

Date 2023.03.01 ~ 08

지친 일상에서 쉬어가는 곳

AI 음악 상담사를 통해 당신의 감정을 위로해주세요

음악 심포 ,

민병창, 서영호, 신제우, 이상욱, 황도희

TABLE OF CONTENTS

1. AI 음악 상담사 '공감이'

4. 한계점 및 발전 가능성

2. 구성원 소개 및 역할

5. 출처

3. 프로젝트 결과

6. Q & A

0) 전체 process

1) 챗봇

2) 감정 분석 모델링

3) 노래 추천 알고리즘

4) 데이터베이스 연동 및 활용

5) 결과 시연

1. AI 음악 상담사 '공감이'를 소개합니다.

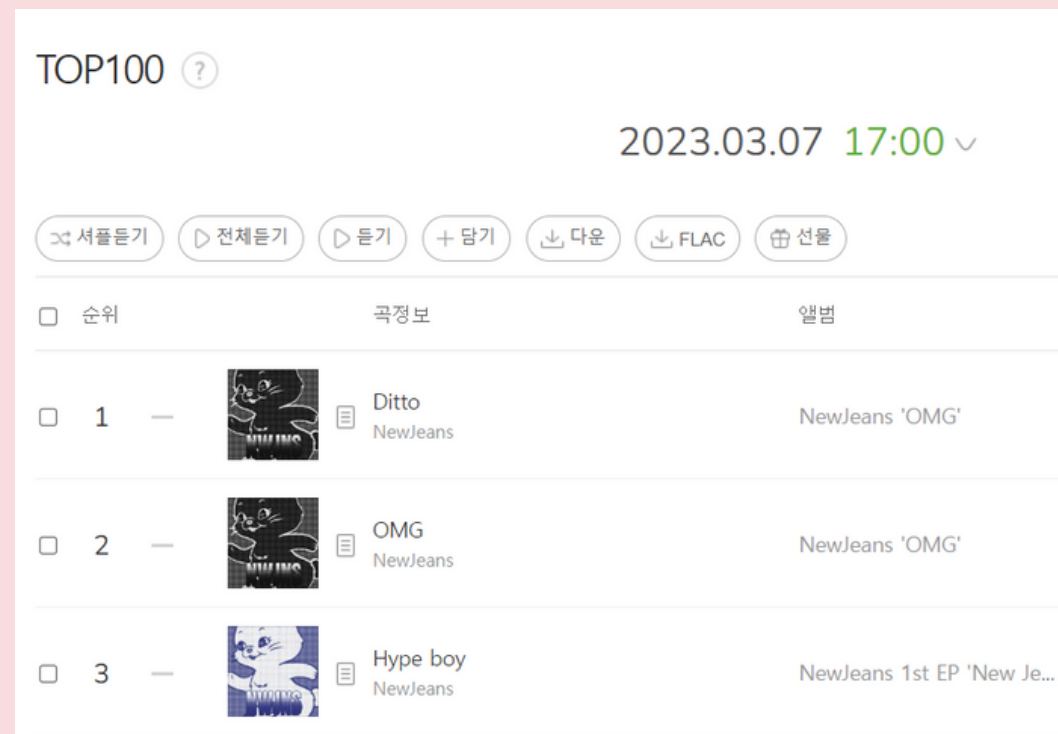
<2022 세계행복지수 >



1. AI 음악 상담사 '공감이'를 소개합니다.

< 기존의 노래 듣는 방식 >

Melón



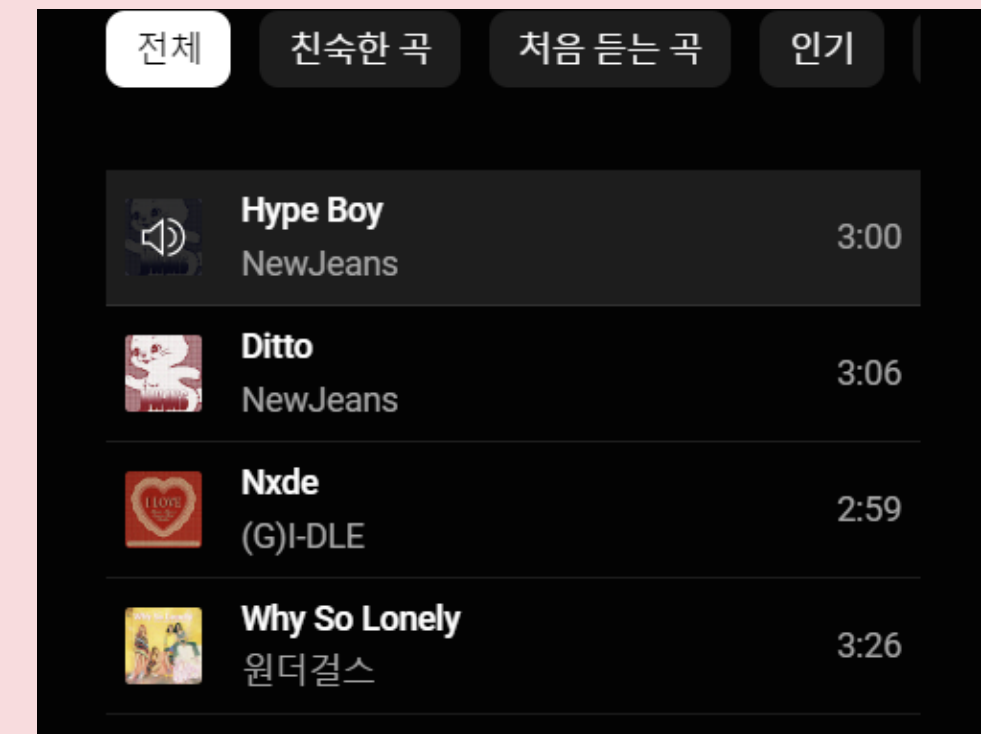
자신에게 맞는 음악을
여러 번 클릭하여 재생 목록에 추가

genie



자신의 감정을 혼자 알아채고
자신에게 맞는 노래를 검색

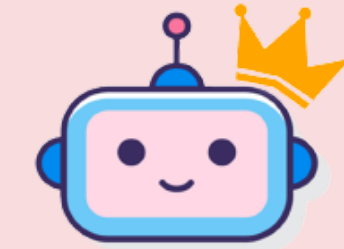
Music



유사 노래 추천

1. AI 음악 상담사 '공감이'를 소개합니다.

Melón genie



우리 공감이는요

자신의 감정을
혼자 알아채고



커뮤니케이션을 통해
자신의 감정 돌아보고
공감 받는 시간

자신에게 맞는 노래 검색
&
SELF MUSIC LIST 형성
(개인의 감정은 개인의 몫)



개인의 감정 상태를
시각화해주는 '감정 지수'
&
감정에 맞는 노래 추천

유사한 노래 추천



유사한 노래 추천

2. 구성원 소개 및 역할

신제우

- KOBERT 모델링
- 깃 관리
- 발표

서영호

- 전처리
 - 감정분석 ML 모델링
- GirdSearchCV 사용하여
CatBoost 튜닝

황도희

- 챗봇 질문지 작성
 - 음악 크롤링
 - 감정분석 ML 모델링
- RandomSearchCV, GirdSearchCV
사용하여 CatBoost, RandomForest,
SVC 학습

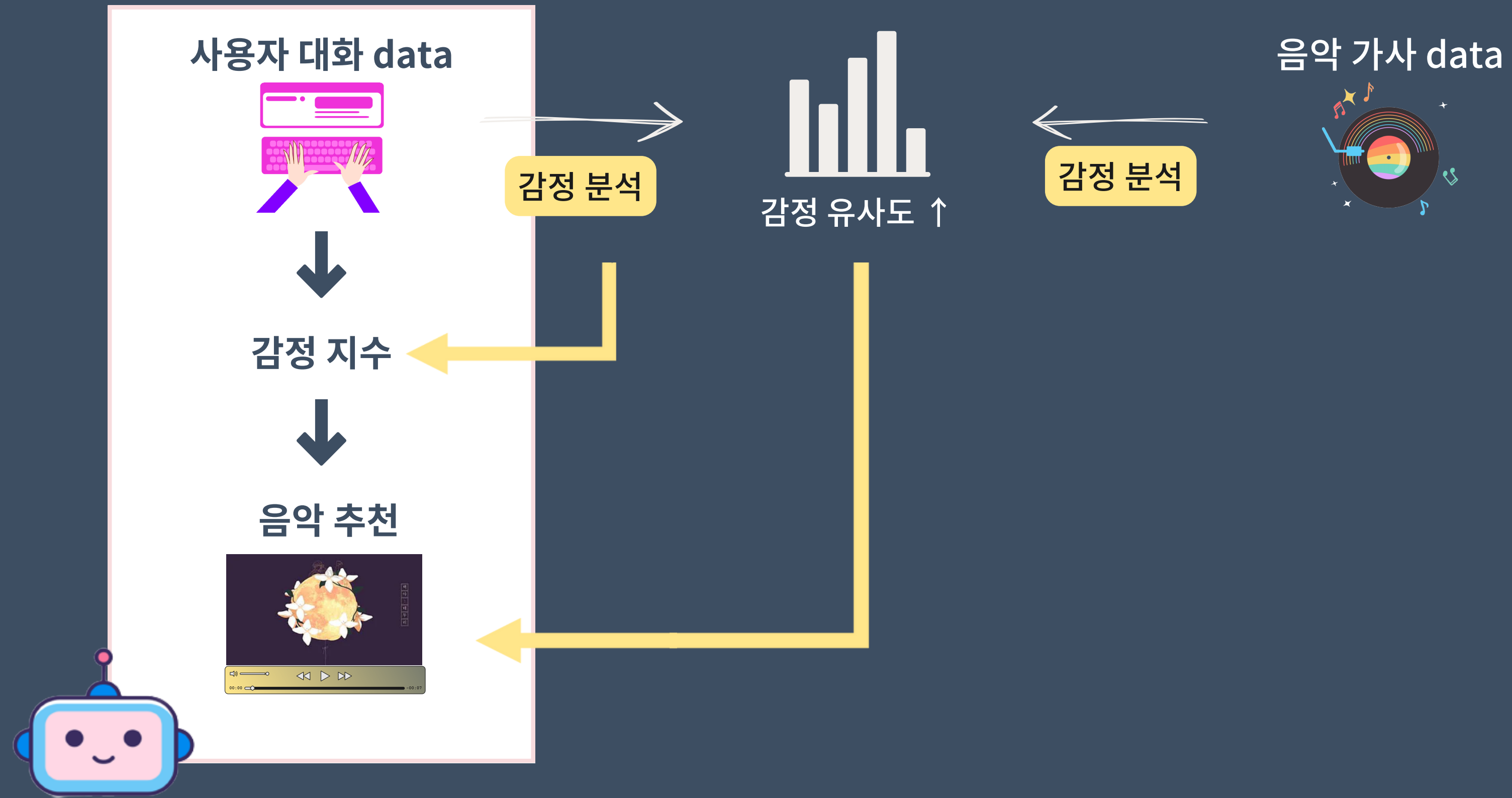
이상욱

- 웹 개발
- 챗봇 개발 및 웹 적용
프론트, 백엔드 개발,DB 연동
- 감정분석 모델링
- ML, DL 베이스 코드 제작
- 데이터 정제&가공
- JSON 특정 데이터 추출
자연어 데이터 가공

민병창

- 감정분석 DL 모델링
- LSTM,GRU 튜닝
- subword토큰화 적용
 - 노래 추천 알고리즘
- cosine 유사도와 random 사용

3. 프로젝트 결과 0) 전체 process



3. 프로젝트 결과 0) 전체 process

기술 스택(앞의 특징점을 가지게 만든 모델 설명)

1) 챗봇

챗봇 작동 운영 방식 + 기술

챗봇의 특징점 : 높은 공감 능력, 감정이 얼마나 차고 있는지 시각화, 입력받은 텍스트 저장

2) 노래가사

크롤링을 통해 가수와 노래 제목 텍스트파일로 멜론에서 가사 받아옴

3) 감정 분석

데이터 정제 방식 간단히

여러 실행 모델 중 top3 보여주고 최종 모델은 선택한 이유 설명

실행했던 유사도 중 왜 지금의 유사도 선택했는지

텍스트 분석과 노래 분석 예시 + 유사도 측정해 연결하는 예시 보여주기

감정 분석 특징점

지금의 내 감정과 유사도가 높은 노래들을 추천 등 (더 고민 필요)

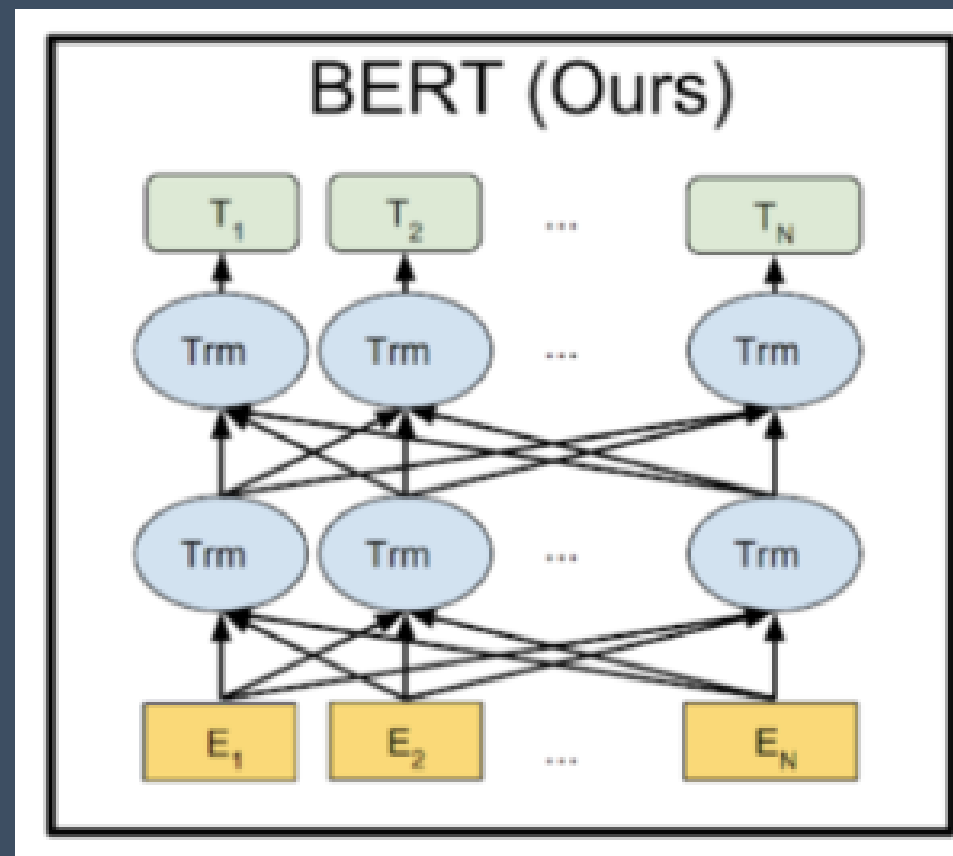
4) 노래추천

3. 프로젝트 결과 1) 챗봇

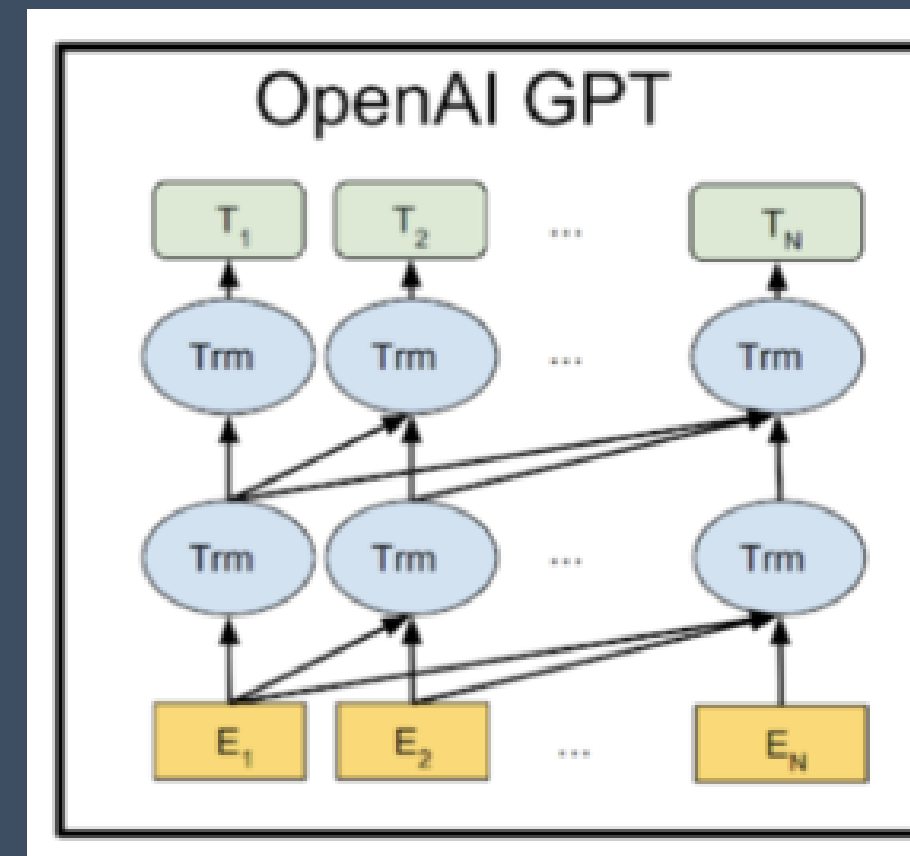
[BERT 모델 선정 이유]

문장을 생성하는 GPT 챗봇 모델보다는

사용자의 문장과 데이터 문장의 일치도를 판단할 때 정확도가 높았던 BERT 모델이 유리하여 선정



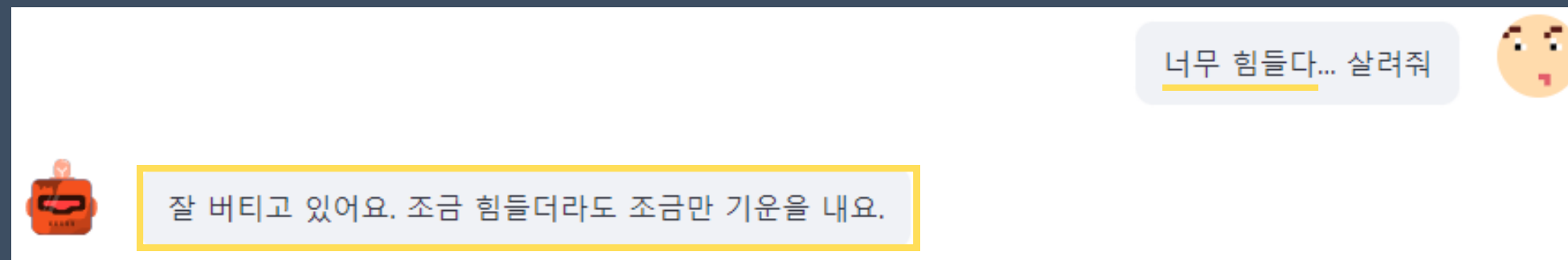
양방향으로 학습하기 때문에
더 많은 정보를 반영하여
더 우수한 학습 성능을 보임



단방향으로 학습하기
때문에 문장 생성이 가능

3. 프로젝트 결과 1) 챗봇 - KO-ROBERTA 모델

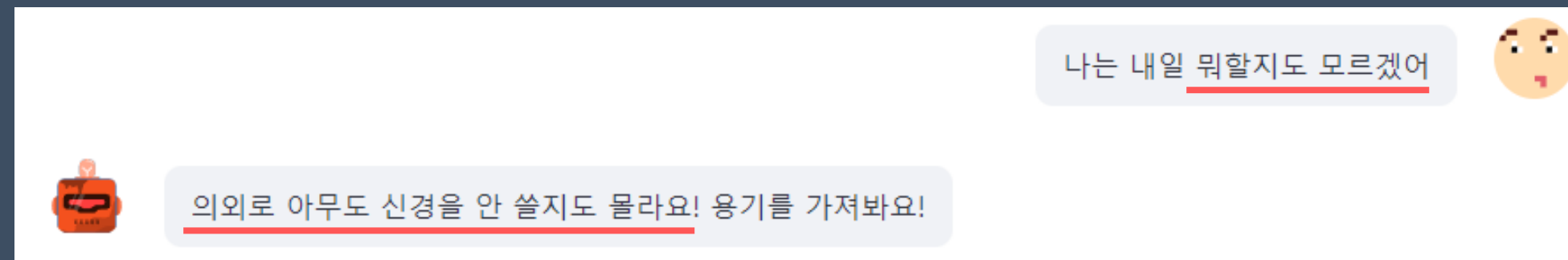
KO-ROBERTA 모델 ✓



[KO-ROBERTA 모델 선정 이유]

- (1) KO-BERT 과 달리 **dynamic masking** 사용
 - dynamic masking : 모델 학습 도중, 고정된 마스킹 문장을 보는 것이 아닌 계속 다른 마스킹 문장을 봄
- (2) 더 큰 batch size, byte 단위 level BPE를 통해 unknown 토큰 없이도, **적당한 크기의 서브워드 사전 학습**을 진행할 수 있음
- (3) 실제 챗봇 구현 시 **조금 더 자연스러운 응답을 선택**하는 모델이었음

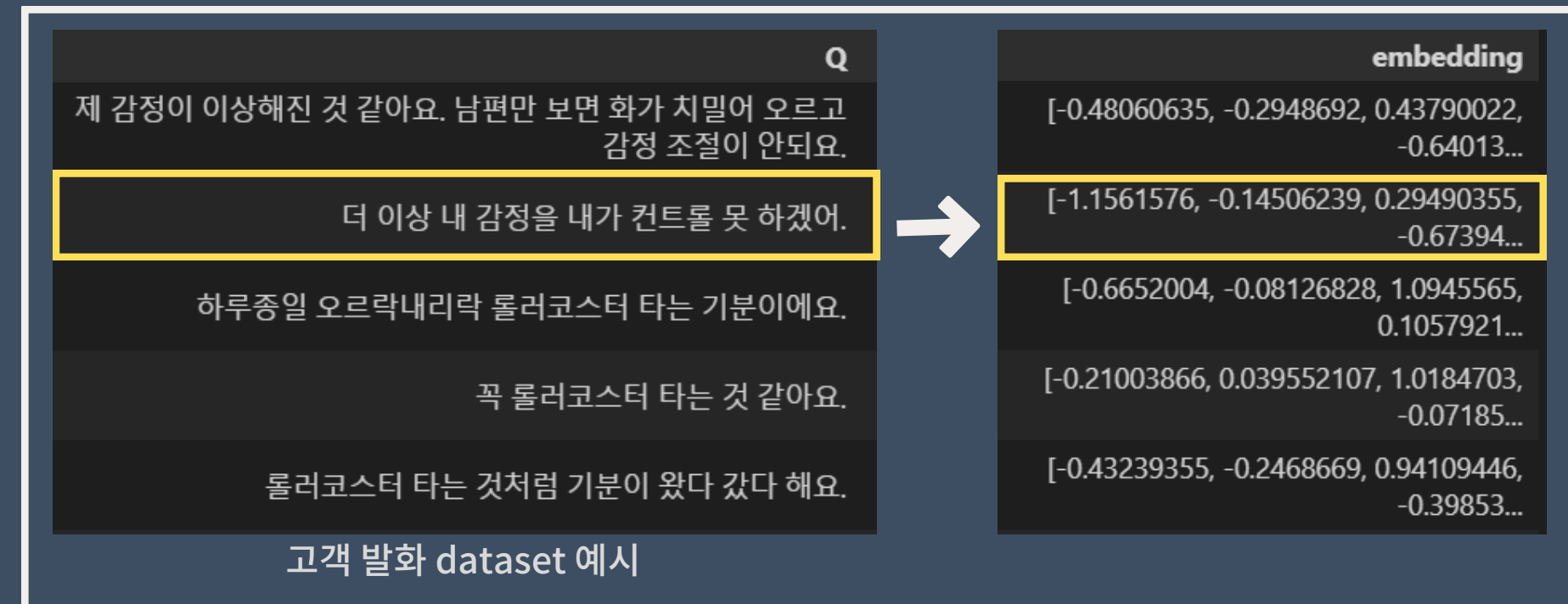
KO-BERT 모델



3. 프로젝트 결과 1) 챗봇 - KO-ROBERTA 모델

[챗봇 진행 과정]

(1) KO-ROBERTA 사전학습 모델 통해 데이터 임베딩



(2) 사용자의 채팅 데이터를 같은 방식으로 임베딩

내 감정을 컨트롤할 수가 없어!!

[-1.151251, 0.14405211, ...]

(3) 코사인 유사도 비교

(4) 가장 유사한 문장의 대응 값을 사용자에게 답변으로 출력

저도 그 기분 이해해요. 많이 힘드시죠?

3. 프로젝트 결과 1) 챗봇

[챗봇 진행 과정]

[앞선 유사도를 통한 선택법 문제]

가진 데이터 중 채팅과 가장 유사한 답변조차 유사도가 많이 낮을 경우, 제대로 된 응답 불가

[우리만의 방식 도입]

코사인 유사도가 가장 큰 데이터의 유사도가 특정 수치를 넘지 않을 땐, 준비해둔 대화의 흐름을 해치지 않는 답변 8개 중 랜덤으로 한가지를 답변

[챗봇 진행과정 고찰]

문제를 근본적으로 해결하기 위한 좋은 방법이라 생각하지 않았으며, 코사인 유사도가 낮은 문장의 경우에는 DB에 쌓아두면서 사용자의 발화는 실제 Q&A 데이터에 직접 추가시키는 방향으로 개발하였으며 이는 성능을 발전을 도모할 수 있을 거라고 생각함

3. 감정 분석 모델 2) 감정 분석 모델링

[개발환경 및 활용한 라이브러리]

ML 모델 RF, CatBoost, Xgboost

DL 모델 LSTM, GRU

Pre-trained 모델 KoBERT

[사용 데이터]

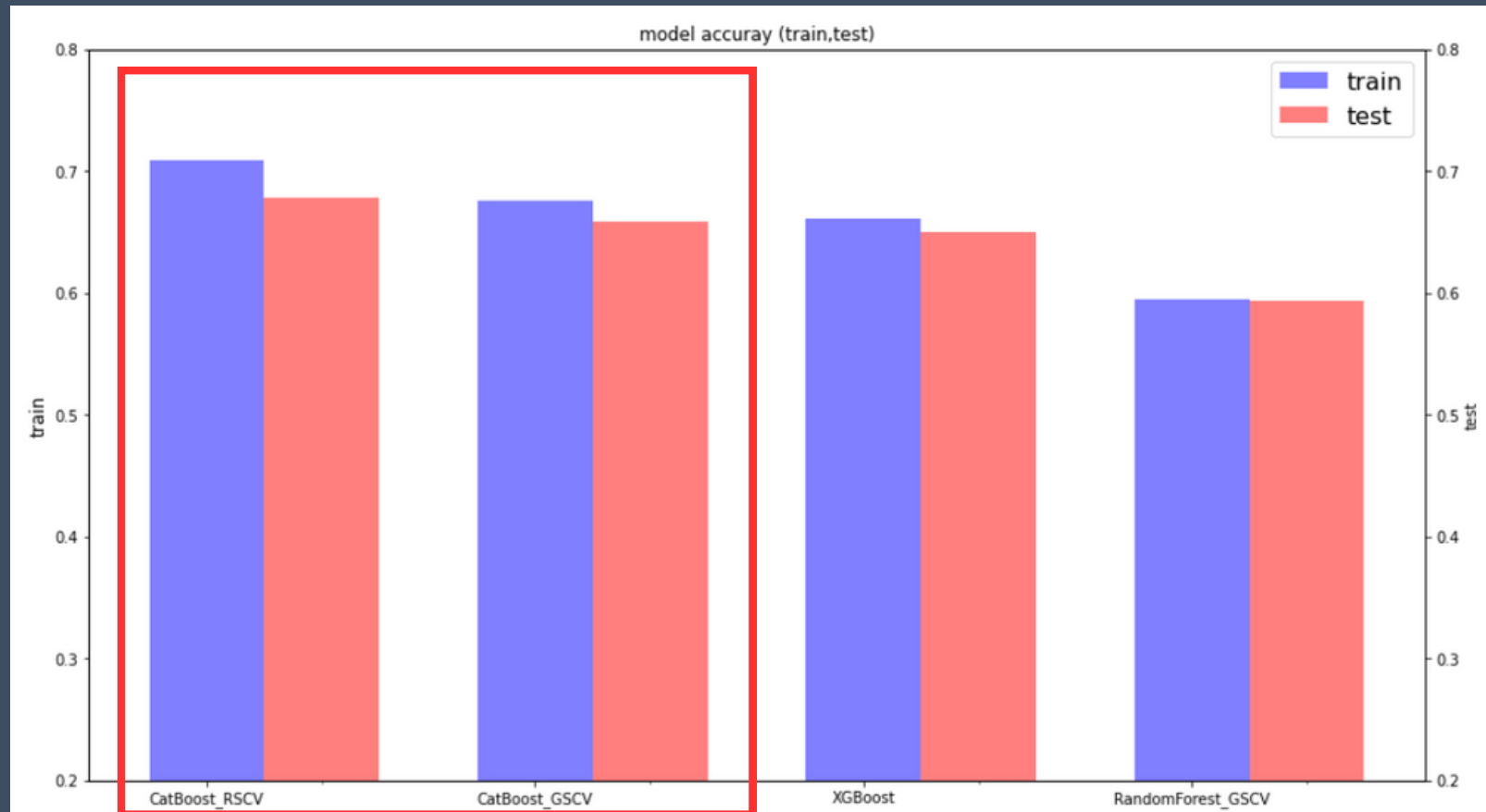
감성대화 말뭉치, 노년층 대상 감성분류 데이터 (AI 허브)

[기본 전처리]

불용어 제거	특정 단어들을 불용어(stopwords) 정의 및 제거하는 함수 구현
품사 추출	OKT로 명사, 동사, 형용사, 부사 품사를 추출 하는 함수 구현 & 단어의 어간 및 단어의 표준화 수행
Tokenizer 단어 인덱싱	Tokenizer 를 활용해 가장 빈도가 높은 2,000개의 단어만 선택하고, 원 -핫인코딩 진행
y 값 원-핫인코딩	학습을 위해 train 데이터에서 하나의 열을 y로 지정하여 y 값도 원-핫인코딩 진행
Zero 패딩	대화 데이터의 99%를 포함하는 60개를 maxlen으로 지정하고, 나머지는 zero padding 사용

3. 프로젝트 결과 2) 감정 분석 모델링 - ML

[ML accuracy 비교]



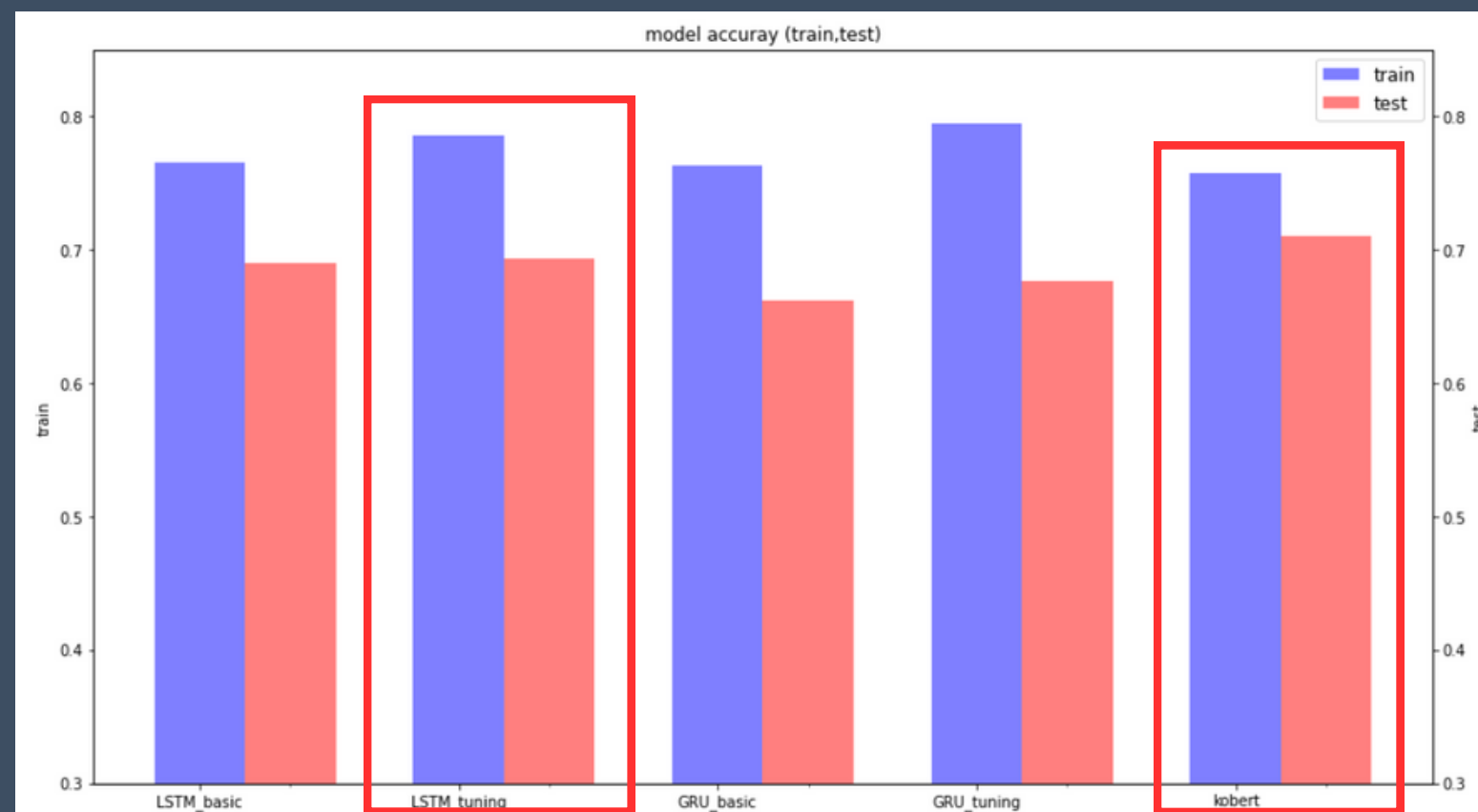
< 시도 모델 中 상위 4개의 모델 >

Catboost, XGboost, SVM, RandomForest
Decision Tree, Logistic regression, LGBM, GBM

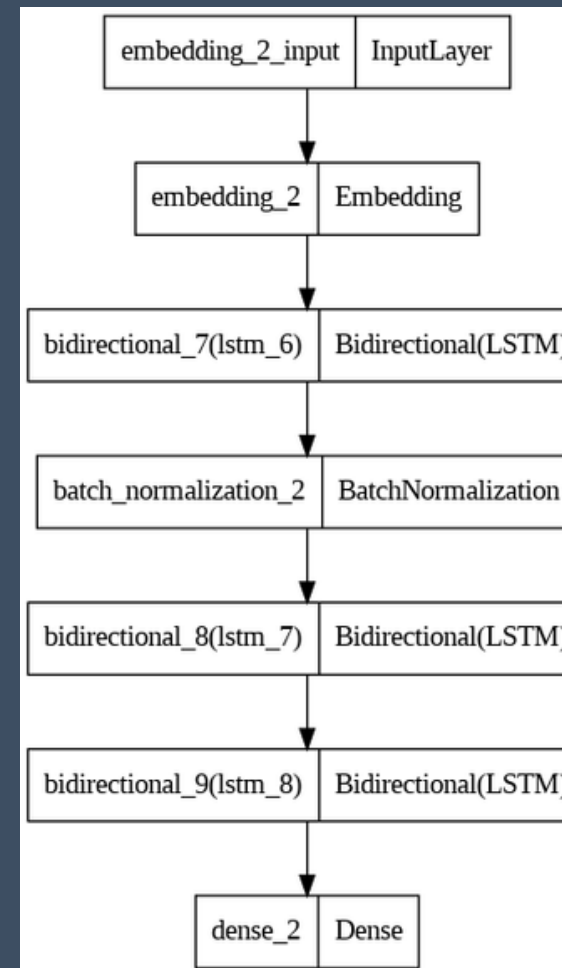
튜닝 모델	Best Model의 parameter	결과
Catboost	random_strength = 30 iterations = 1380 n_estimators = 800 max_deth = 6 learning_rate = 0.1 colsample_bylevel = 0.6	BEST train score : 0.7141 test score : 0.6754

3. 프로젝트 결과 2) 감정 분석 모델링 - LSTM

[DL accuracy 비교] DL 모델 중 LSTM과 KOBERT를 중점적으로 시도



< DL 모델 비교 >

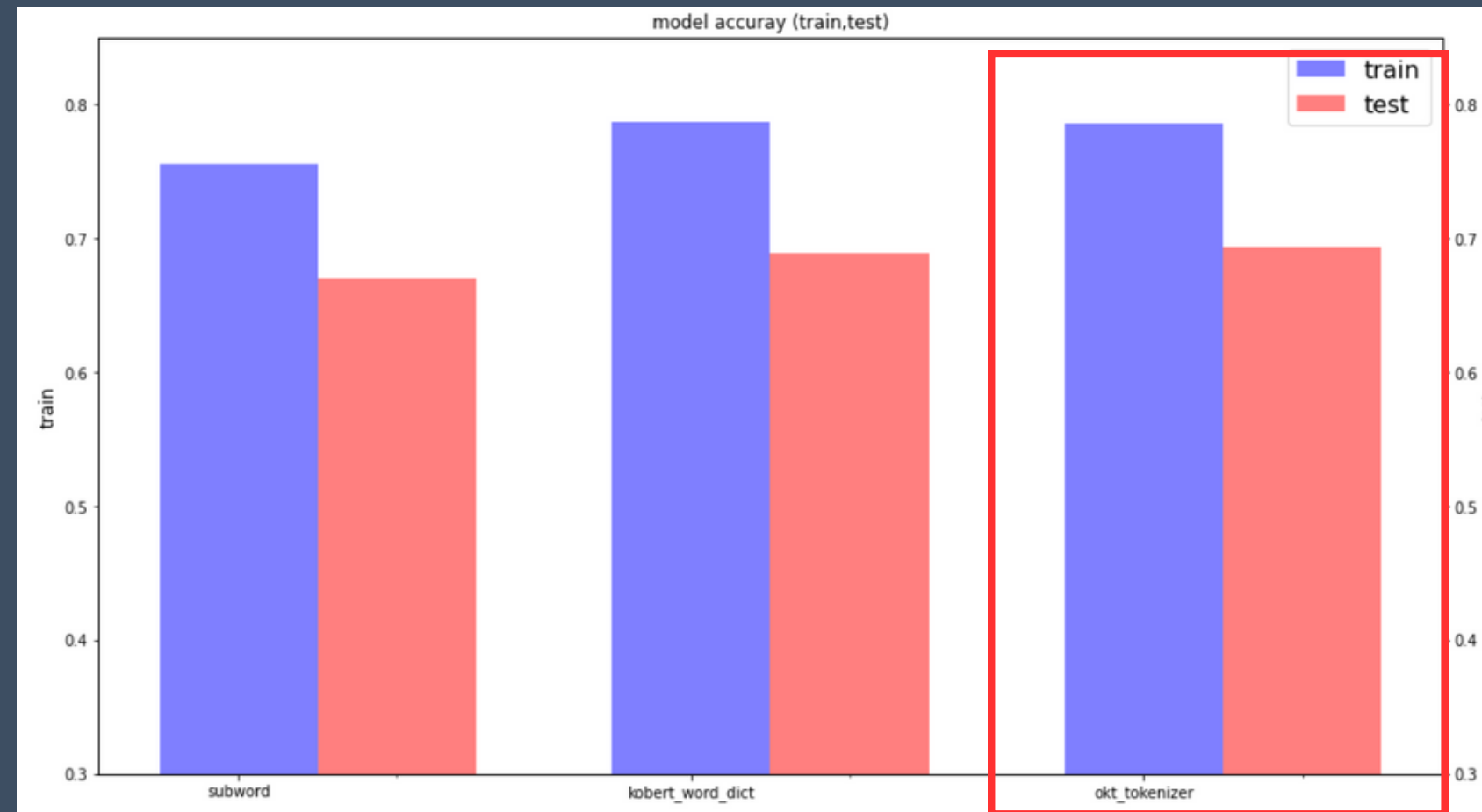


< LSTM BEST MODEL >

Best Params	결과
optimizer = Adam units = 16 ~ 32 layer = 3 ~ 4 Dropout = 0.5 Batchnormalization O	train_acc = 0.785 test_acc = 0.693 param수가 너무 많아지거나 학습 데이터 수가 많을 경우 (layer 개수 및 units 수를 늘려도) 성능이 오르지 않음을 확인 따라서, DL 모델의 경우 학습 데이터 마다 적당한 param 수를 아는 것이 필요

3. 프로젝트 결과 2) 감정 분석 모델링 - LSTM

[LSTM Best Model의 토큰화 변경 결과]



< LSTM Best Model의 토큰화 변경 비교 >

전처리 종류		train	test
기본 토큰화 (okt tokenizer)	+ 원핫 인코딩 + 128개로 임베딩	0.785	0.693
subword 토큰화		0.7554	0.6703
kobert_word_dict 토큰화		0.7866	0.6883

결과 : 토큰화 별로 학습된 모델 결과 비슷

3. 프로젝트 결과 2) 감정 분석 모델링 - KOBERT

BERT란?

2018년에 등장한 State of the Art

- KoBERT는 보다 한국어를 집중적으로 다룬 BERT
- 많은 BERT 모델 중에서도 다중 분류가 가능한 강점을 가짐



3. 프로젝트 결과 2) 감정 분석 모델링 - KOBERT

J팀 BERT 모델 - '제우' -> 59

M팀 BERT 모델 - '제우' -> 355

J팀 모델로 M팀의 tokenizer를 사용하면...?

3. 프로젝트 결과 2) 감정 분석 모델링 - KOBERT

[KoBERT Tokenizer 소개]

KoBERT Tokenizer – Sentencepiece

- (1) PRE-TOKENIZATION 을 필요로 하지 않음
 - 기존의 TOKENIZER와 다르게 띄어쓰기를 하나의 CHARACTER로 취급
- (2) CHARACTER-COVERAGE 설정이 가능하다
 - 너무 많은 CHARACTER를 다루어야 할 때 효과적으로 다룸
- (3) 속도가 빠르다
 - 알고리즘이 굉장히 최적화 되어 있음

3. 프로젝트 결과 2) 감정 분석 모델링 - KOBERT

[KoBERT 결과]

Best Params	결과
<pre>[] # Setting parameters max_len = 64 batch_size = 64 warmup_ratio = 0.1 num_epochs = 10 max_grad_norm = 1 log_interval = 200 learning_rate = 5e-5</pre>	<p>train acc : 0.7092881045836209</p> <p>test acc : 0.7021824039229103</p>

3. 프로젝트 결과 2) 감정 분석 모델링 - 최종 모델 선정

[모델 비교]

모델	Train / Test ACC
CatBoost	Train ACC : 0.714 Test ACC : 0.675
LSTM	Train ACC : 0.785 Test ACC : 0.693
KoBERT	Train ACC : 0.709 Test ACC : 0.702

최종 모델 : CatBoost

- LSTM : 다양한 튜닝 방법을 시도했지만 과적합으로 인해 미선택
- KoBERT : 성능은 제일 좋았지만, 모델 용량이 크다는 단점으로 웹 업로드 과정에서 서버 내 메모리 제한 이슈 발생
- streamlit 툴로 짜여진 챗봇은 hugging face api를 통해 실행

=> 과적합이 덜하고, 제약사항이 없는 CatBoost를 웹 사용모델로 선정

3. 프로젝트 결과 3) 노래 추천 알고리즘



3. 프로젝트 결과 4) 데이터베이스 연동 및 활용

123 emotion4	ABC song_sim	ABC song_dif	123 user_preference
0.162478268	백예린 - Antifreeze		1
0.162478268	백예린 - Antifreeze		0
0.1369036904	검정치마 - 나랑 아니면		1
0.1369036904	검정치마 - 나랑 아니면		0
0.1369036904		프롬 - 서로의 조각	1
0.1369036904		프롬 - 서로의 조각	0

- 사용자가 **좋아요/싫어요** 눌렀을때 업데이트 되는 테이블
- 테이블 데이터를 통해 비슷한 감정을 지닌 유저 간 노래를 추천하는 **새로운 알고리즘 구현 가능**
-> 지속적으로 같은 감정을 지닌 사용자가 좋아했던 음악을 추천하여 **조금 더 다양한 음악을 추천** 할 수 있음.

ABC sentence	ABC time
따봉기능	2023-03-08 05:21:19.400747+09:00
따봉기능	2023-03-08 05:22:12.598725+09:00
나는 우산을 가진 힙합	2023-03-08 05:23:22.092190+09:00
점수 잘주세요	2023-03-08 05:34:48.811144+09:00
점수 잘 주시면 잘해드릴게요	2023-03-08 05:34:57.659126+09:00

- 추후 서비스의 발전을 위해 데이터베이스를 웹에 **연동**
- 코사인 유사도가 특정 수치 이하인 텍스트들을 한 테이블에 적용
-> **챗봇 데이터를 업데이트** 시키면서, 챗봇의 성능을 높임

3. 프로젝트 결과 5) 결과 시연

<https://sangwookwoo-music-resting-place-chat-botchatbot-k1rq1h.streamlit.app/>

4. 한계점 및 발전가능성

[한계점]

1. 음악에 대한 감성분석을 **가사**로만 **진행**하여 멜로디 등 다른 요소를 반영 할 수 없었음.
2. 모델의 정확도가 예상보다 높지 않았음.
3. Torch 기반 사전학습 부족으로 KoBert 모델 튜닝 시도 하지 못함.

[발전가능성]

1. 챗봇이 대답하지 못한 내용을 DB에 저장하여 적절한 **응답 데이터**를 **추가**함으로써 더 정확한 챗봇의 구현 가능
2. 다른 음악 어플과 차별성을 주기 위해, 노래 좋아요, 싫어요 데이터를 쌓아 **개인 맞춤** 추천 가능

5. 출처

- SDSN 2022 세계 행복 지수
- Bert와 gpt의 차이: scatter lab tech :<https://tech.scatterlab.co.kr/transformer-review/>
- 웰니스 심리 상담 데이터셋(AI허브)
- 감성분석 말뭉치
- 챗봇 이미지 : https://www.flaticon.com/kr/free-icon/chatbot_2040946
- 멜론 : <https://www.melon.com/>
- 지니 : <https://www.genie.co.kr/>
- YouTubeMusic : <https://music.youtube.com>

6. Q&A

