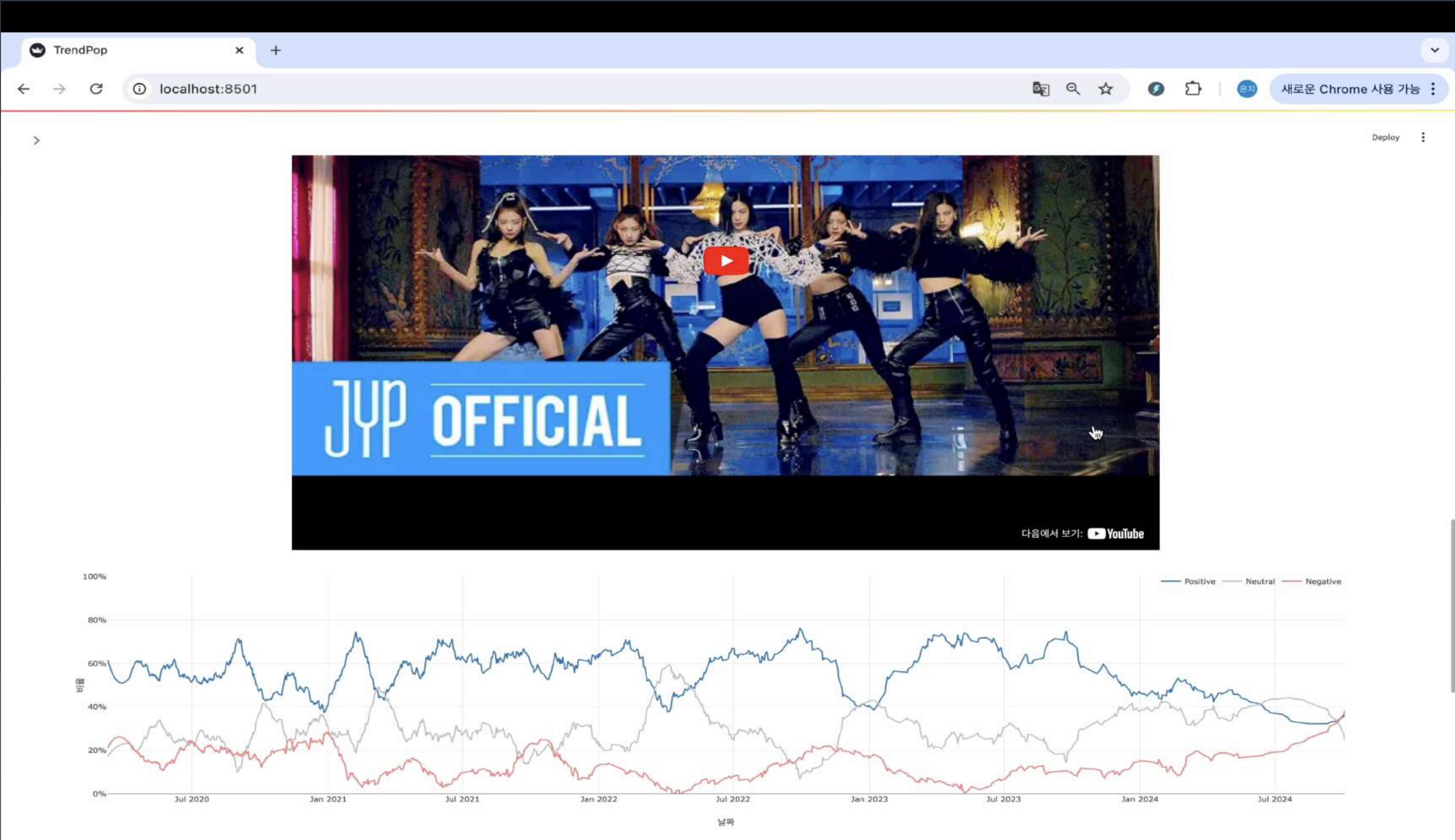
Demo Output



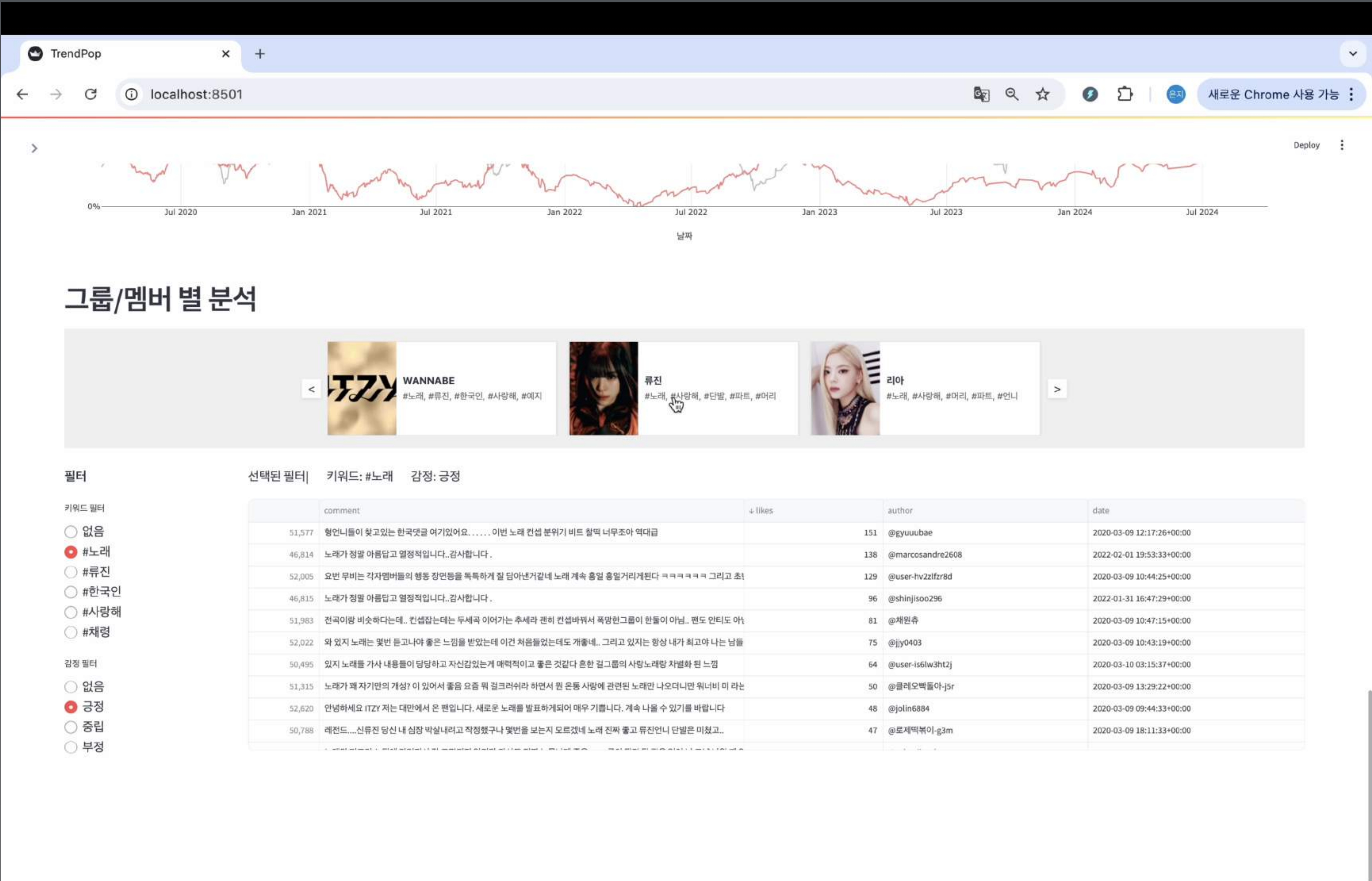
분류된 감정의 비중 추세 시각화



원하는 시기의 관련 뉴스 요약



특정 뮤직비디오 감정반응 추세 시각화



뮤직비디오/멤버별 키워드 및 감정분류에 따른 필터링 + 좋아요 순 정렬기능

성과

TRENDPOP 대시보드는 팬덤의 감정 변화와 주요 이슈를 흐름을 파악할 수 있어, 아티스트와 소속사가 적절한 전략적 대응을 할 수 있도록 지원합니다. 특히, 특정 이벤트 시 팬들의 긍정적 또는 부정적 반응을 감지하고, 뉴스 요약을 통해 감정 변화의 원인을 파악하여 마케팅 전략을 조정할 수 있는 유용한 인사이트를 제공할 수 있습니다. 또한, 팬덤의 주된 관심사와 키워드를 세분화하여 팬들과의 상호작용을 강화하고, 데이터 기반 마케팅 전략을 세울 수 있는 기반을 마련했습니다.

한계와 발전 방향

하이퍼파라미터 튜닝과 더 많은 데이터 레이블링을 통해 현재 감정 분석 모델의 성능을 더욱 끌어올릴 계획입니다. 또한, 실험적으로 RAG를 활용하여 댓글의 맥락을 반영하는 방식을 도입하는 방안도 검토하고 있습니다.

다음으로, 댓글의 90% 이상이 한국어가 아닌 다른 언어로 작성된 만큼, 감정 분석의 대상을 다국어로 확장하는 것이 중요한 과제입니다. 이를 위해 다국어 모델을 개발하거나 기계 번역을 바탕으로 영어 모델을 활용하는 방법을 고려 중입니다.

주제 분류 모델 역시 발전이 필요합니다. 댓글에 특정 키워드가 직접적으로 언급되지 않더라도, '비주얼', '댄스', '지지', '타그룹 언급' 등 사전 정의된 주제로 분류할 수 있는 모델을 개발할 계획입니다. 이를 통해 팬덤의 반응을 보다 정교하게 분석할 수 있을 것입니다.

마지막으로, 실시간 및 배치 방식으로 댓글 데이터베이스를 자동으로 최신 상태로 유지하고, 실시간 모니터링이 가능한 환경을 구축하려 합니다. 또한, 실제 서비스 구현을 위해 모든 과정이 서버에서 안정적으로 이루어지고, 사용자에게 신속하게 정보를 제공할 수 있는 인프라도 마련할 것입니다.

Contribution

- 유튜브 댓글 감정 분류 아이디어 제안: 유튜브 댓글을 통한 팬덤의 감정 분석 아이디어를 제안하고, 해당 분석이 가져올 비즈니스 임팩트와 댓글에서 추출할 수 있는 정보를 정의했습니다.
 - 분석 목적 및 정보 제공 방식 구체화: 감정 분류와 키워드 추출 모델을 결합하여 댓글로부터 어떤 정보를 추출하고, 사용자에게 어떤 인사이트를 제공할 수 있을지 정의했습니다.
 - 데이터 수집 및 전처리 설계 및 구현: Selenium과 YouTube API를 바탕으로 크롤러 초안을 잡고, 각 뮤직비디오에서 수집된 댓글 데이터를 전처리한 뒤 하나의 파일로 병합하는 로직을 설계하고 구현했습니다.
 - 모델 선정: 감정 분류 모델의 국내외 선행 연구를 검토하고, 다양한 모델을 표로 정리하여 KcELECTRA를 베이스 모델로 선정했습니다.
 - Ground Truth 정의 및 평가 기준 설정: 3명의 코더가 1,000개의 댓글을 동일하게 라벨링하도록 제안해 그 신뢰도를 계산하고, 코더 간 차이가 큰 경우를 고려해 만장일치로 긍정/중립/부정으로 분류한 733개의 댓글을 성능 평가의 기준으로 삼았습니다.
 - 성능 평가 및 비교 대시보드 제작: OpenAI와 Gemini API를 통해 테스트셋에 대한 성능 평가를 빠르게 진행하고, 파인튜닝한 BERT 기반 모델과 성능을 비교할 수 있는 대시보드를 제작했습니다.
 - 성능 개선 및 레이블링 프로그램 제작: 모델 성능을 개선하기 위해 베이스 모델이 잘못 분류한 댓글의 '실패 원인 분석'을 수행하고, 지도 학습을 위해 각각 4,000건씩 총 12,000건을 라벨링하는 방안을 제안했습니다. 또한, 화살표 입력만으로 레이블을 지정할 수 있는 레이블링 프로그램을 제작하여 레이블링 효율을 높였습니다.
 - 키워드 추출 및 사전 구축: 키워드 추출을 위해 빈도 기반과 TF-IDF 기반 방법을 테스트하고 구현했으며, 명사 추출을 위해 한국어 형태소 분석기 Okt와 Mecab을 시도해보고 최종적으로 Okt를 선정했습니다. 또한, 잘못 파싱된 단어를 걸러내기 위해 불용어 및 대체어 사전을 구축하고, 나무위키를 바탕으로 멤버별 닉네임 사전을 구축해 멤버명을 인식할 수 있도록 했습니다.
-

- 주제 분류 및 군집화 시도: 댓글의 주제 분류를 위해 LDA 기반의 토픽 모델링과 Sentence Transformer를 활용한 k-means, DBSCAN 기반 군집화를 시도하였으나, 각 토픽/군집의 응집도 및 해석가능성이 떨어져 최종 서비스에 탑재하지는 않았습니다.
- 뉴스 요약 기능 구현: 팬덤 감정 변화의 배경 정보를 제공하기 위해 구글과 네이버 뉴스 API의 한계를 고려하여 크롤링을 통해 뉴스 제목과 링크를 데이터베이스로 구축했습니다. OpenAI 임베딩 모델을 활용해 쿼리와 유사도가 높은 기사 3개를 선정하고, 해당 기사의 내용을 요약하여 제목과 링크와 함께 제공하도록 구성했습니다.
- 댓글빈도와 감정을 시각화하기 위해 지수가중이동평균과 Savitzky-Golay 필터, Kalman 필터 등을 테스트해보고, 빈도에 대해서는 log-scale의 지수가중이동평균, 감정추세에 대해서는 Savitzky-Golay 필터를 선택하여 구현했습니다.

updated 2024.10.31

a journey from social science to machine learning practitioner

Thank You

Data Scientist
Shin Jaesol

ysys143@naver.com

010-7196-8188

for more information, visit my blog socml.tistory.com and github <https://github.com/ysys143>