유튜브 댓글 기반
마케팅 의사결정 지원 시스템
TrendPop

4팀: 이건모 | 고은지 | 신재솔 | 이지선

목차

1.팀원소개
2.프로젝트 목적과 필요성
3. 진행 과정
4. 화면 구성
5. 시스템 설계
6. 선행 연구
7. 개발 과정
    7-1. 모델링 선정 | 라벨링
    7-2. 파인튜닝 과정 | 성능평가
    7-3. 키워드 추출
    7-4. 댓글 주제 분류
    7-5. Gemini 활용
    7-6. 이슈사항 및 해결 방안
8. 시연
9. 향후 계획과 기대효과

팀원 소개

눈꽃여왕 고은지
#모델링 #전처리
#파인튜닝

빠리시간 신재솔
#크롤링 #키워드
#Gemini

갓생인생 이건모
#프론트 #전처리
#파인튜닝

고양이손 이지선
#기획 #UI 설계
#QA

프로젝트 목표

[연예공화국]<5>전 세계 한류 첨병 BTS→블랙핑크→뉴진스

권성훈 기자 cdrom@imaeil.com

매일신문 입력 2023-12-16 07:00:00 수정 2023-12-17 18:12:29

> 5년 만에 글로벌 문화 영향력 7위에 입성, 24단계 도약
> '불어라 한류' 연간 생산유발효과 10조원 안팎
> LPGA 황금세대 '박세리 키즈' 이후처럼 한류도 위기 올 수 있어

대한민국은 연예 강국이다. 전 국민이 연예인(셀럽)에 열광하고, 어릴 때부터 꿈이 대다수 '연예인'이다.

세계 문화대국 한국의 위상이 갈수록 높아지고 있다. 글로벌 문화 영향력은 2017년 세계 31위에서 2021년 7위로 도약했다. 이제는 미국과 유럽 문화강대국(영국, 프랑스, 이탈리아, 스페인) 그리고 이웃나라 일본과도 어깨를 나란히 할 정도로 소프트파워를 갖게 된 것이다.

프로젝트 목표

[연예공화국]<5>전 세계 한류 첨병 BTS→블랙핑크→뉴진스

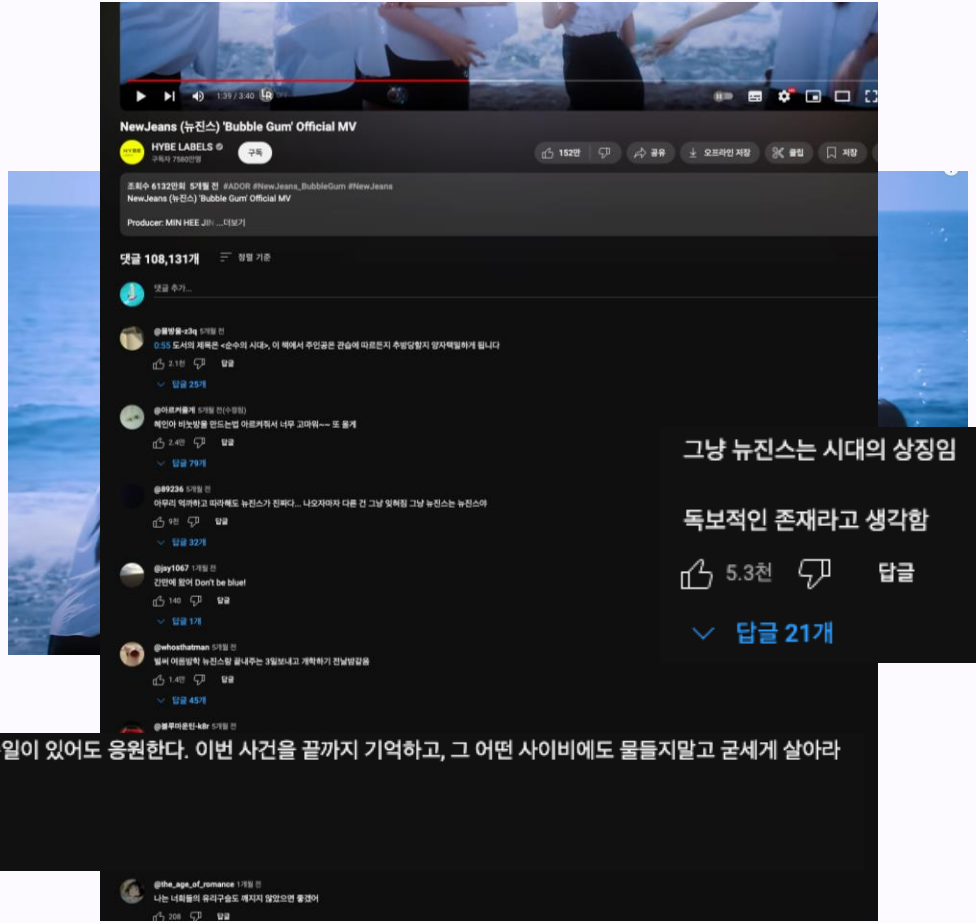
권성훈 기자 cdrom@imaeil.com

매일신문 입력 2023-12-16 07:00:00 수정 2023-12-17 18:12:29

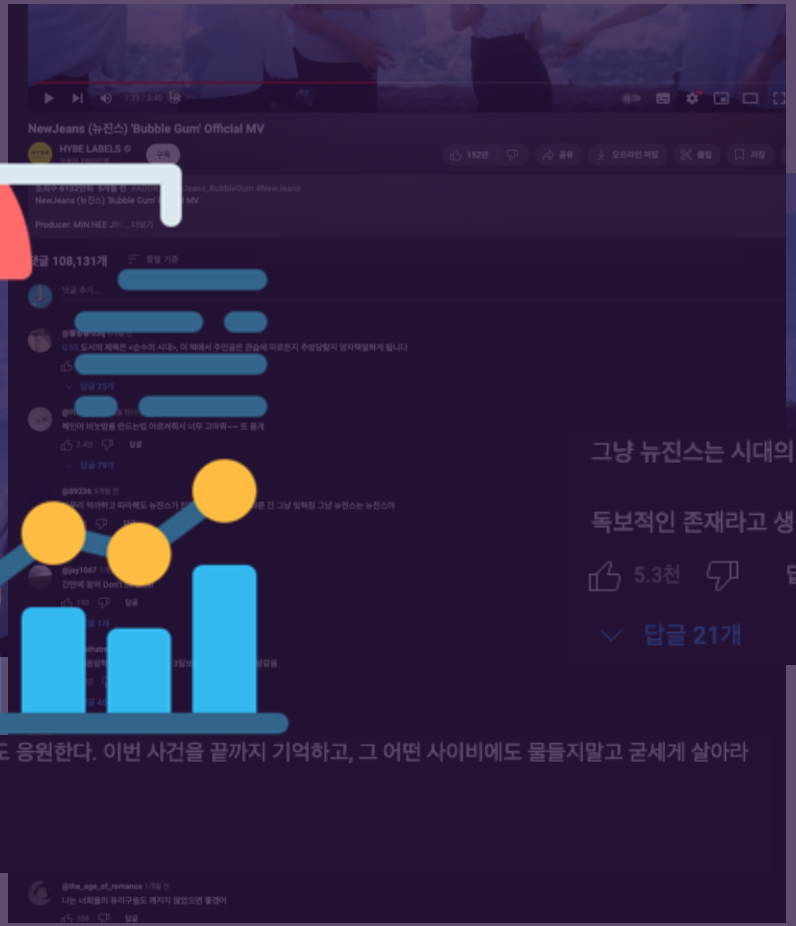
> 5년 만에 글로벌 문화 영향력 7위에 입성, 24단계 도약
> '불어라 한류' 연간 생산유발효과 10조원 안팎
> LPGA 황금세대 '박세리 키즈' 이후처럼 한류도 위기 올 수 있어

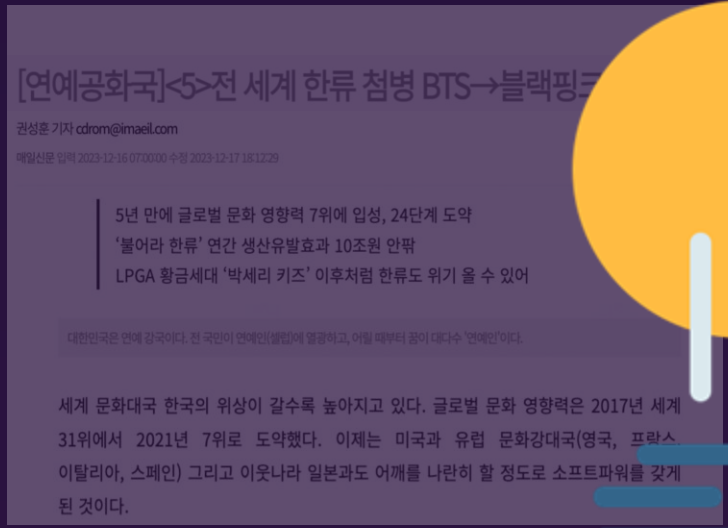
대한민국은 연예 강국이다. 전 국민이 연예인(셀럽)에 열광하고, 어릴 때부터 꿈이 대다수 '연예인'이다.

세계 문화대국 한국의 위상이 갈수록 높아지고 있다. 글로벌 문화 영향력은 2017년 세계 31위에서 2021년 7위로 도약했다. 이제는 미국과 유럽 문화강대국(영국, 프랑스, 이탈리아, 스페인) 그리고 이웃나라 일본과도 어깨를 나란히 할 정도로 소프트파워를 갖게 된 것이다.



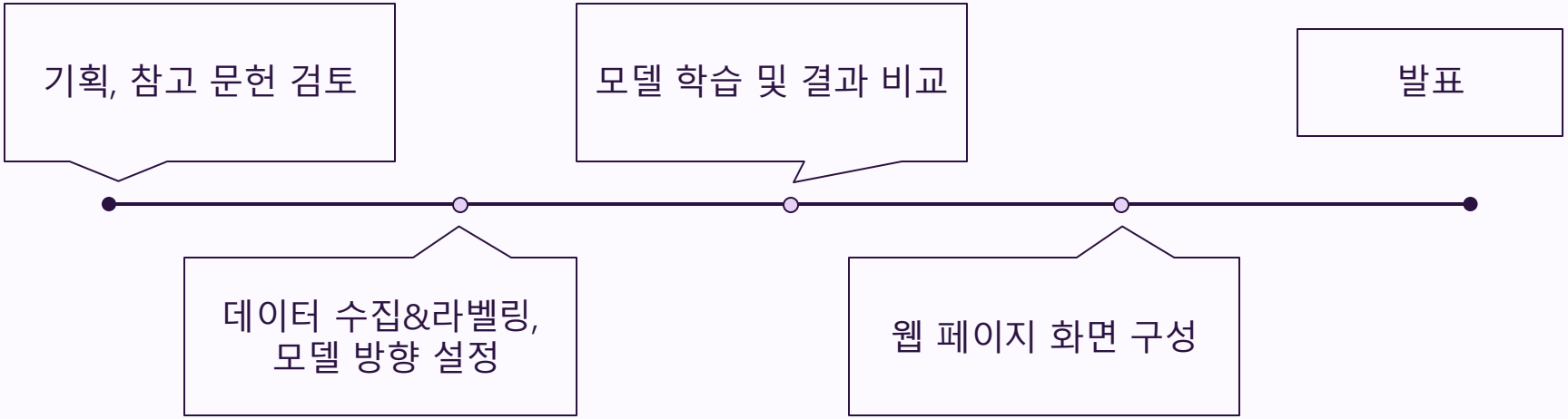
프로젝트 목표


NewJeans (뉴진스) 'Bubble Gum' Official MV
HYBE LABELS
댓글 108,131개
그냥 뉴진스는 시대의 상징임
독보적인 존재라고 생각함
5.3천
답글
답글 21개

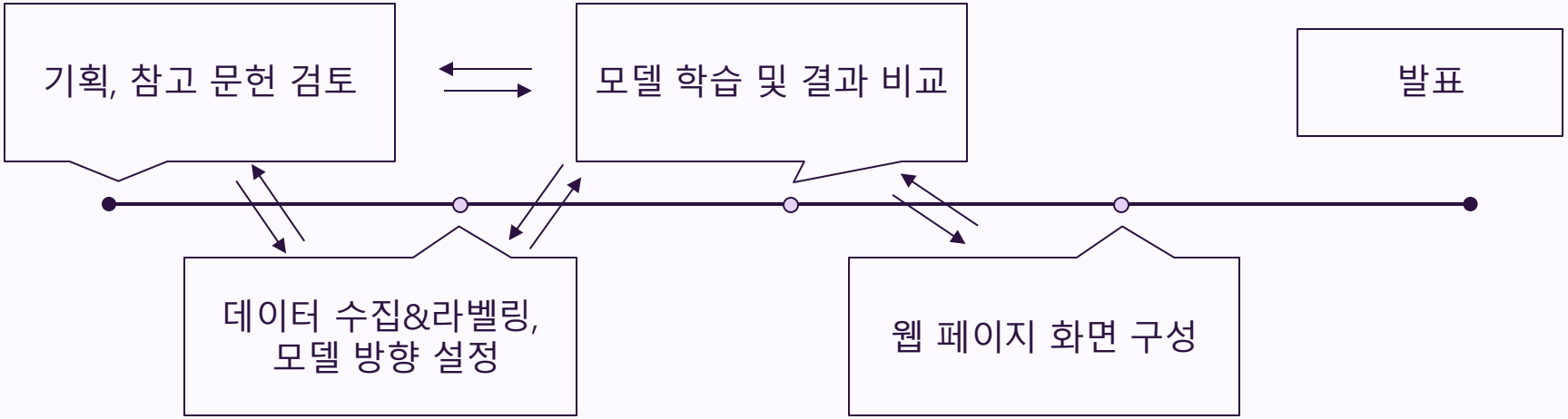

[연예공화국]<5>전 세계 한류 첨병 BTS→블랙핑크
권성훈 기자 cdrom@imaeil.com
5년 만에 글로벌 문화 영향력 7위에 입성, 24단계 도약
'불어라 한류' 연간 생산유발효과 10조원 안팎
LPGA 황금세대 '박세리 키즈' 이후처럼 한류도 위기 올 수 있어
세계 문화대국 한국의 위상이 갈수록 높아지고 있다. 글로벌 문화 영향력은 2017년 세계 31위에서 2021년 7위로 도약했다. 이제는 미국과 유럽 문화강대국(영국, 프랑스, 이탈리아, 스페인) 그리고 이웃나라 일본과도 어깨를 나란히 할 정도로 소프트파워를 갖게 된 것이다.

니네 바이블로 뜬줄 알고 있었는데 이번에 사실을 알게 됐다.. 무슨일이 있어도 응원한다. 이번 사건을 끝까지 기억하고, 그 어떤 사이비에도 물들지말고 굳세게 살아라
658
답글
답글 7개

진행 과정(예상)



진행 과정(실제)



화면 구성

Top 10 아티스트 그룹 선택

1 전체 댓글 반응 감성 분석 차트

2 Gemini로 뉴스 요약 검색(RAG)

뮤직비디오 선택

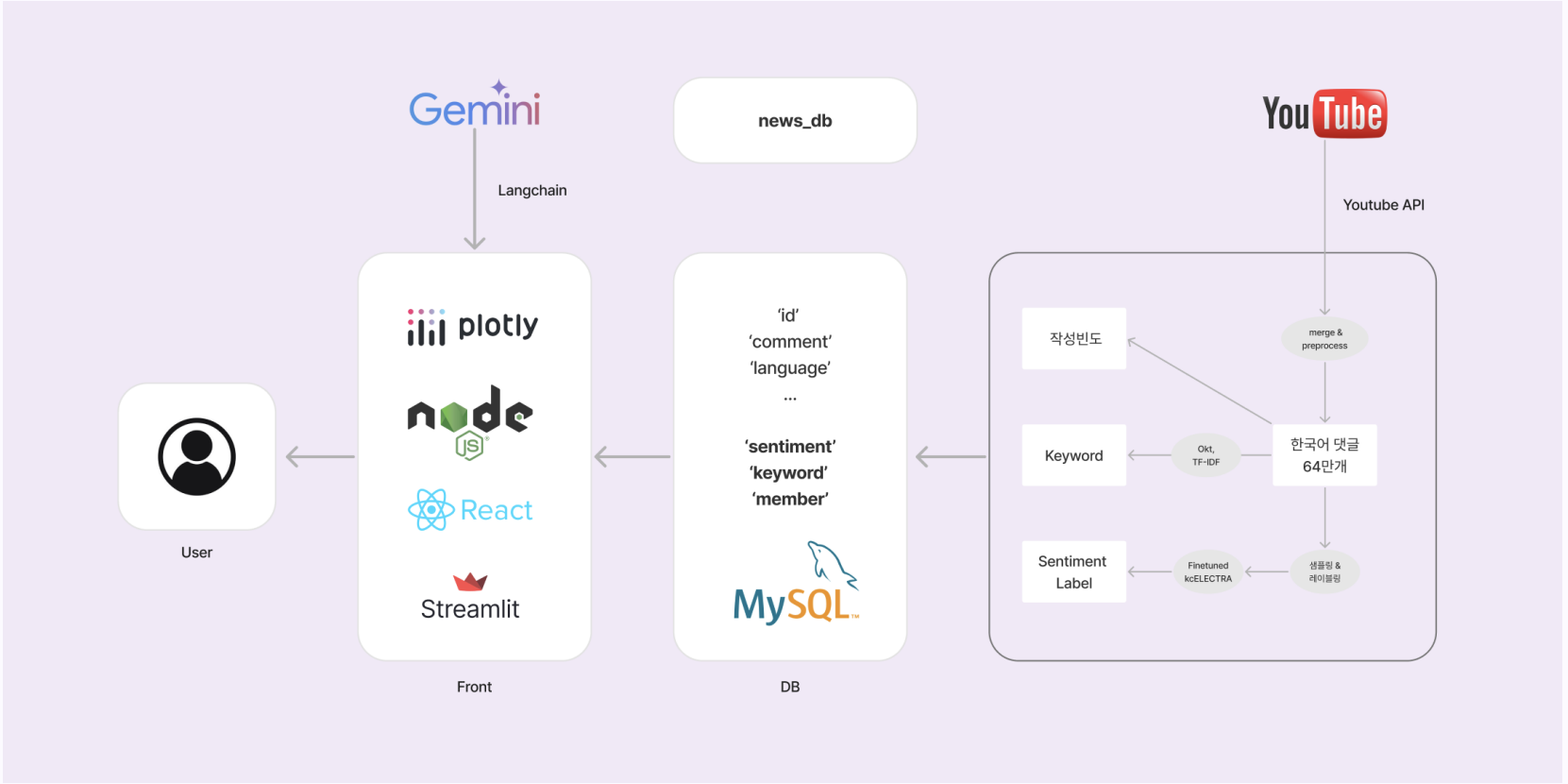
3 뮤직비디오 별 반응 감성 분석 차트

멤버 선택

4 그룹/멤버 별 키워드 분석 기능

5 키워드, 감정으로 댓글 필터링

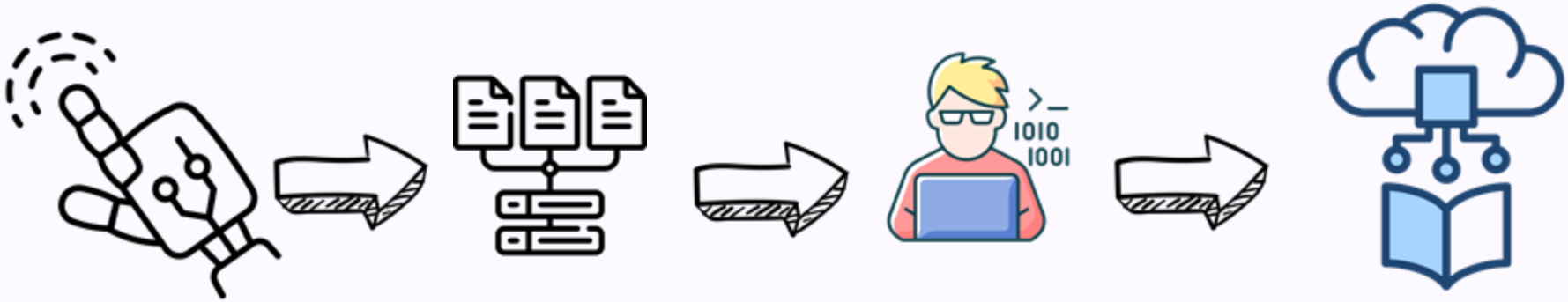
시스템 설계


Gemini
Langchain
news_db
YouTube
Youtube API
plotly
node
React
Streamlit
User
Front
'id'
'comment'
'language'
...
'sentiment'
'keyword'
'member'
MySQL
DB
작성빈도
merge & preprocess
Keyword
Okt, TF-IDF
한국어 댓글 64만개
Sentiment Label
Finetuned kcELECTRA
샘플링 & 레이블링

선행 연구

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, 지도학습	
2013 KOSAC - 한국어 감정 및 의견 분석 코퍼스		2013		Lexicon	형태소분석	svm, 긍부정 극성, 주관적(se		10묶음 교차 검정법 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 전기차		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	쇼핑몰 후기 분석 데이터	Lexicon	KSS 문장분리기, 한 어절 이	의존 구문분석, 규칙 기반 접근	비지도 or 지도 X 의존 관계 활	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) 'KO	NN, Prompt	형태소분석기, 동사 형용사 추	BERT(KcBERT, RoBERTa GPT(3.5, 4)	사전학습, Zero-shot prompt	F1 score, Pearson 상관계수
2016 한국어 감정분석 코퍼스를 활용한 양상정보 기반의 감정분석 연구	신효필, 김	2016			표현유형 태그, seed태그주석	선형 svm,scikit-learn	확률자질 (PES, PPOSS, PS REF-TYPE-SPEECH, REI	10겹 교차검증법, precision, recall, f1, accuracy
LLM과 앙상블 머신러닝모델을 융합한 한국인 감정분석 모델 설계	김현지	2024	AIHub 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, 애널리스	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/모음만 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) AIHub 감정 대화 데이터셋 2) 라디오 청취자 문자 사연을 7가지 - 행복', '슬픔', '놀람', '분	NN	15,831문장 -> 5,531문장 비문법적인 노이즈를 가진 문장 '내용 없음(사진)', 중복된 사연	KoBERT	KoBERT 파인튜닝 모델1 - AIHub 데이터로 학습 모델2 - 라디오 청취자 데이터	Accuracy(라디오 청취자 test set)
[감정 분석 모델] 한국어 감정 분석 데이터셋 KOTE 논문 리뷰 / Python에서 K		2024	KOTE 데이터셋	ML	최소 길이는 10, 최대 길이는 가장 긴 텍스트의 상위 10%의 길이는 404, 평균은 57.32, 중	KoELECTRA 를 KOTE로 파		
뉴스 기사 제목의 감정 라벨링 기법 연구	하재룡	2022	코로나19 관련 뉴스 기사 제목 Naver 포털에서 월별로 약 20 긍정, 부정 수작업 레이블링	Lexicon, NN	관용구,특수문자 등 제거	1) KNU Lexicon 2) KoELECTRA	모델1 - 도메인 단어 사전을 가 모델2 - KoELECTRA 파인튜 모델3 - 추가 파인튜닝(NSMC 모델4 - 세가지 모델을 적절한	Accuracy, Precision, Recall, F1-Score
딥러닝을 활용한 감정 분석 과정에서 필요한 데이터 전처리 및 형태 변형	서혜진, 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					
2019 자연어 처리 기술을 이용한 감정분석 기법에 관한 연구	윤태성	2019	현대 한국어의 어휘빈도 자료		형태소분석기, konlpy.org,			

모델링 선정 | 라벨링



LLM labeling

KcELECTRA Finetuned by Kotedataset

Labeling

Fine-tuning

Group: NewJeans
Title: OMG

울분이 터져나온다 나라를 두번 잃을수는 없다

Sentiment:

List

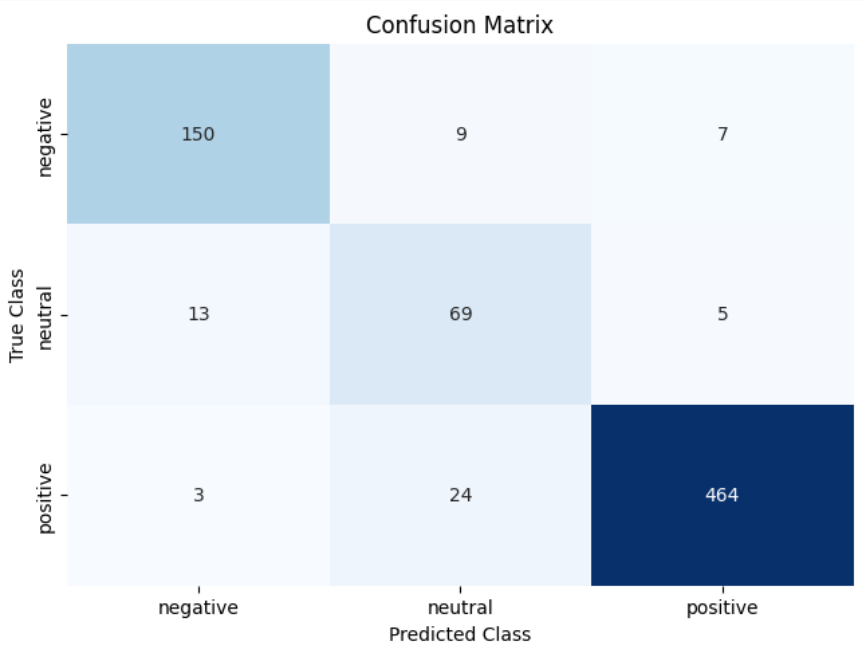
종료

직원들아 탈 하이브는 지능순..
타돌팬이지만 갠적으로 잘됐으면하는 그룹 다 이쁘고 노래도 잘하고 노래가 잘 어울리네요 듣기편하고 화이팅!
놀아줘클럽혜인언니다
난지금까지는DIVE아니였지만 지금부터DIVE될거같아
신곡 미쳤네ㅠㅠㅠ
와미친…
역시 독보적 뉴진스
오늘 주주총회는 열리지 않습니다. 오늘 민희진이 위험하다고 퍼뜨리면 안됩니다. 역공당할 수 있어요. 물론 사이비 단체의 비
뉴진스 안좋아했는데 방시혁땜에 뉴진스 팬됨ㅋㅋㅋ
노래 진짜 구리다
→ 울분이 터져나온다 나라를 두번 잃을수는 없다
노래 진짜 넘좋자나 ㅠㅠㅠ
르세라핌 노래들은 전부 다 띵곡이야.. 더 대박나자아
흠흠. 닝닝를 매우 좋아한다
[timecode] 언니.....

파인 튜닝 | 성능 평가

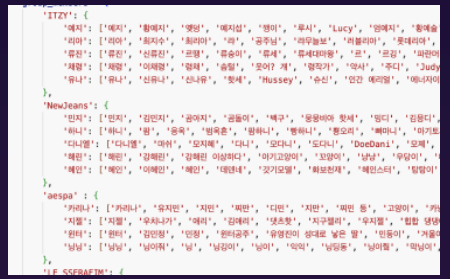
Performance Metrics by various models

Model	Accuracy	Precision	Recall	F1-Score
KcELECTRA 12000 sample finetuned	0.918011	0.851625	0.880576	0.864481
gpt-4o with prompt	0.905914	0.824174	0.841612	0.832502
gpt-4o	0.90457	0.829732	0.866307	0.844036
KcELECTRA 17000-oversampled finetuned	0.895161	0.820019	0.852437	0.833488
gpt-4 finetuned with 283 examples	0.889785	0.824535	0.741701	0.768945
koELECTRA(Max)	0.880376	0.800417	0.781084	0.790099
gpt-4o-mini with prompt	0.879032	0.781747	0.810755	0.793331
koELECTRA base model	0.876344	0.840351	0.723631	0.749794
gpt-4o-mini	0.86828	0.780447	0.794009	0.777354
kote-ELECTRA 200-finetuned	0.643817	0.399864	0.413798	0.37505

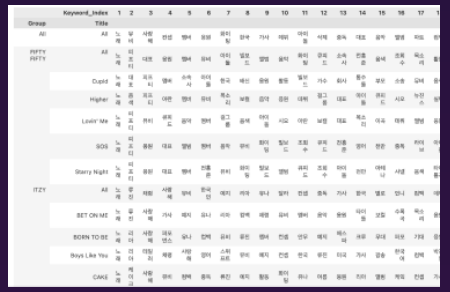


키워드 추출

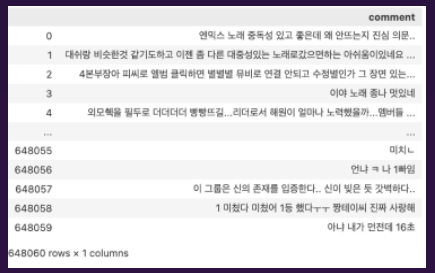
닉네임 사전기반 멤버명 인식



키워드에서 타 멤버 언급제거



댓글 데이터



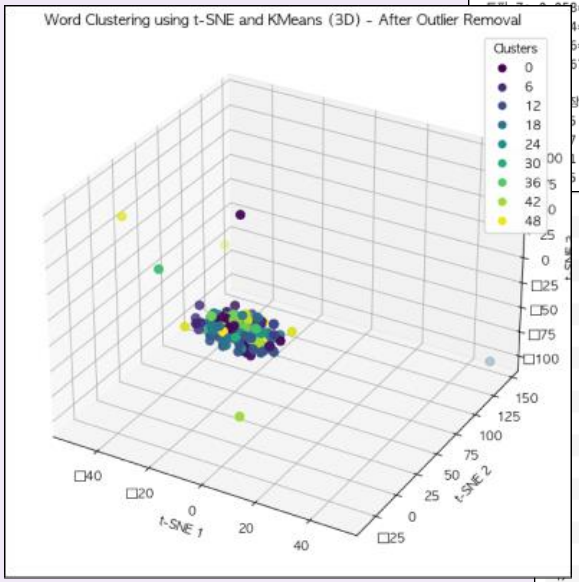
명사추출 : Okt와 Mecab의 비교

키워드 추출

불용어, 사용자정의 단어 추가

키워드 추출 : Count, TF-IDF

댓글 주제분류



Clustering

```
=== 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*"믹스" + 0.034*"화이팅" + 0.019*"활동" + 0.015*"이제" + 0.012*"
토픽 3: 0.074*"릴리" + 0.050*"파트" + 0.046*"설윤" + 0.042*"규진" + 0.039*"지우
토픽 4: 0.052*"데뷔" + 0.034*"그룹" + 0.027*"하니" + 0.017*"중간" + 0.016*"한국
토픽 5: 0.297*"노래" + 0.039*"뮤비" + 0.030*"이번" + 0.027*"느낌" + 0.027*"중독
토픽 6: 0.033*"그것" + 0.029*"보컬" + 0.028*"타이틀곡" + 0.027*"수록" + 0.025*"
토픽 7: 0.058*"느낌" + 0.028*"스타일" + 0.028*"지니" + 0.027*"역시" + 0.026*"
토픽 8: 0.044*"대박" + 0.074*"가사" + 0.020*"본부" + 0.019*"신인" + 0.019*"마음
토픽 9: 0.056*"음악" + 0.032*"걸그룹" + 0.025*"대중" + 0.020*"생각" + 0.020*"
토픽 10: 0.057*"축하" + 0.065*"컴백" + 0.060*"최고" + 0.043*"당신" + 0.030*"

=== 각 댓글과 관련 있는 토픽 ===
0.5 (노래), 확률: 0.4087
0.7 (스타일), 확률: 0.3717
0.1 (사랑), 확률: 0.2831
0.5 (노래), 확률: 0.5500
```

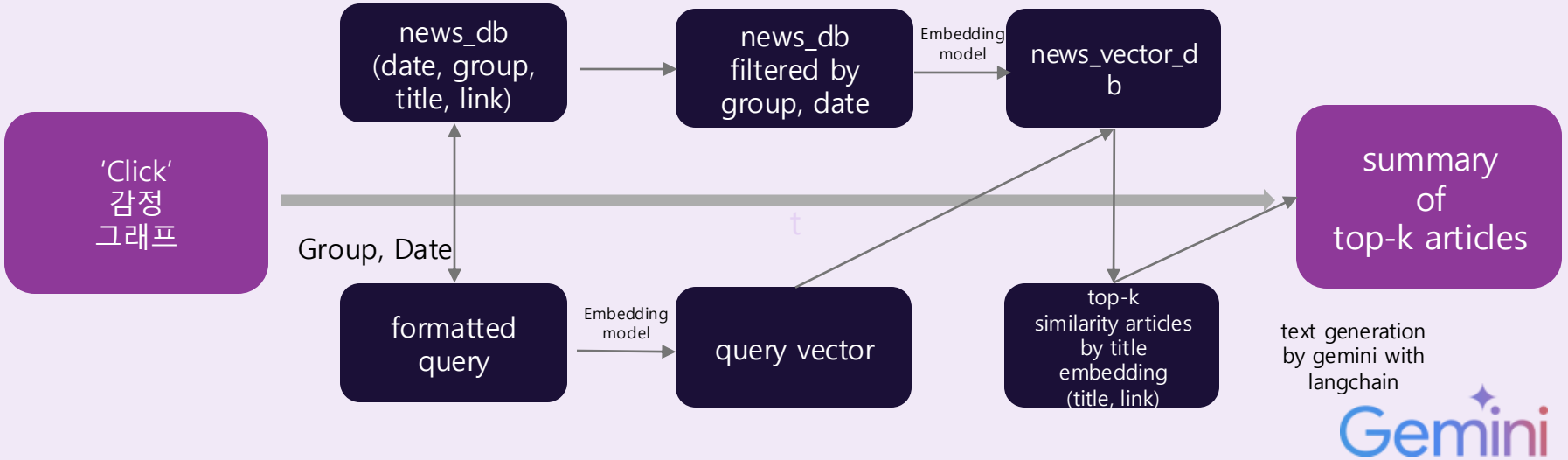
엔믹스 노래 중독성 있고 좋은데 왜 안뜨는지 진심 의문	
대쉬랑 비슷한것 같기도하고 이젠 좀 다른 대중성있는 노래로갔으면하는 아쉬움이있네요 ...	
본부장아 피씨로 앨범 클릭하면 별 뮤비로 연결 안되고 수정별인가 그 장면 있는 곳에...	
이야 노래 좋나 멋있네	
외모핵을 필두로 더 뱅뱅뜨길리더로서 해원이 얼마나 노력했을까멤버들 모두 착하고 이...	
노래는 좋은데 가사는 개짜치는 별	
해원이 외모체크 이후로 찐팬됨 멤버 모두 착하고 예쁨 실력도 탑이네	
와 설윤이랑 지우랑 비주얼 폭발하는 노래 하나 타이틀로 내자 엔믹 요즘 슬슬 물...	
이어폰 까고 들으니까 무슨 연극같음	
구림	
이 노래가 위 라고위에서위 안에 들어갈줄 알았는데	
노래가 먼이따구나 대형기획사에서 낸 뮤비 조회수가 중소 걸그룹보다 못하네 걍 예능활...	
사랑해요	
규진이 교정못음 왜이리이쁘지	멤버
중간에 엔믹스 나오는 부분좀 그만했으면 좋겠어 맥잖거	노래
소녀감성 다 필요없어령생 힙합해줘	노래
팝 아티스트 최초 빌보드 라틴 뮤직 워크 참석 유튜브로 직접 검색 언제나 반전매력과...	가사
규진 설윤 이쁘당	멤버
솔직히 실력으로는 엔믹스가 원탑 아니나	믹스
대박 대박 안 말하고 싶은데도 막 나오네요	가사
베이스 미침	사랑

Topic Modeling(LDA)

G	I
comment	키워드
와씨 이번 컴백곡 미쳤다 (아직 안봄) 보고 왔는데 개지리누 춤연습 영상 벨리 벨리 몰라와라!!	컴백, 댄스, 응원
오늘만 기다렸다	컴백
아니 노래들이 점점 그냥 대중적으로 가자 제발 음악성 어쩌고저쩌고 모르겠고 듣기 편하고 신나는걸로 가자	음악, 컨셉
근데 뉴진스 원래 조회수는 잘 안나옴 데뷔곡부터도 조회수 밀리는데	조회수, 순위
짜증나네.. 하필 ㅈㄲ 어로 노랠 부르다니..	내셔널리티
와 미쳤다(아직안봄)	응원
요즘 지우 설윤 많이 밀어주는 것 같아서 좋을 그렇다고 기존에 밀어줬던 멤버들도 다 맞는 파트 주고 .. 요즘	멤버, 파트, 소속사
1위 가즈아	응원, 순위
새청바지들을 상대하기엔 많이 부족하다 너무 의식해서 오바하지 않길 바람	순위, 컨셉, 타그룹
엔믹스스러워서 좋네 o.o랑 다이스 개좋아했어서...규진님은 진짜 미쳤다	음악, 컨셉
하우스풍 케이팝은 언제나 옳다....	음악, 장르, 하우스
니 인생이 힙씬 멍함	키보드배틀
화이팅	응원
the song is really good keep up the enthusiasm for nmixx. I know nmixx will be more successfu	기타
KYUJIN 사 랑 해 요 !!!!!!!!!!!!!!!	응원
안보고싶으면 보질마셰여 ㅎ	키보드배틀
빌보드 핫 트렌딩 송 1위 추카추카	응원, 순위, 빌보드
NMIXX Like OOH-AHH(OOH-AHH하게)	음악
아니 진심으로 엔씨로서 화나서 그럼 딱고 미친거임?? 4세대 여돌판 거의 전쟁터인데 진심으로 이런걸로 승	음악, 소속사, 컨셉
이게 엔믹스다	응원, 컨셉
사실 내심 이번엔 대중적으로 오나 기대했는데 한번 들으니까 대중적이고 뭐고 우린 이게 맞는거같다ㅋㅋㅋ	음악, 컨셉
뭔 노래인가 했더니 해품달 ost 였네	음악
노래 맛깔난다 진짜.. 이게 엔믹스지	음악, 컨셉
지우 진짜 이쁨!!!	멤버
와아 3분 왜이렇게 짧아요	음악
인급음 1위까지 악 백만뷰 남음~~ vamos!!!!!!!!!!	조회수
NMIXX ()'s Top Popular Songs on YouTube Music! 별별별(See That?) - 35M DASH - 84M Love	기타
이번노래도 대박 나자	음악, 응원
아... 대쉬로 고점박고 이번엔 천장 뚫을 줄 알았더니 여기서 브레이크 걸리네;;. 많이 아쉬운데.... 또다른 곡이	음악, 컨셉
우리 팔들 너무 멋지고 재능이 많아서 매번 놀랍니다!!	응원
이거 노래라고?..	음악, 컨셉, 비난

사전할당 주제어(대주제) 분류
*향후과제

Gemini 활용 이슈요약봇

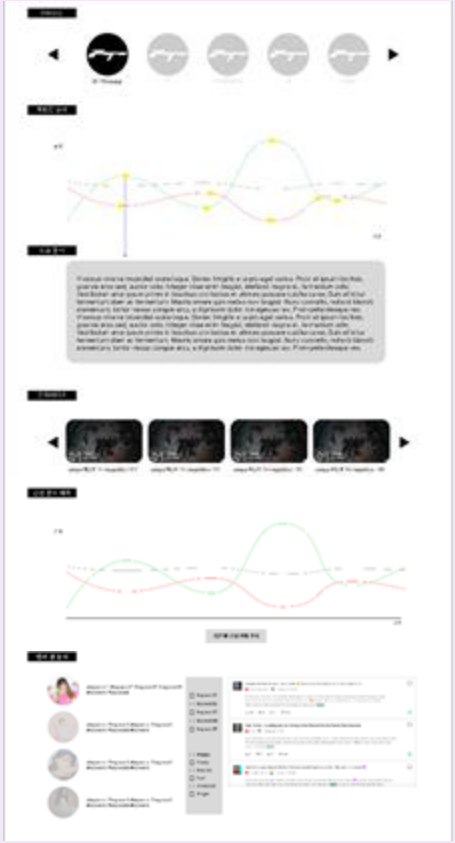


시연

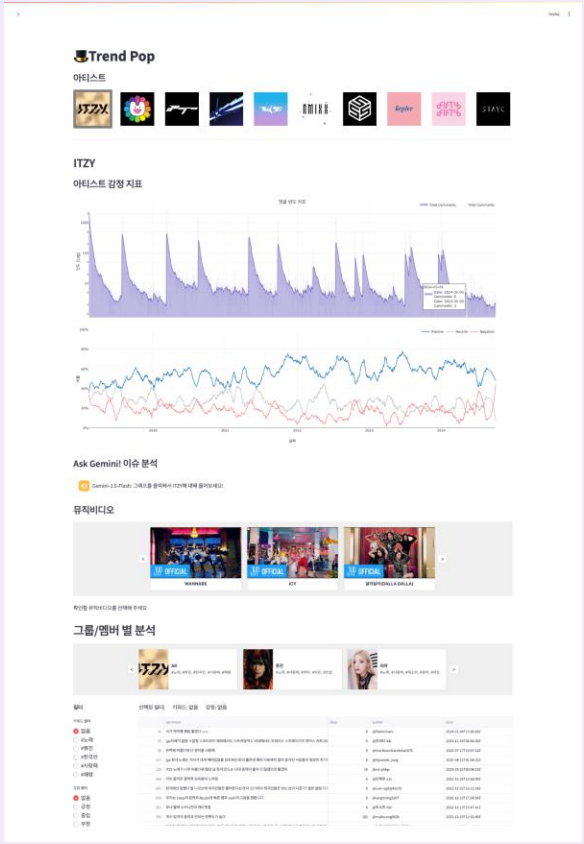
와이어프레임

시연

와이어프레임



구현 완료



이슈 사항 및 해결 방안

실시간 대응 능력 향상

캠페인 성과의 즉각적 평가

트렌드 선도와 맞춤형 콘텐츠 제공

악플러 대응 및 리스크 관리

팬덤과의 소통 강화

향후 계획과 기대효과

[감정분석]
감정분석 모델 고도화

한국어 모델 성능개선

하이퍼파라미터 튜닝
lr, batch_size, n_epochs, decay, dropout, optimizer...

보다 많은 레이블링

(실험적) RAG를 통한 맥락 반영

유튜브 댓글의 특성과 입장, 감정, 맥락에 대한 보다 깊이 있는 연구

[다국어지원]
영어기반모델
(+Translate API)
다국어지원모델

[Realtime]
DB 업데이트 자동화
실시간/배치방식으로
댓글 및 뉴스 수집

[키워드]
키워드 방식보다 더 나은
주제분류 모델 고안

주제분류 지도학습을 위한
레이블 생성

댓글의 특성을 더 잘 반영하는
임베딩 모델 학습

[요약봇]
검색(임베딩) 정확도 개선
속도개선
스트림방식 적용

[웹호스팅 및 서비스]

GCP 및 Vertex AI를 통한
GCP Native Stack 구현

Compute Engine,
Storage,
Vertex AI,
BigQuery

감사합니다.
THANK YOU.
WISH YOU ALL THE BEST!

심사위원님들 피드백
-학습한 데이터 출처
-감정분석 영화 리뷰 데이터를 기준으로 한 이유? 비슷하지 않다
-데이터 전처리 과정을 어떻게 했는가?
-차트에 포커스 표시와 X축 기간 레이블
-감성 사전 만드는 노력!그게 장표에 없다!했던 노력에 대한 흔적을 녹여내라