

종합 설계

EDA X

2. 임베딩 레이어를 '사전학습된 임베딩 벡터'를 사용할 것인가?

- 사용 - FastText-KR
- 미사용 - 기존의 lstm 임베딩 레이어로 구성

3. 학습

3-1. 학습모델

- LSTM 모델
- Bi-LSTM
- Bi-LSTM + attention.....?
- Transformer -Encoder

3-2. 모델 구성

- l2 정규화(과적합 방지) - LSTM에서 사용
- lr epoch단위로 나눠서 할 것인지(과적합 방지) - LSTM에서 사용

-epoch 수

4. 문장 예측

- 입력된 문장의 단어를 추출해서 예측?
- 입력된 문장의 단어 모두를 예측?

EDA 사용의 문제점

지나치게 높은 모델 평가
= 데이터 불균형(?) 문제

해결방법

1. 훈련 세트에서만 EDA 사용

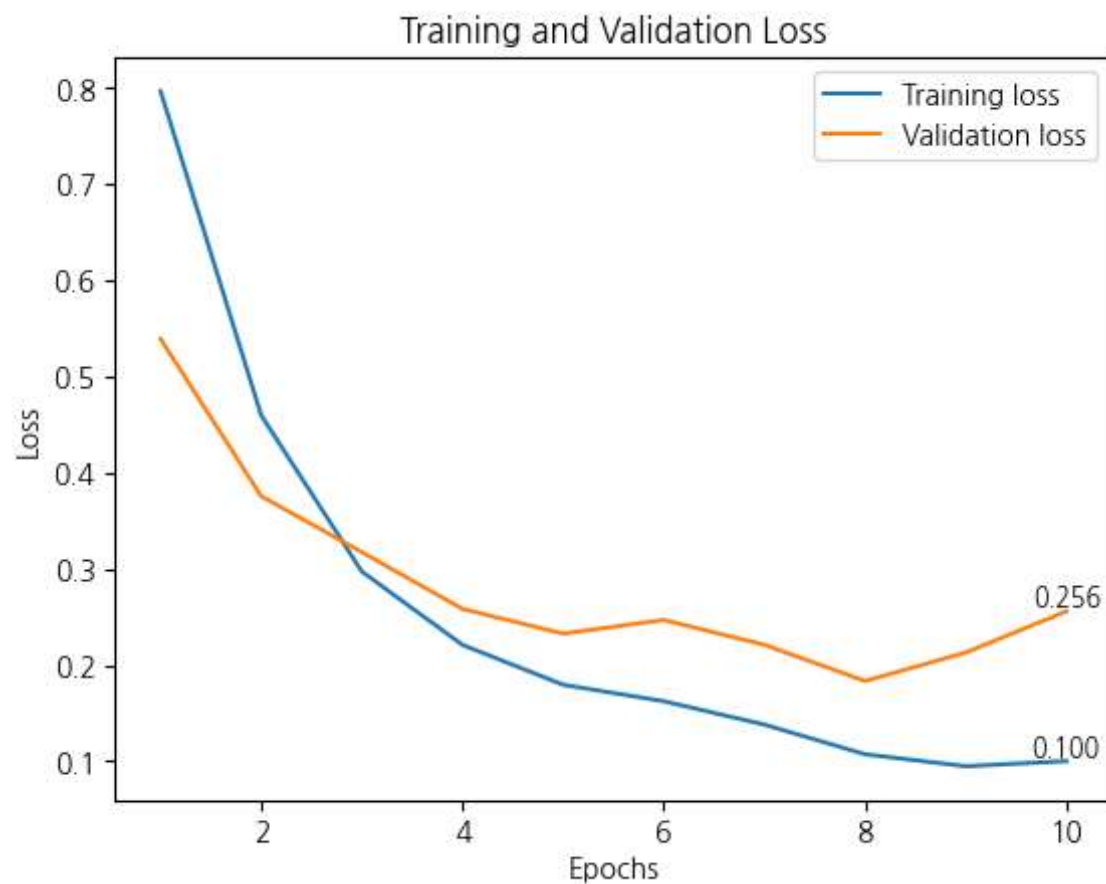
2. 유사한 데이터 포인트가 훈련과 테스트 데이터 셋
모두에 나타나지 않게

➔ sklearn의 GroupShuffleSplit 사용

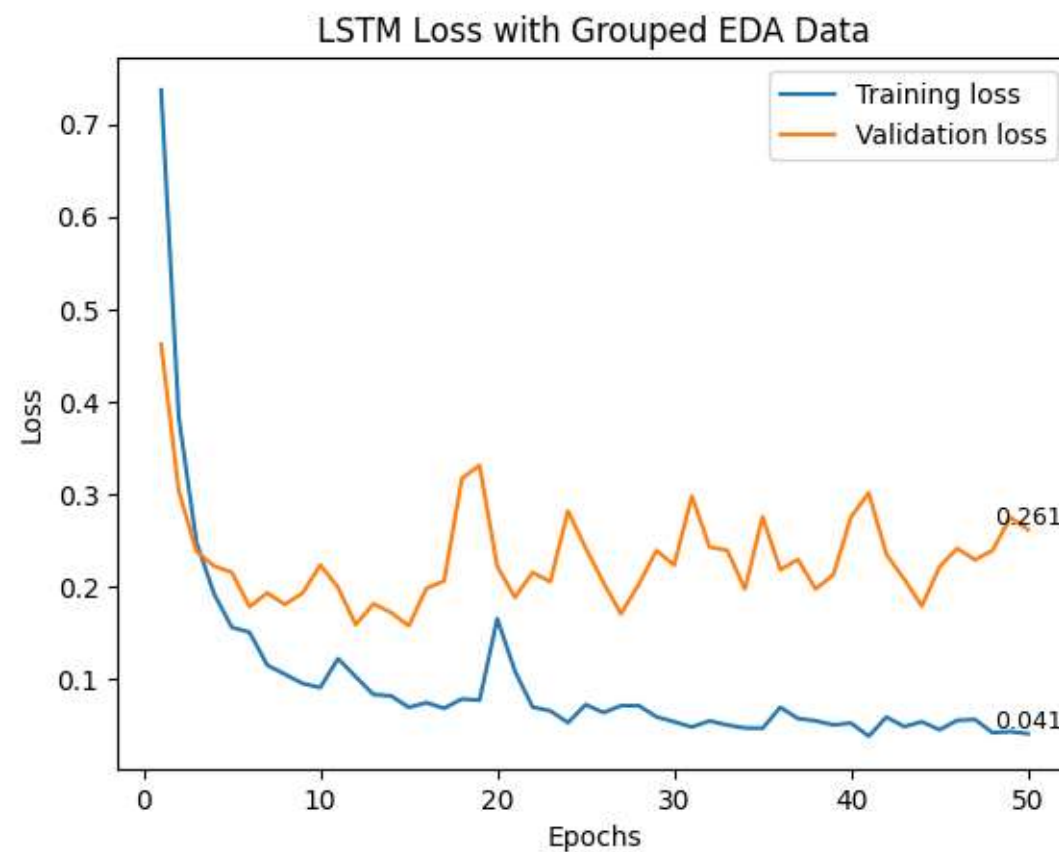
유사한 데이터 포인트의 중복을 방지하기 위해 sklearn의 GroupShuffleSplit을 사용하여 동일 그룹의 데이터가 훈련, 검증, 테스트 세트 중 하나에만 속하도록 분할합니다.

Loss

Base Data
+ EDA

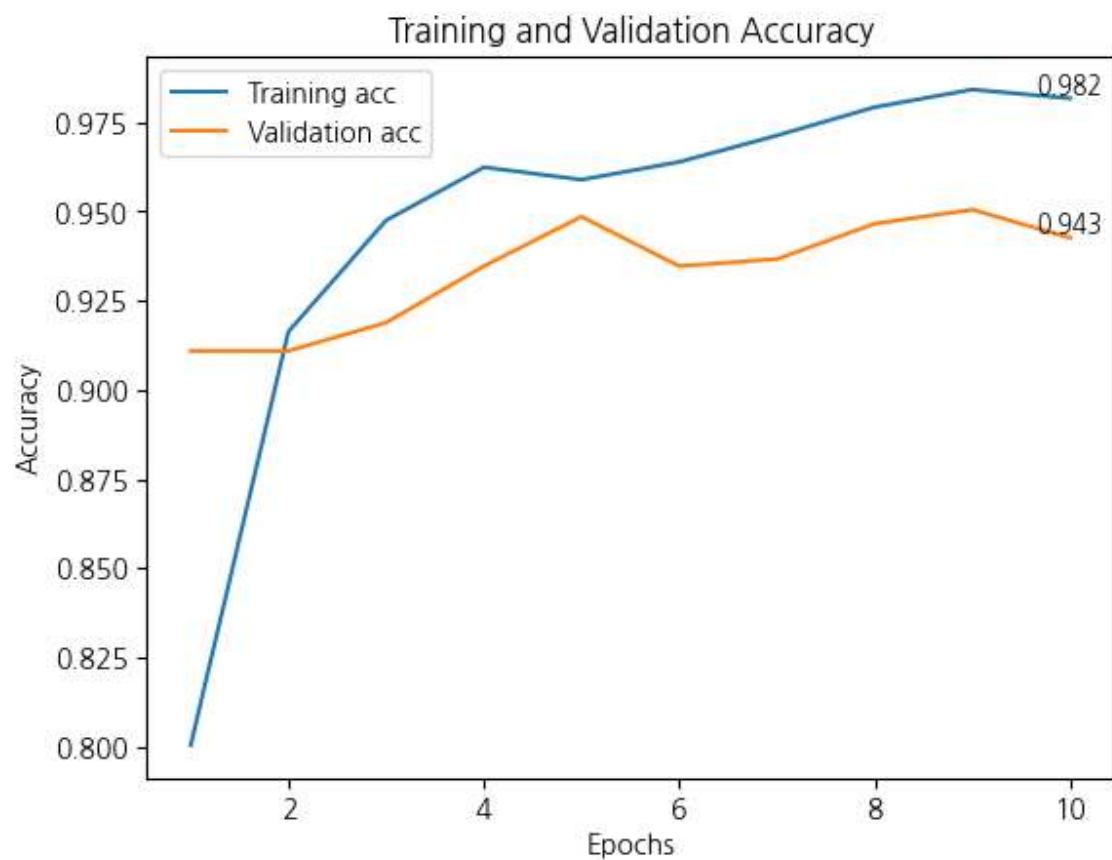


Base Data
+ group EDA

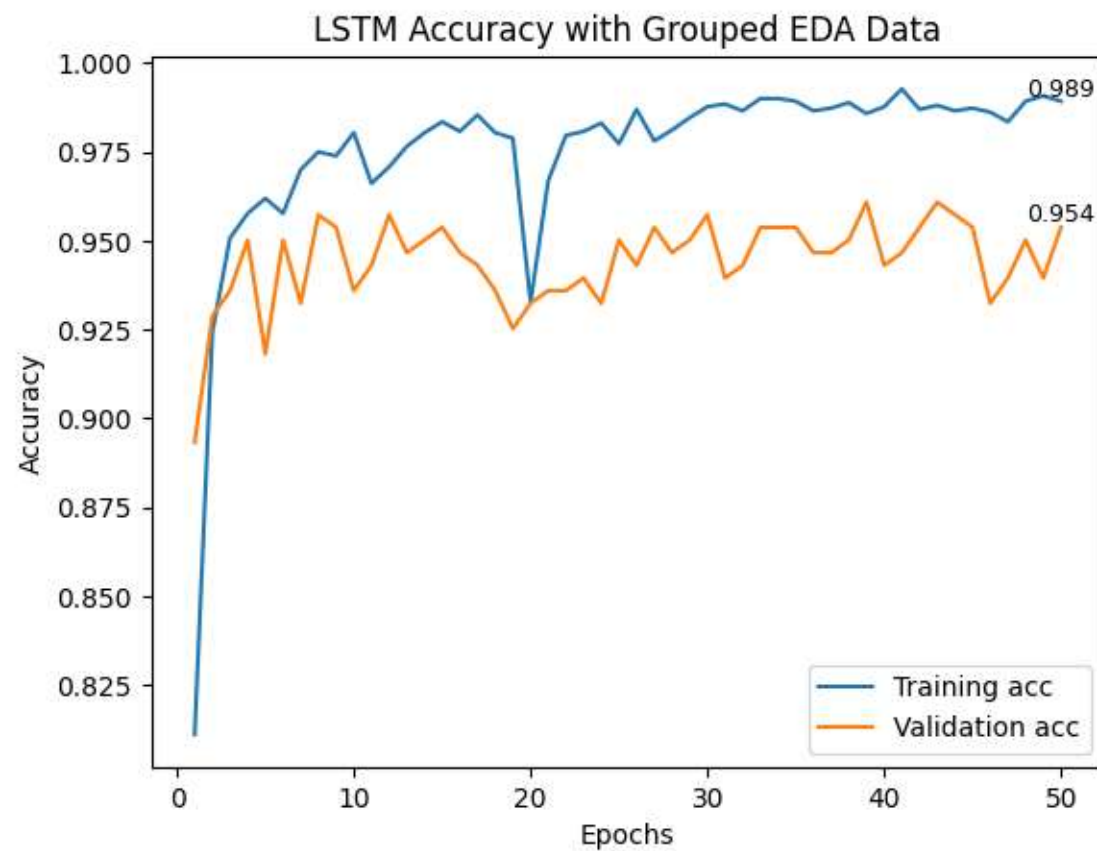


Acc

Base Data
+ EDA



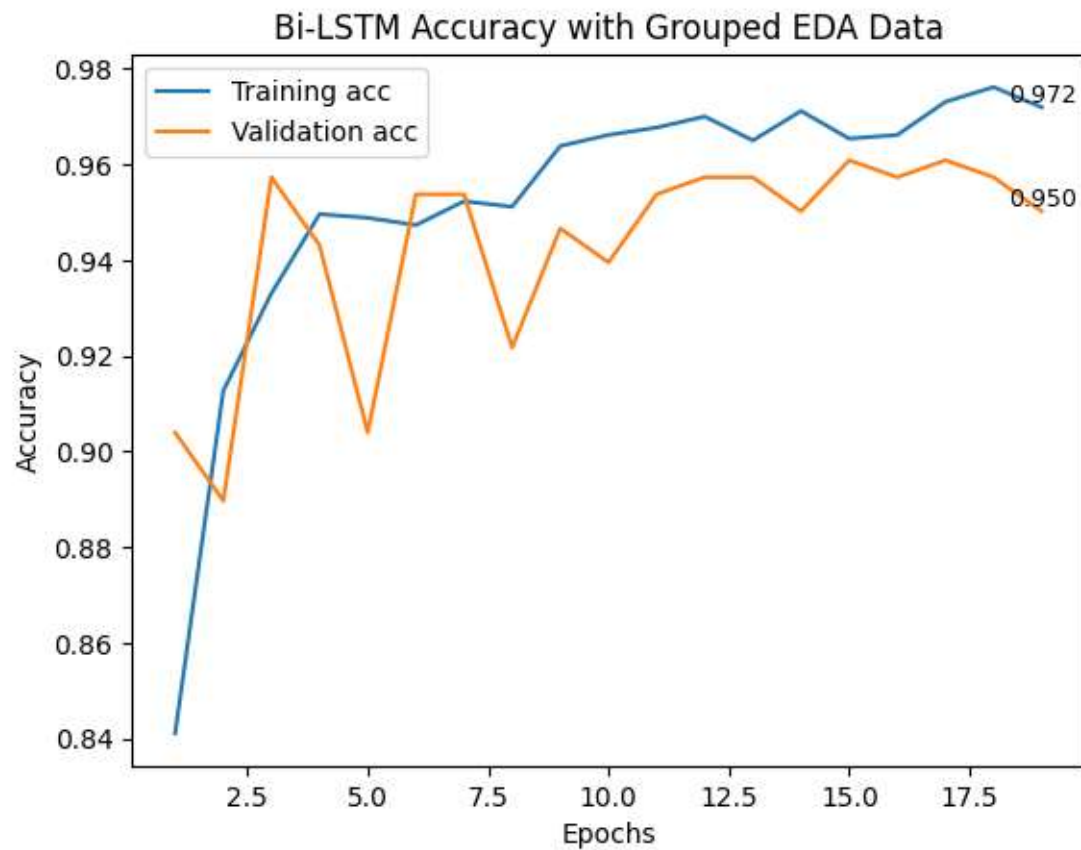
Base Data
+ group EDA



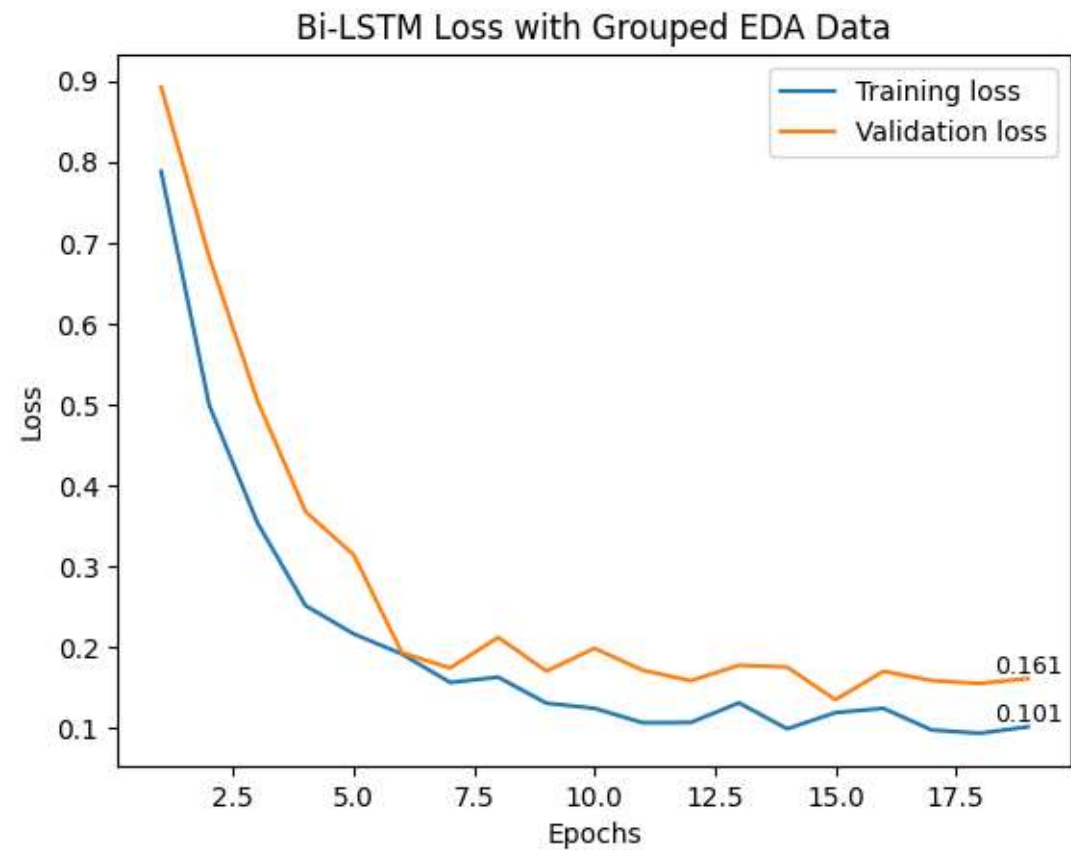
Bi-LSTM 사용

Bi-LSTM

Acc



Loss



Bi-LSTM - 예측

입력문장: <친구>

>> 89.78% 정도로 E 유형입니다.

입력문장: <혼자>

>> 95.73% 정도로 I 유형입니다.

입력문장: <사람>

>> 53.86% 정도로 E 유형입니다.

Bi-LSTM - 예측

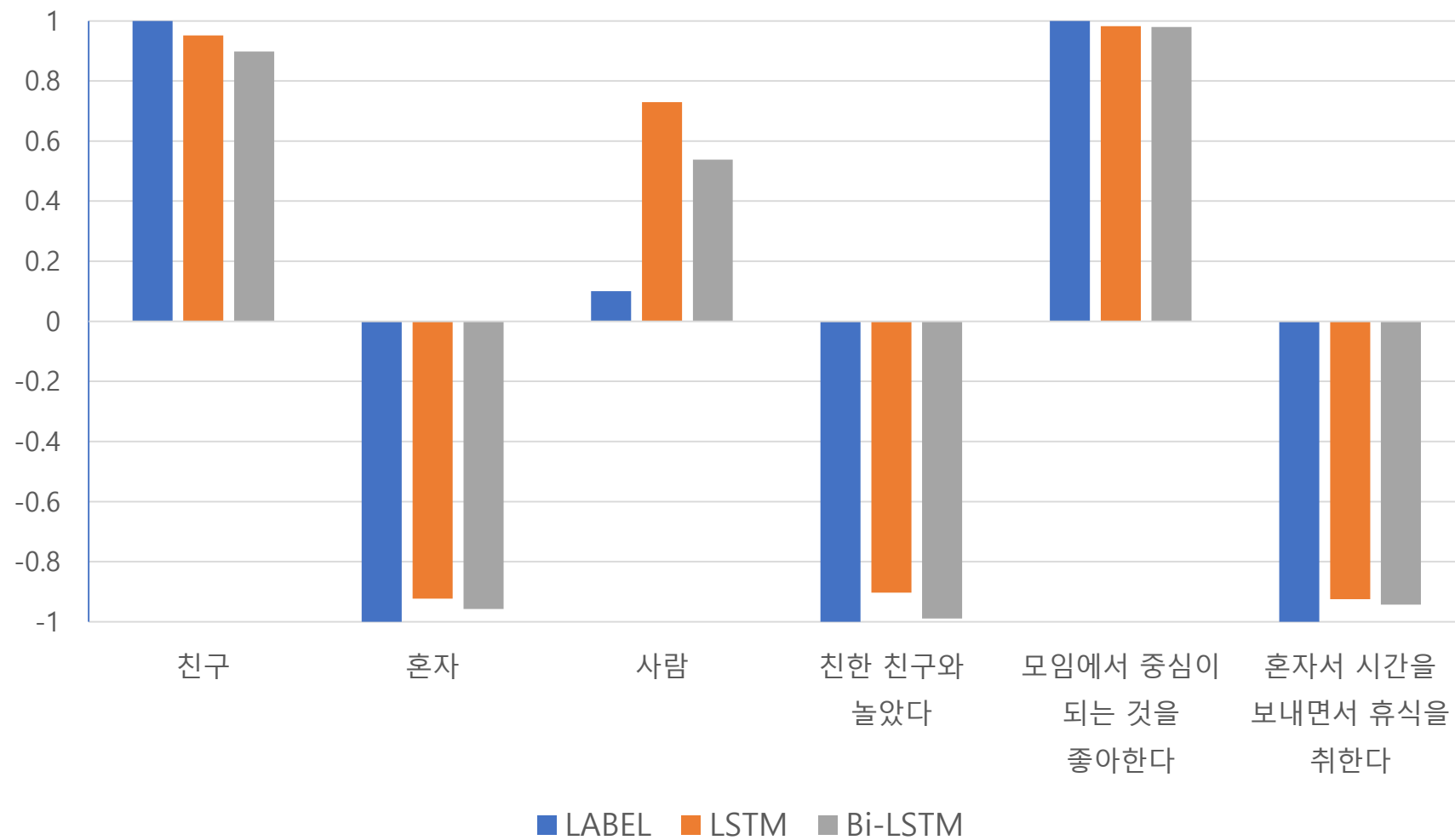
입력문장: <친구>

>> 89.78% 정도로 E 유형입니다.

입력문장: <친한 친구와 놀았다.>

>> 98.91% 정도로 I 유형입니다.

단어 예측



LSTM 모델의 전반적인 문제점

1. EDA 사용의 문제점

1-1. 유사한 데이터 포인트의 중복

1-2. EDA가 적용된 문장 데이터셋의 품질 문제

2. 너무 적은 수의 문장 데이터 셋

3. <MBTI 예측시> OOV 처리문제

유사한 데이터 포인트의 중복을 방지하기 위해 sklearn의 GroupShuffleSplit을 사용하여 동일 그룹의 데이터가 훈련, 검증, 테스트 세트 중 하나에만 속하도록 분할합니다.

Dataset

write & create

Question
Data

Answer Data

Label

Answer Analysis



KoNLPy
(Morpheme Analysis)

Word Embedding Vector
values for each Key-word

Sentiment
Analysis

LSTM

Frequency

Modifier
(adverbs)

MBTI prediction Model

LSTM

Bi-LSTM
+ Attention

Transformer
- Encoder

Data Augmentation

UI

Bot

E/I Question

Answer Analysis
Model

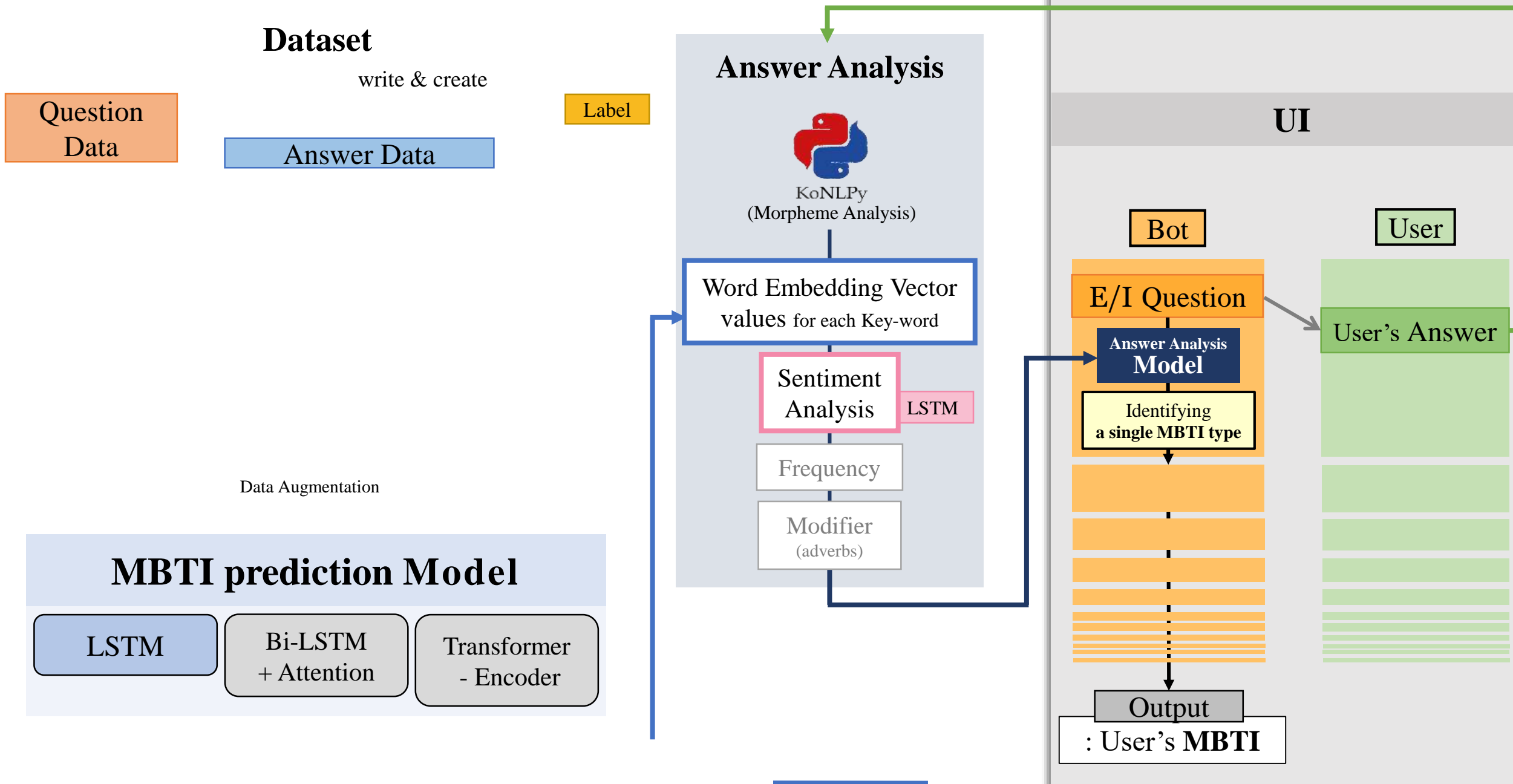
Identifying
a single MBTI type

Output

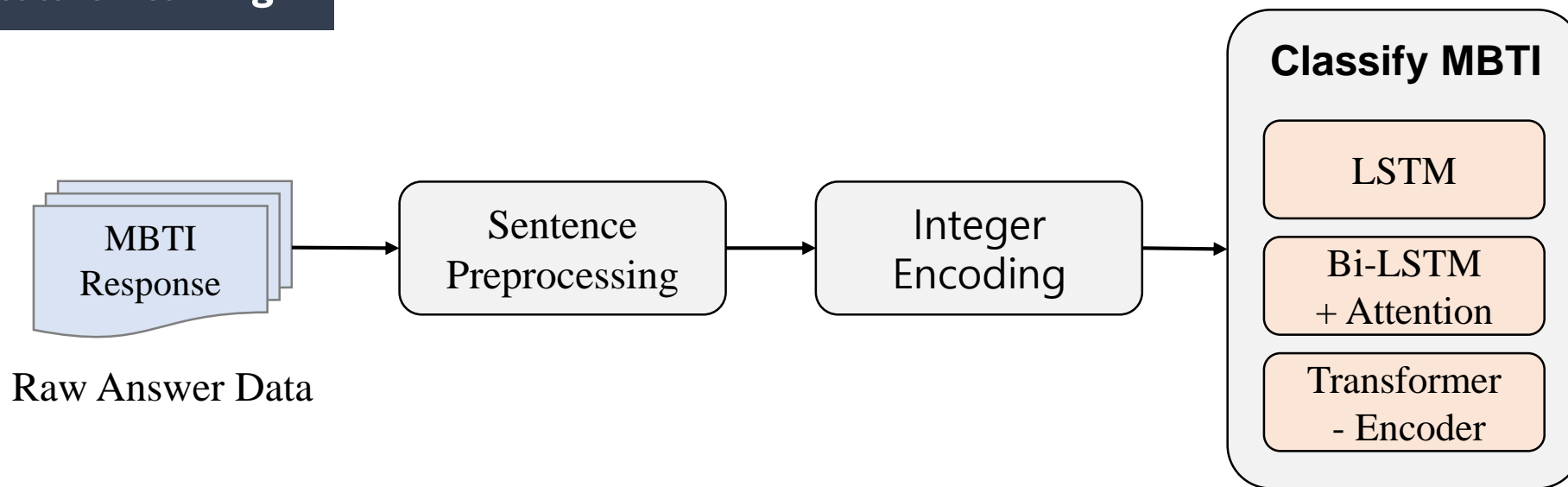
: User's MBTI

User

User's Answer



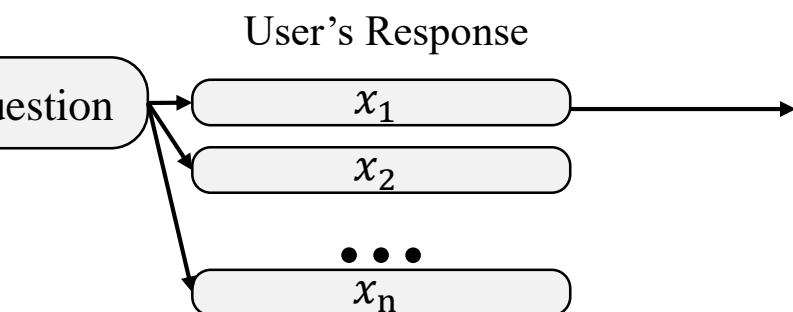
MBTI Feature Learning



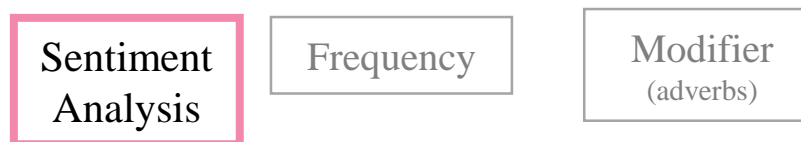
MBTI Response Analysis

UI

Input Data



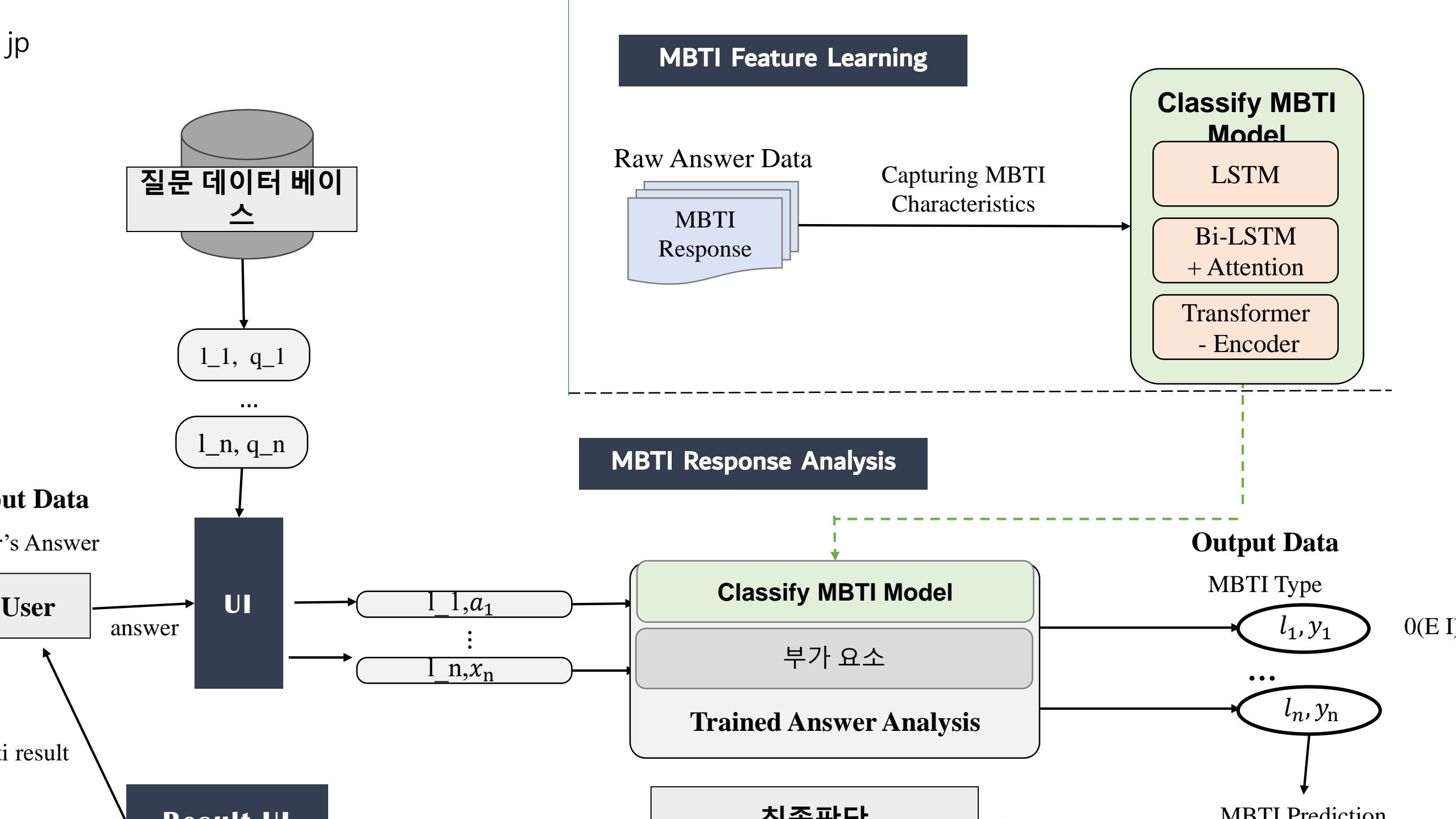
Input Data



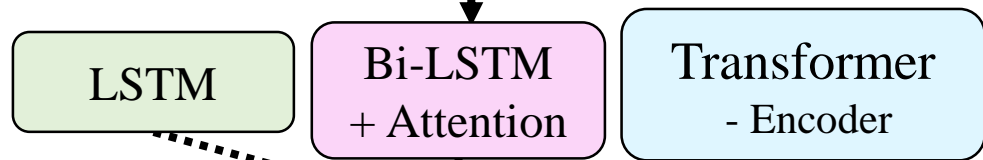
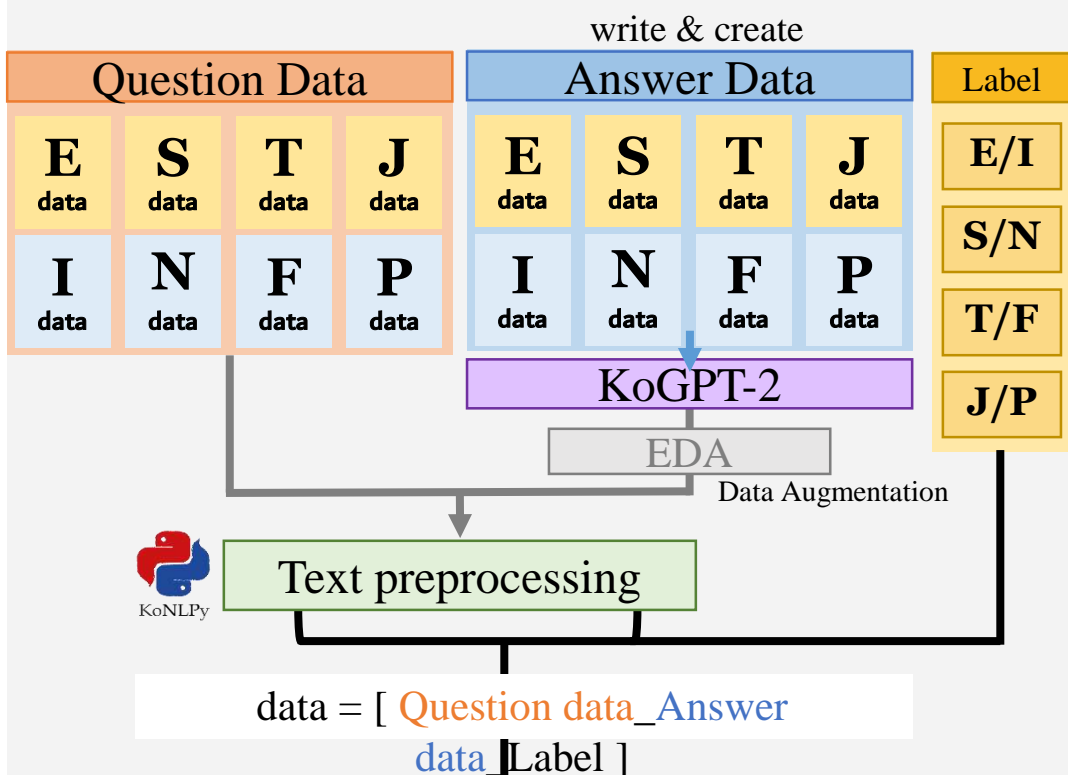
Output Data

MBTI Type





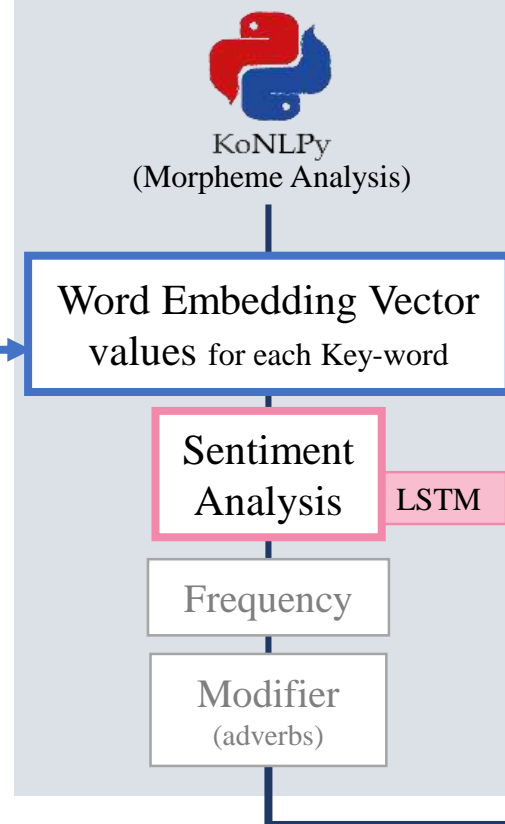
Dataset



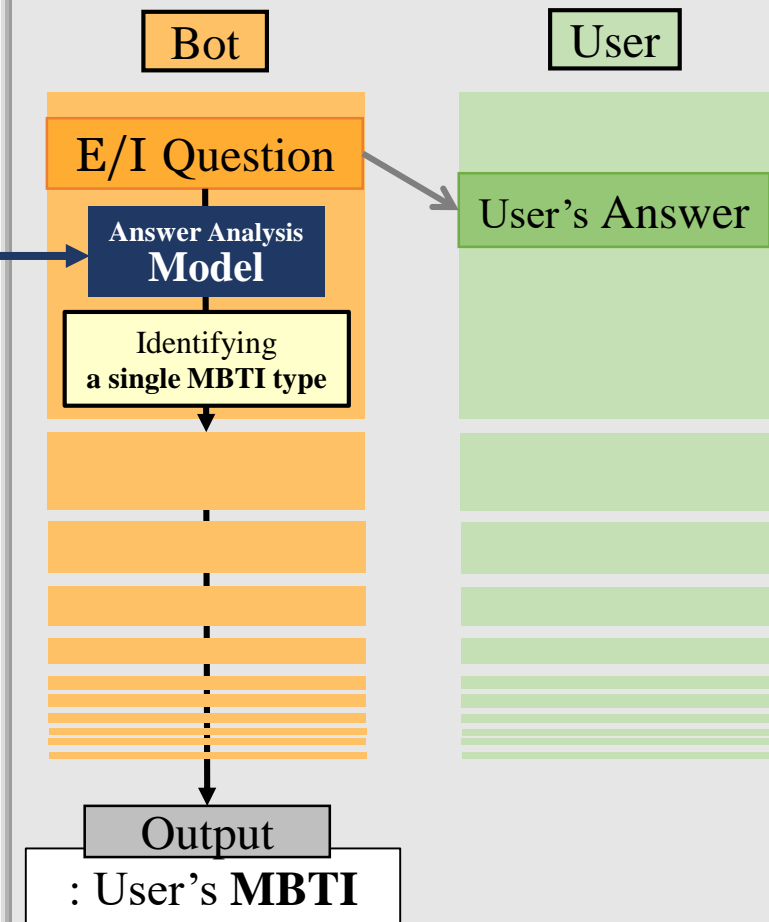
4-BINARY CLASSIFIERS Model

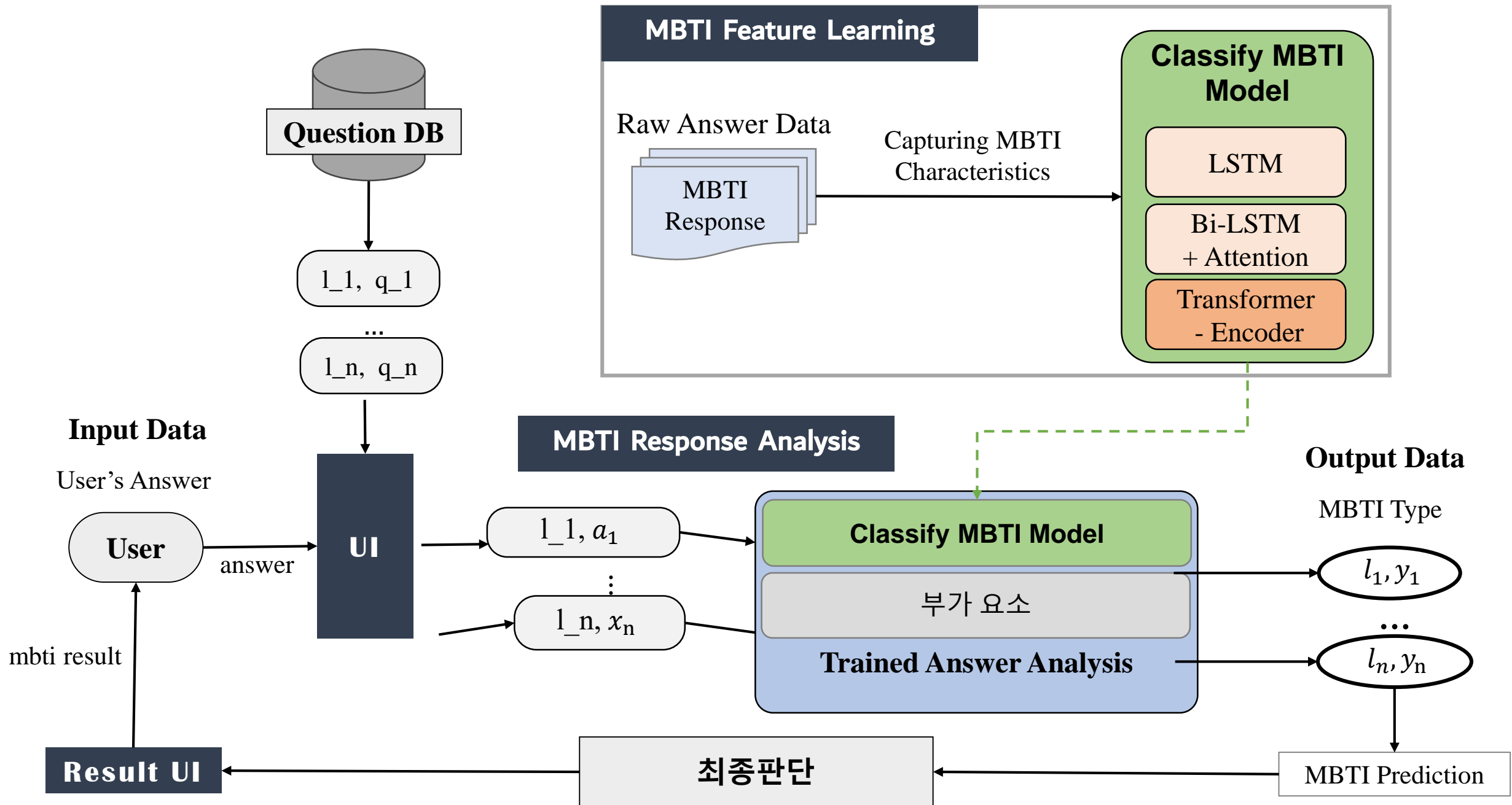


Answer Analysis Model

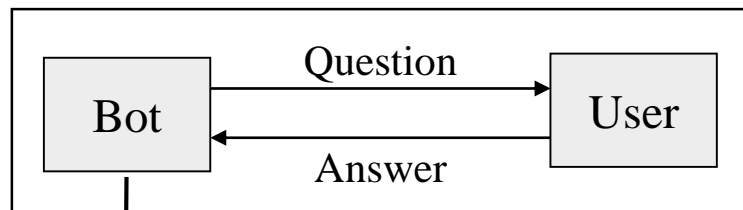


UI





UI



MBTI Feature Learning

Raw Answer Data



Capturing MBTI Characteristics

Classify MBTI

LSTM

Bi-LSTM
+ Attention

Transformer
- Encoder

MBTI Response Analysis

Input Data

Bot's Question

User's Response

q_1

x_1

...

x_n

Answer Analysis

Decision Model

Sentiment
Analysis

Frequency

Modifier
(adverbs)

Output Data

MBTI Type

y_1

...

y_n

MBTI Prediction

MBTI Feature Learning

Input Data

Raw Answer Data

MBTI
Response

Capturing MBTI
Characteristics

Classify MBTI

LSTM

Bi-LSTM
+ Attention

Transformer
- Encoder

UI

색깔 조합

질문!!!!

그러면 좀 큰 틀의 구조도를 보여주고

저 위에 만들어 놓은거 각부분 설명할때 써도 되나?

++ ??? 우키오빠한테 물어볼거!!

그럼 개괄식으로 이렇게 작성하면 세부적인 부분은 어떻게
설명을 하는가?

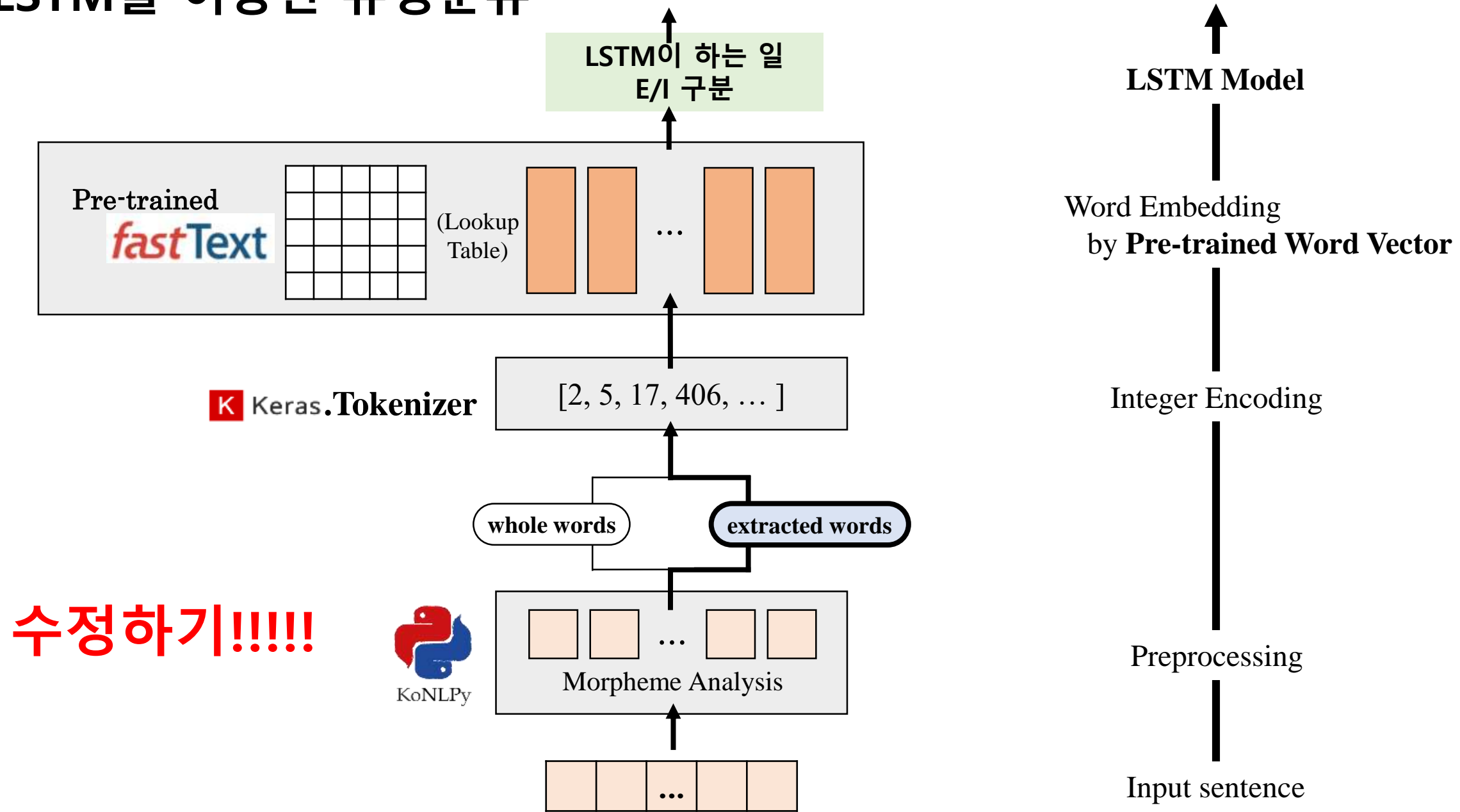
[교수님 질문]

fasttext 의 입력은 word일텐데 나는 정수 인코딩을 넣어버리면 ,, 그게 정확히 사전학습된 모델의 벡터를 잘 가져오는 게 맞는것인가?

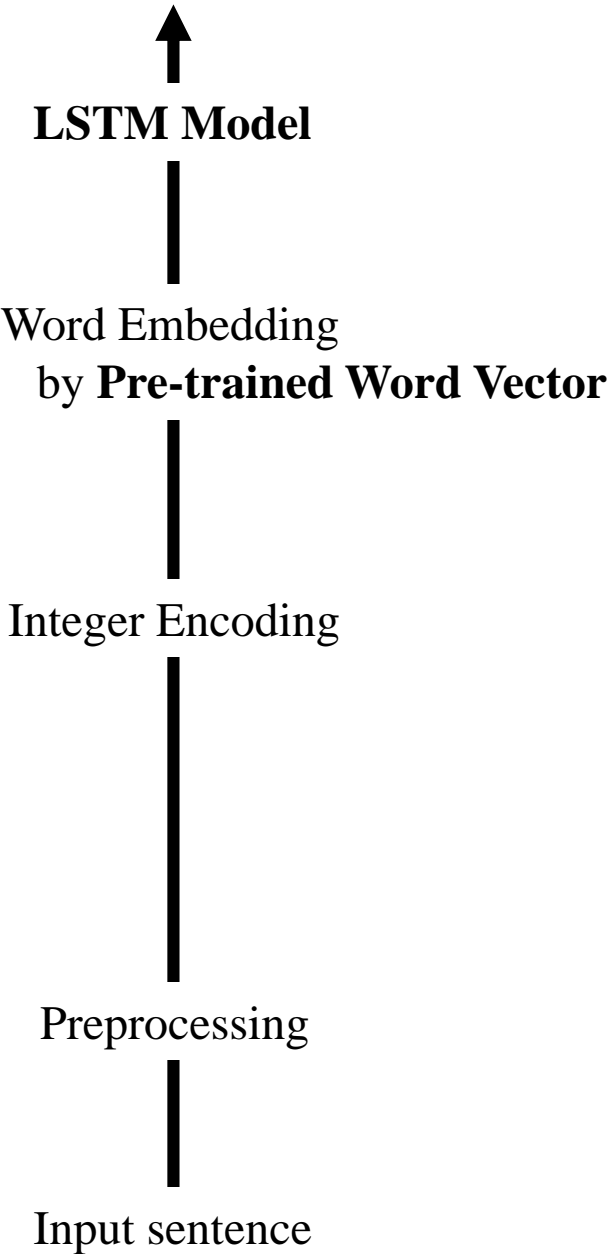
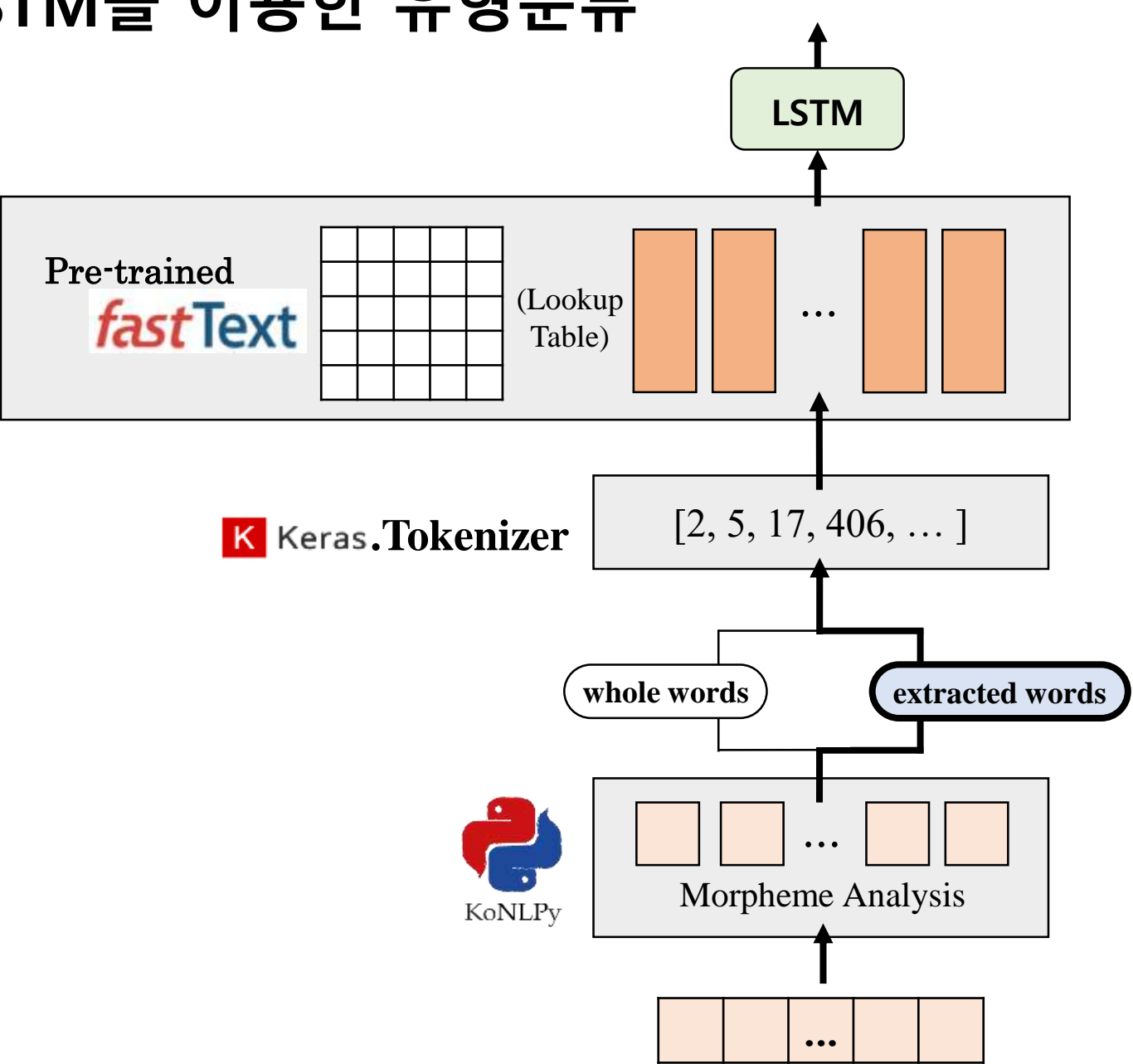
2. 구조도 수정
더 개괄식으로 작성하기

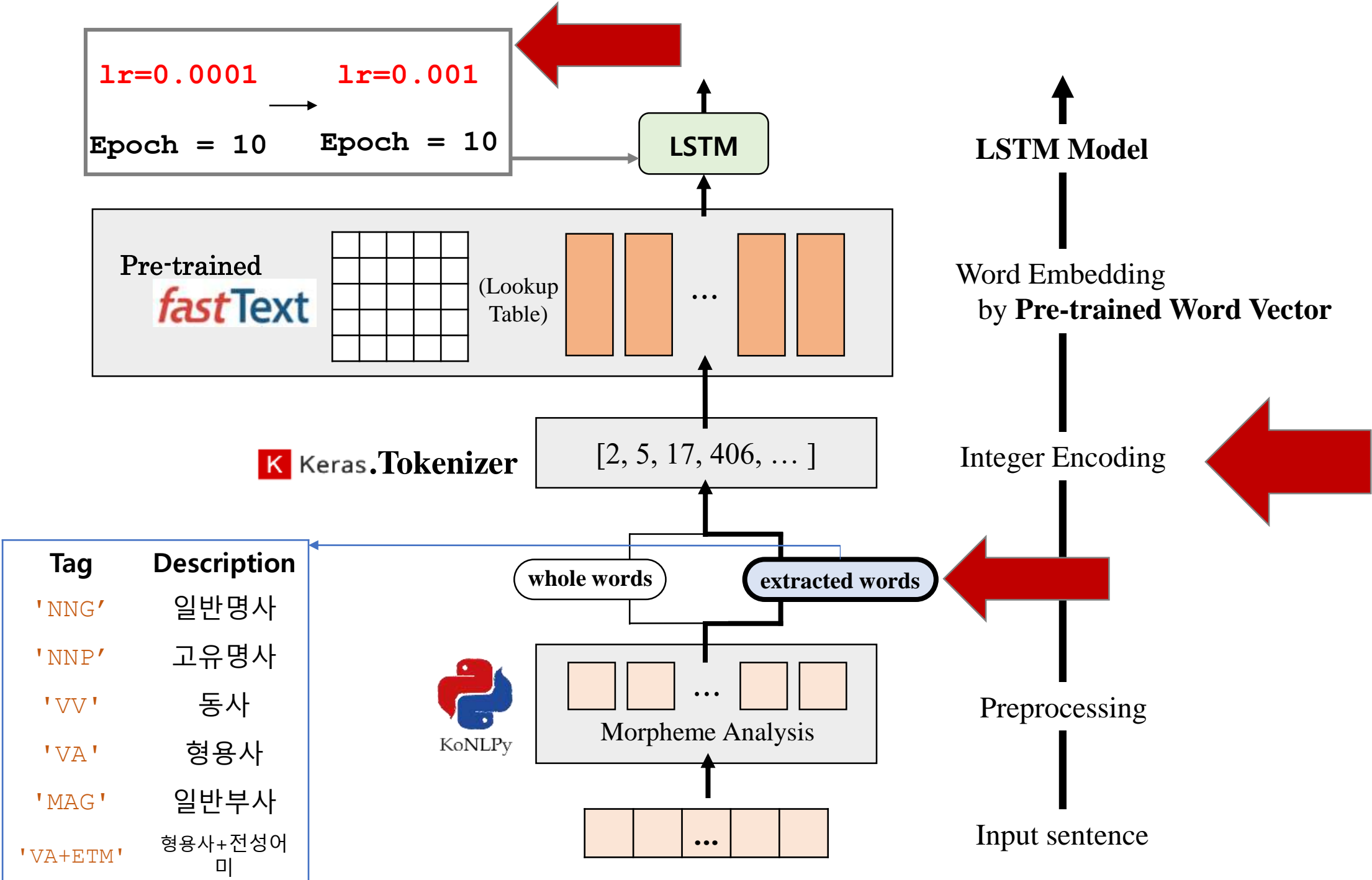
3. Weighted Sum
: 가중치의 곱을 더하여
일정 비율만큼 각 모델들을 사용해주는 것

LSTM을 이용한 유형분류



LSTM을 이용한 유형분류

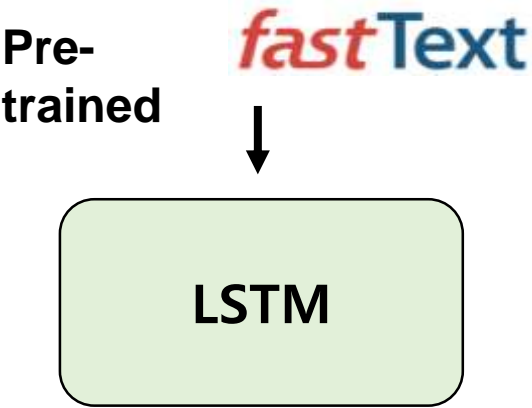




성능 비교

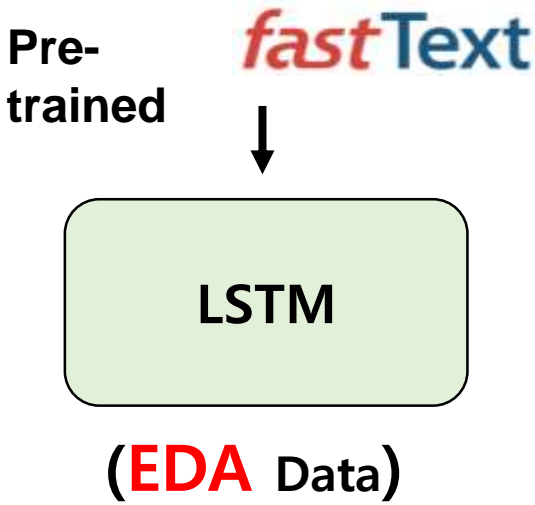
1.

Data	약 1000개
Embedding vector	사전학습 모델
학습 모델	LSTM



2.

Data + EDA	약 3000개
Embedding vector	사전학습 모델
학습 모델	LSTM



EDA 기법 적용

num_aug = 3



유의어 교체
alpha_sr = 0

랜덤 삽입
alpha_ri = 0

랜덤 교체
alpha_rs = 0.1

랜덤 삭제
alpha_rd = 0.1

sentence = 새로운 사람을 만나면 항상 두근거린다

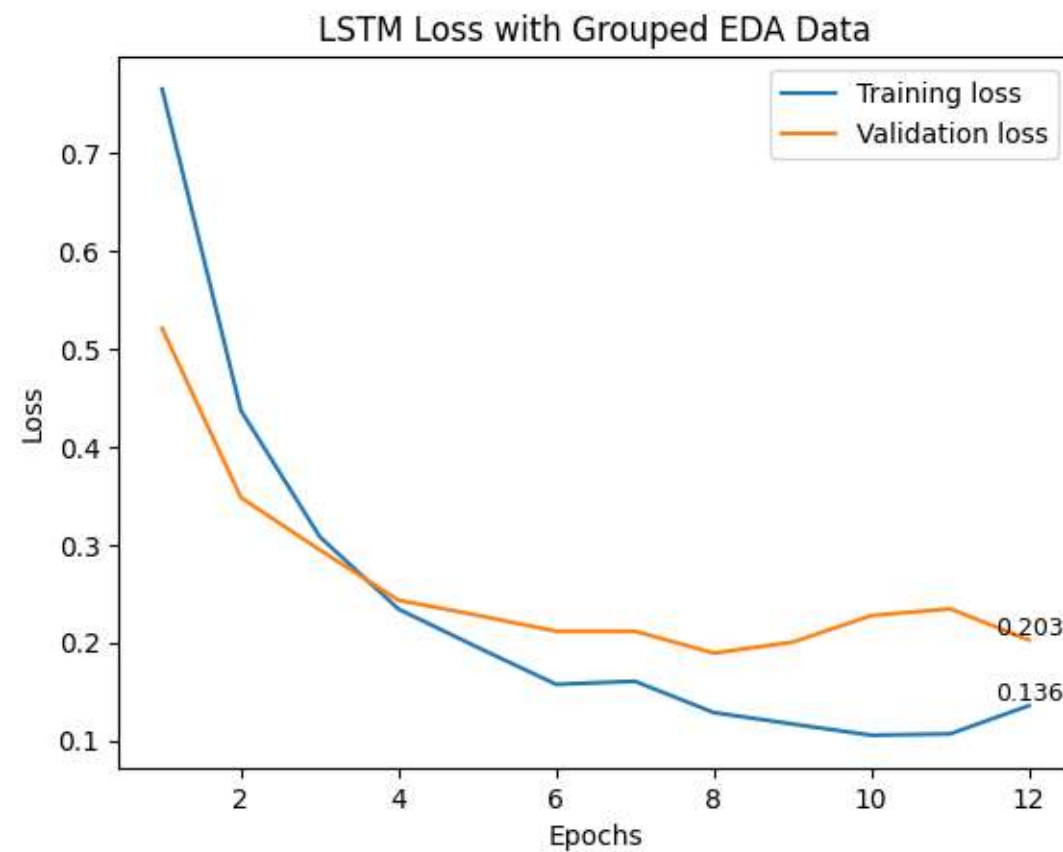
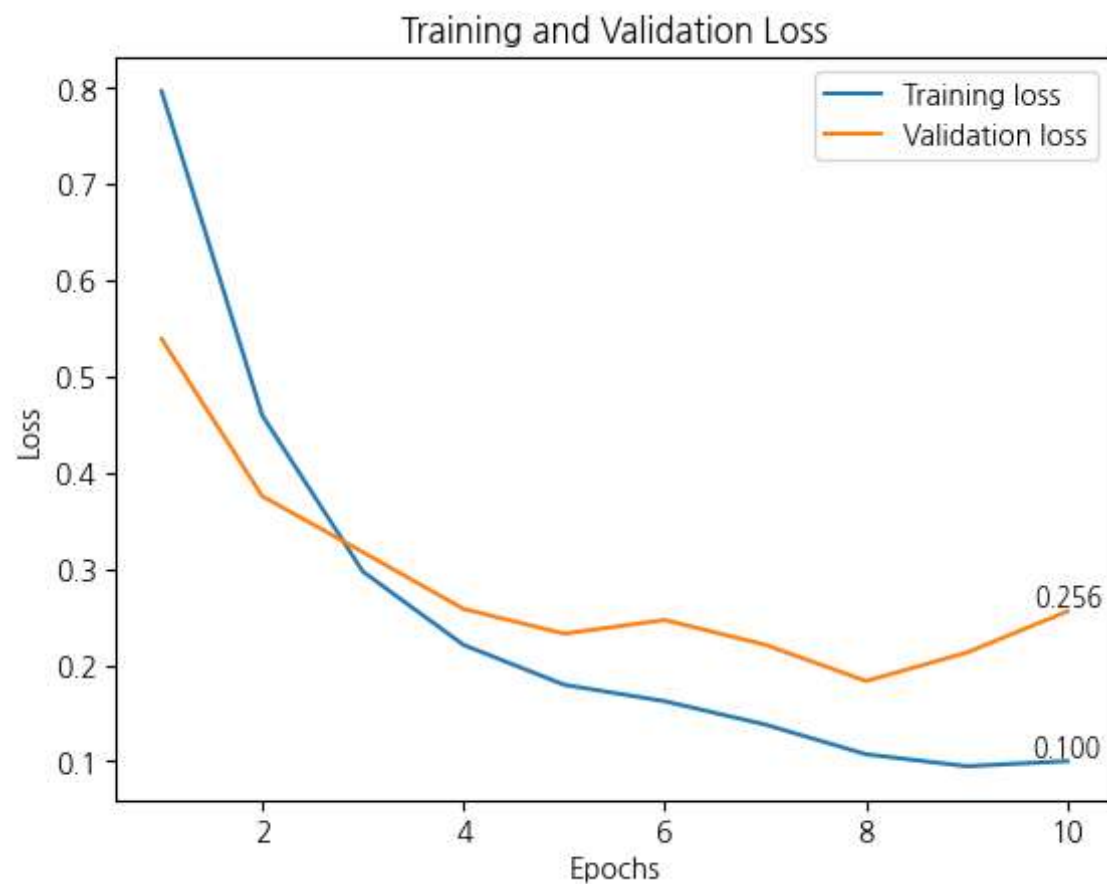
새로운 사람을 만나면 항상 두근거린다
새로운 만나면 항상

sentence = 대부분의 경우 먼저 말을 걸려고 해요

대부분의 경우 먼저 말을 걸려고 해요
대부분의 우 먼저 말을 걸려고 해요
먼저 경우 대부분의 말을 걸려고 해요

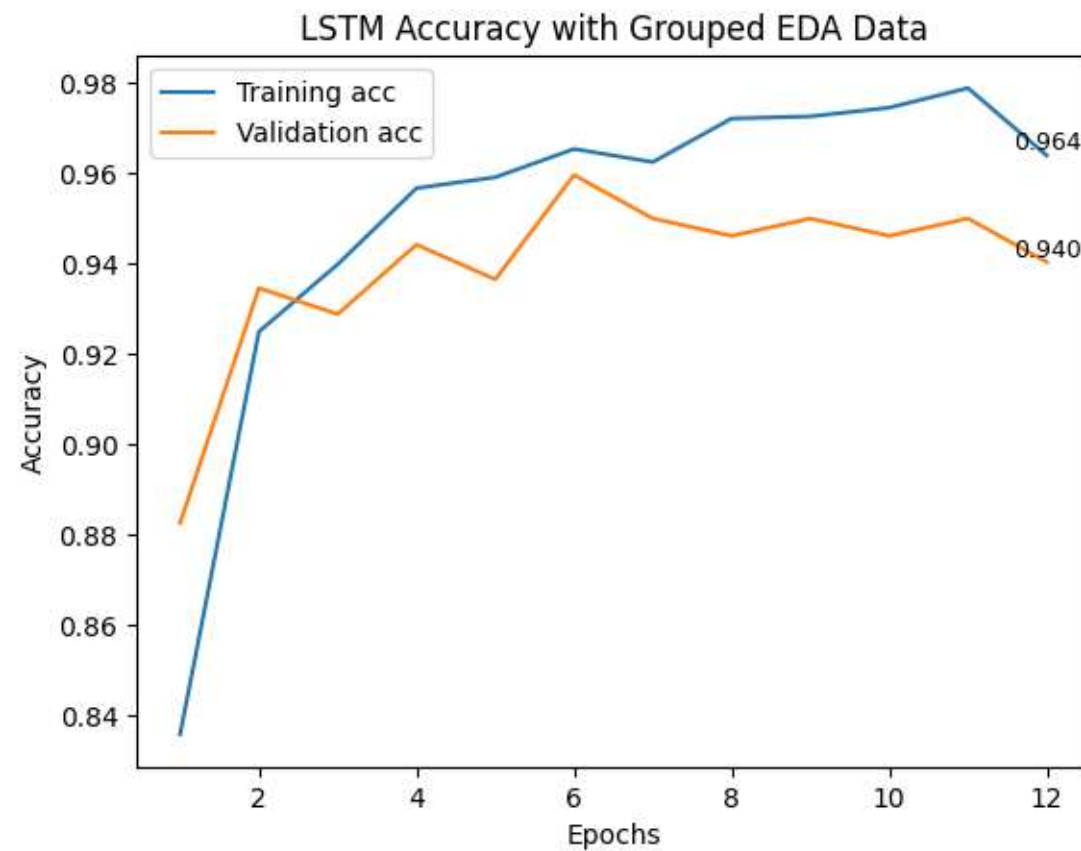
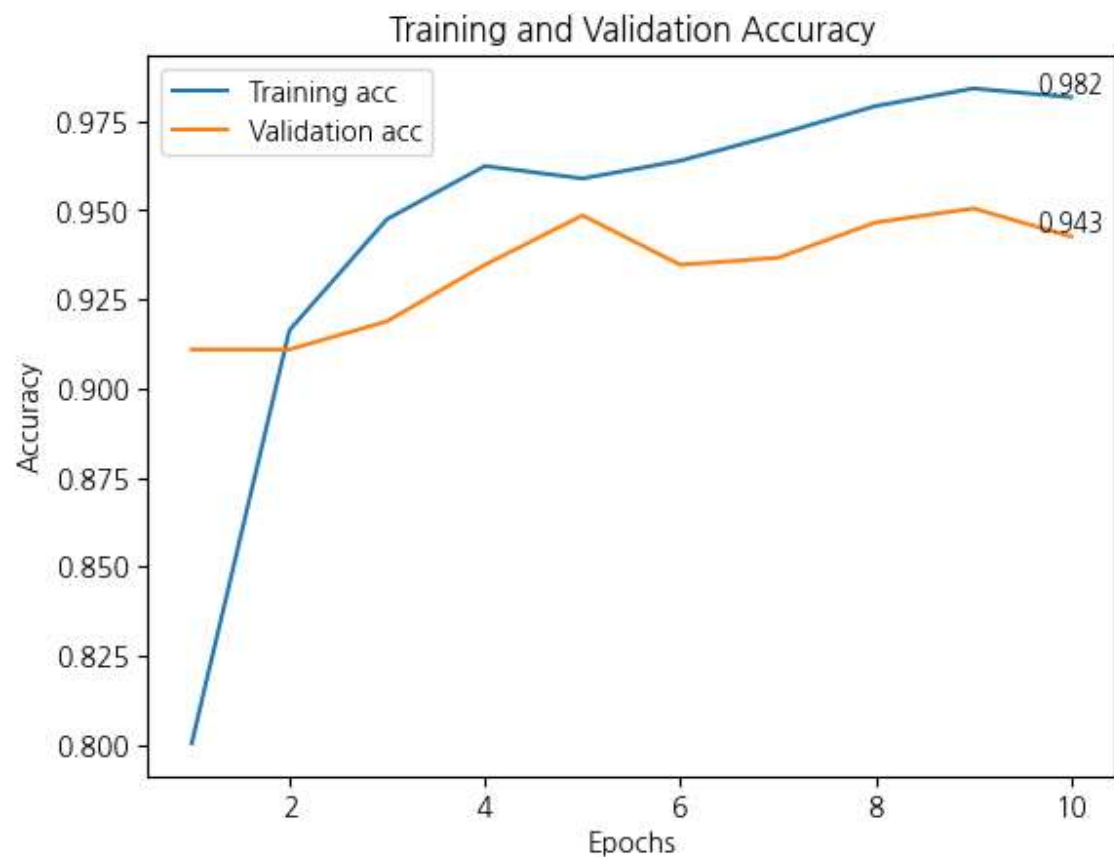
Loss

Base Data + EDA



Acc

Base Data + EDA



높은 성능의 이유?

1. Extract words

(명사, 어간, 동사, 형용사 등)

ex) 좋다, 좋아, 좋아하다, 좋고 -> “좋(다)”

나는 사람 만나는 것을 좋아한다.

로

train

evaluation

~~2. 사전 학습된 모델 사용~~

3. EDA 기법 때문..



문장 A -(EDA기법 적용)-> A1, A2, A3

A1-train에 들어감

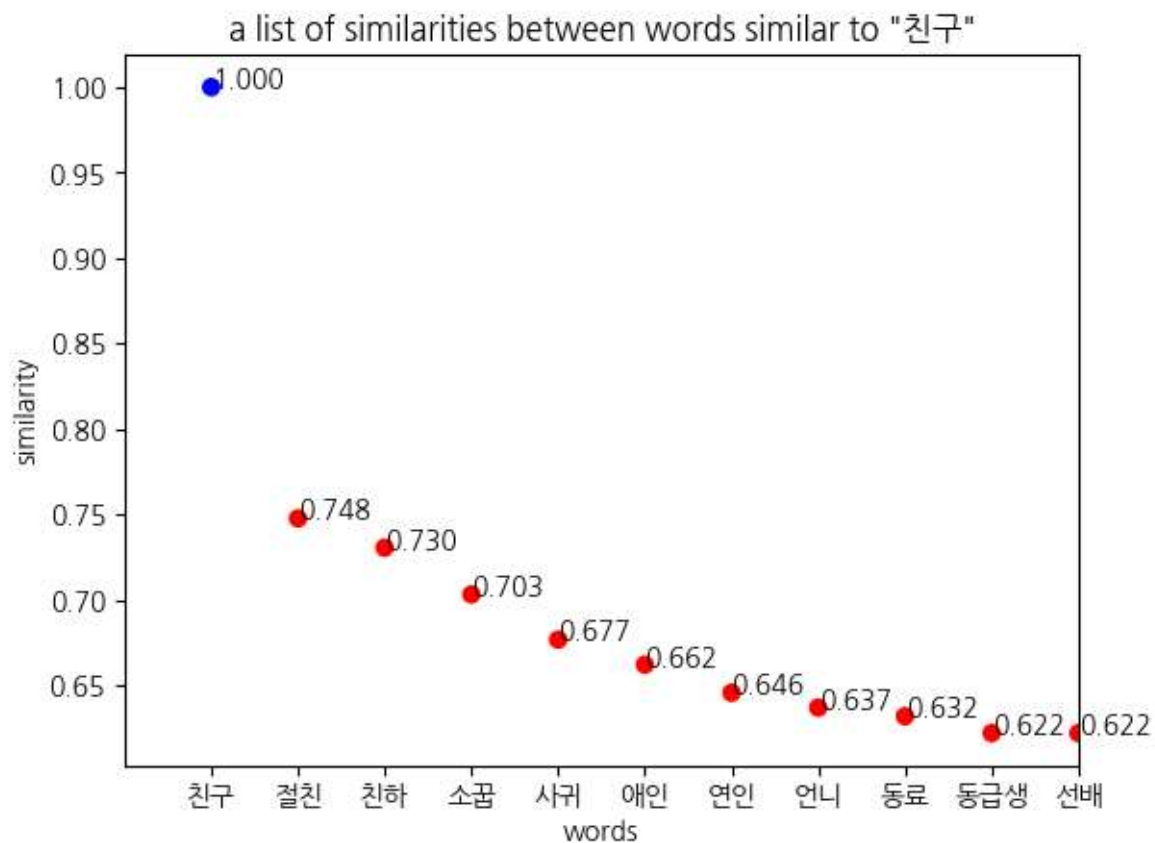
A2-test에 들어감

A3- validation에 들어감..

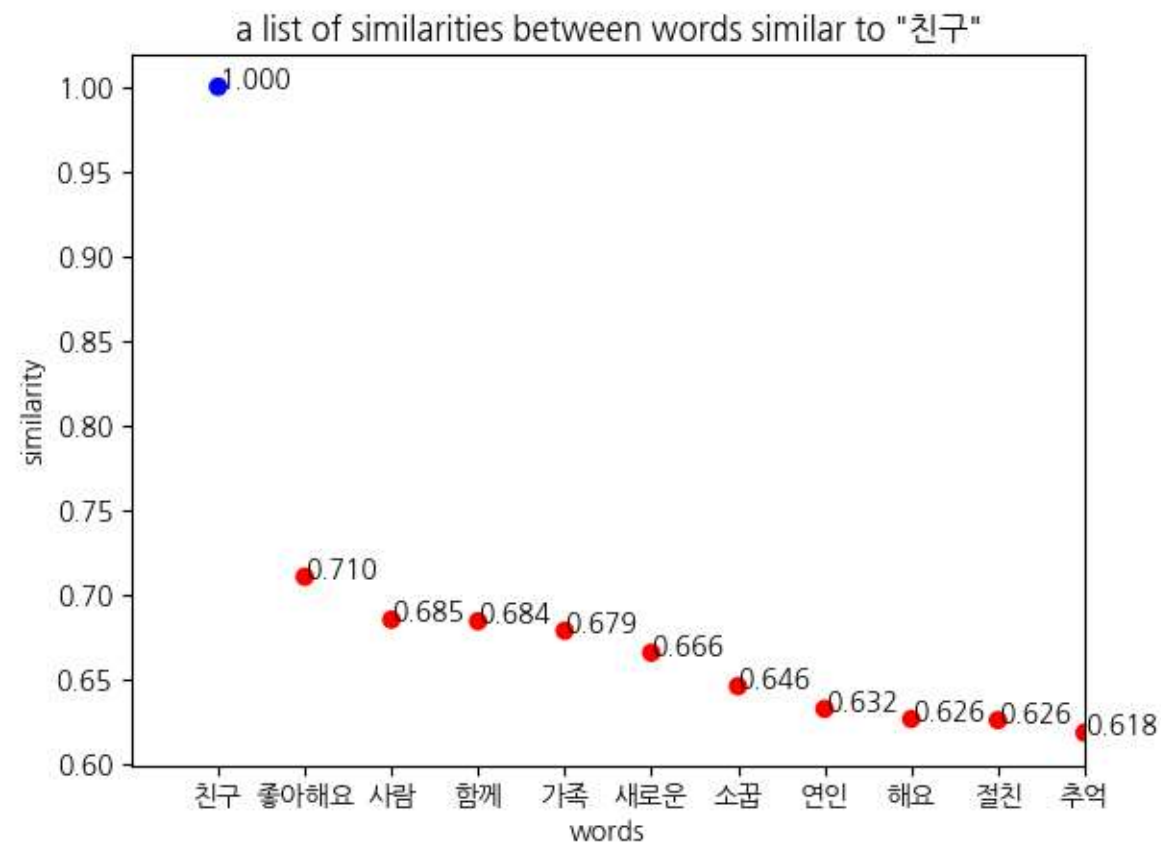
그럼 비슷한 문장이 거의 들어간거니까
아무래도 성능이 좋을 수 밖에..

Similarity

Pre-trained FastText - KR

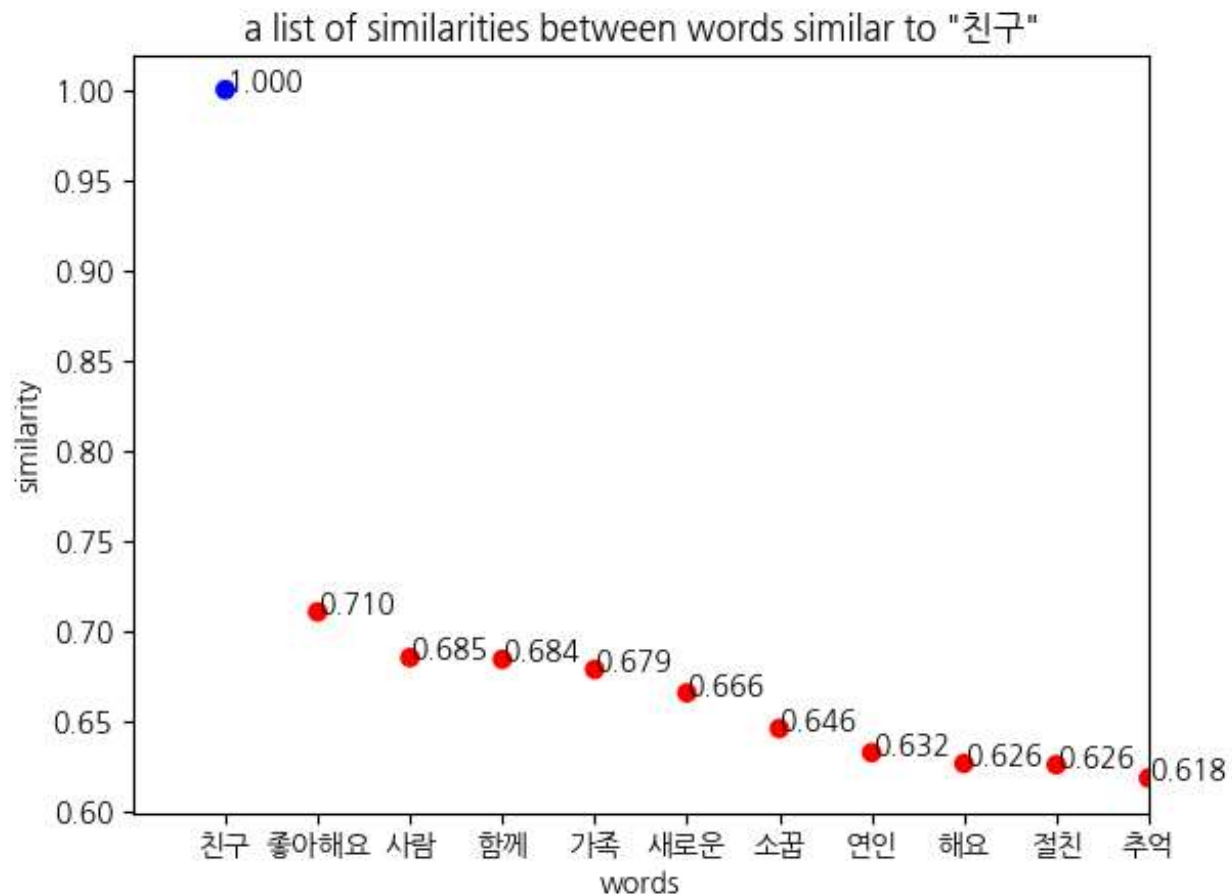


new_FastText model

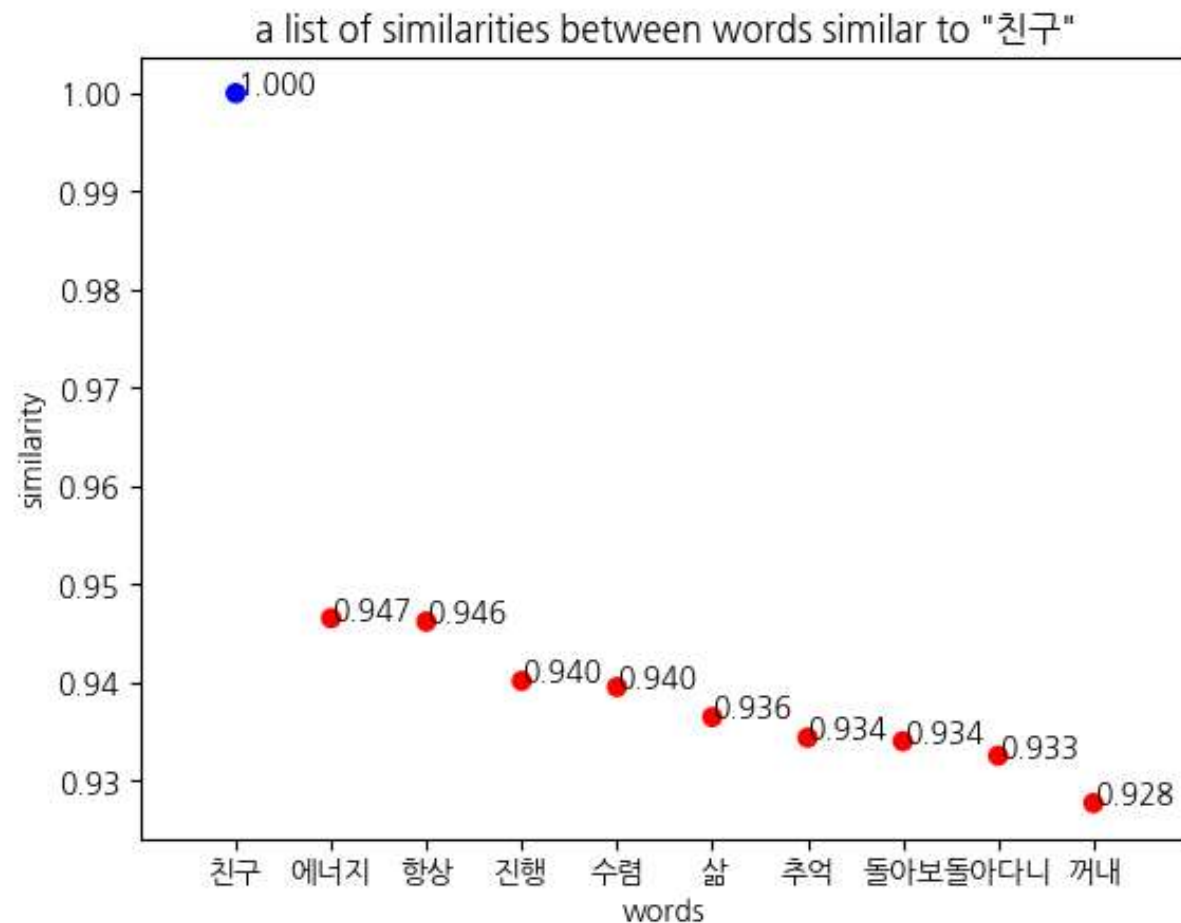


Similarity

09.05



09.19



Prediction for **a single word**

어느정도 예측이 나오기는 하는데
거의 고정된 예측만 나옴!

단일 단어이기 때문에 40~50사이의 어중간한 값
나오는 듯 함!

Input: 친구

Predicted label: 0.5188979506492615

51.89% 정도로 E입니다.

Input: 혼자

Predicted label: 0.4132554233074188

40.33% 정도로 I입니다.

Prediction for a single word

단일 토큰에 대한 예측 어려움...

-> LSTM 모델은 문장 전체의 정보를 사용하여 학습함

따라서 단일 토큰만 제공된다면
해당 토큰의 문맥적, 구조적 정보가 부족하여
모델의 예측이 어려워짐

Prediction for a sentence

문장입력

입력문장: <그래도 주말에는 집에 있는 게 좋아요..>

out of voca

그래도

에

는

에

는

아요

토큰화

00V 토큰 확인: 00V 주말 00V 00V 집 00V 있 00V 게 좋 00V

정수 인코딩

[[1, 32, 1, 1, 24, 1, 4, 1, 641, 9, 1]]

모델 예측값

예측값 : 0.02986094

MBTI 판별

>> 97.01% 정도로 I 유형입니다.

pre \geq 0.5 -> E

pre < 0.5 -> I

Prediction for **a sentence**

예측값 : 0.996568

입력문장: <생일에는 주변 친구들과 신나게 파티를 열어서 놀아요~>
>> 99.66% 정도로 E 유형입니다.

00V 토큰 확인: 생일 00V 00V 주변 친구 들 00V 신나 게 파티 00V 열 00V 놀 00V
[[234, 1, 1, 87, 14, 61, 1, 421, 641, 83, 1, 275, 1, 198, 1]]

Prediction for a sentence – OOV 문제

FastText 성질 없어짐
-> oov처리 x

예측값 : 0.12785016

입력문장: <엠티에서는 그 누구보다도 앞장선다.>
>> 87.21% 정도로 1 유형입니다.

OOV 토큰 확인: OOV OOV OOV OOV OOV OOV OOV OOV
[[1, 1, 1, 1, 1, 1, 1, 1]]

훈련 데이터의 OOV 비율에
따른 예측 값

Prediction for a sentence (using extracted words)

입력문장: <생일에도 집에서 혼자 시간을 보내곤 해요.>

>> 98.84% 정도로 I 유형입니다.

00V 토큰 확인: 생일 집 혼자 시간 보내

[[222, 24, 8, 6, 42]]

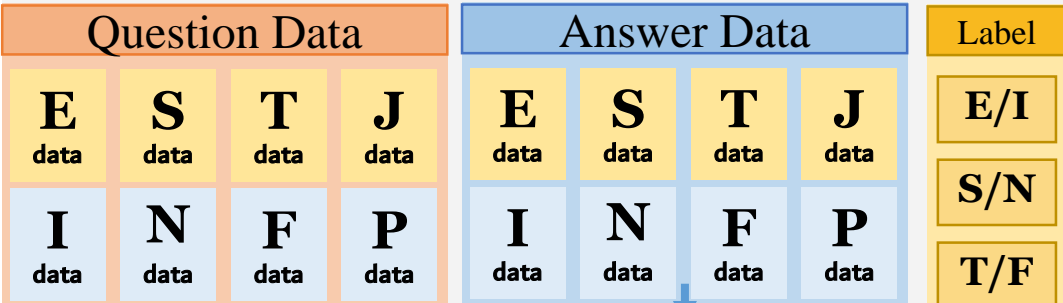
입력문장: <먼저 말을 거는 편이다.>

>> 98.64% 정도로 E 유형입니다.

00V 토큰 확인: 먼저 말 00V

[[20, 49, 1]]

Dataset



KoGPT-2

EDA

Data Augmentation



Text preprocessing

data = [Question data_Answer data_Label]

LSTM

Bi-LSTM
+ Attention

Transformer
- Encoder

4-BINARY CLASSIFIER

Model

E/I
model

S/N
model

T/F
model

J/P
model

Answer Analysis Model



KoNLPy
(Morpheme Analysis)

Word Embedding Vector
values for each Key-word

Sentiment
Analysis

LSTM

Frequency

Modifier
(adverbs)

Weighted Sum
: 가중치의 곱을 더

softmax

UI

Bot

User

E/I Question

Answer Analysis
Model

Identifying
a single MBTI type

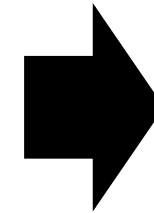
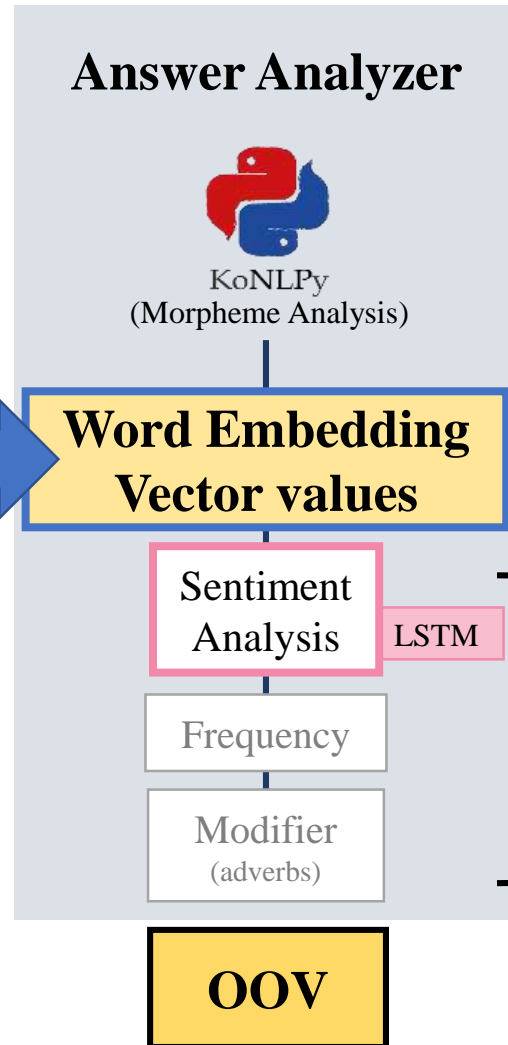
User's Answer

Output

: User's MBTI

Answer Analyzer

MBTI 예측
LSTM 모델



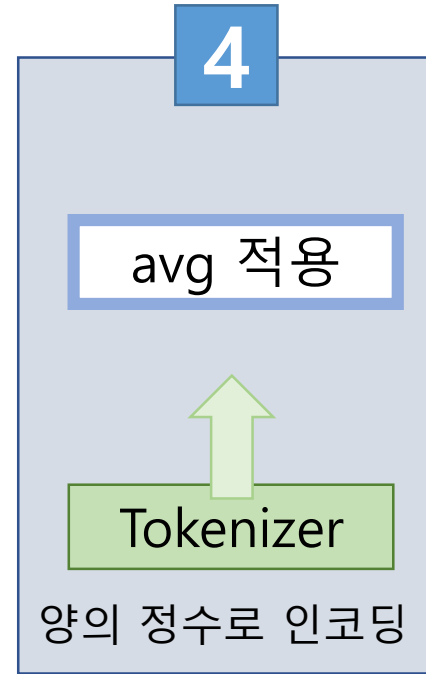
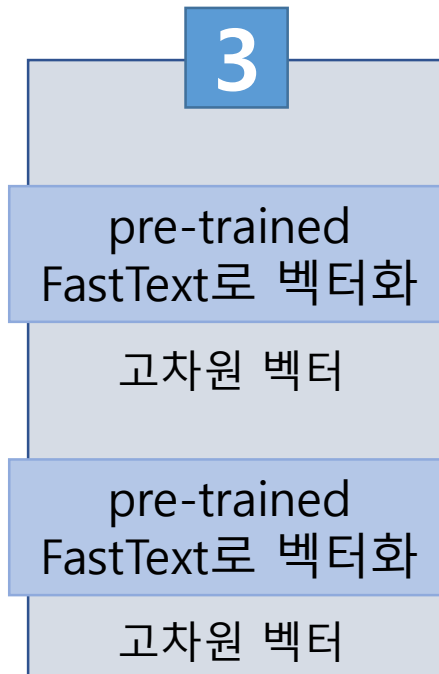
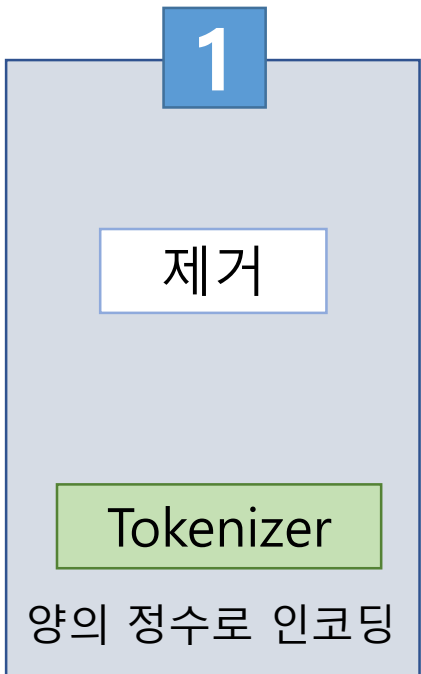
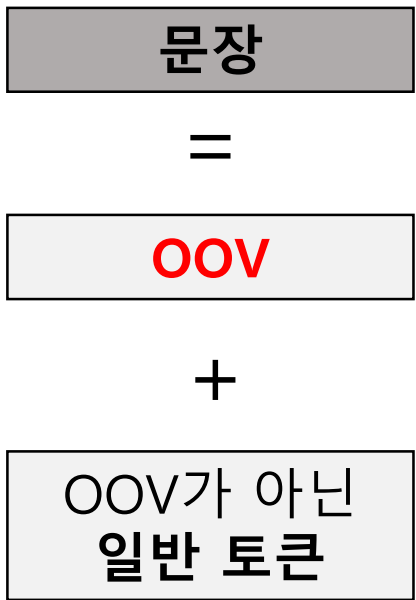
<최종>
MBTI 분석

정확도를
높이는 요소

Answer Analyzer – OOV

입력 문장 = "나는 친구를 만날 때 가장 행복하다."

형태소 단위로 토큰분리



LSTM 모델 학습의 한계

1. 문맥 학습의 한계 -> 예측이 어느정도만 되고 그렇게 잘 되지는 않는다!
2. EDA 기법의 분산 문제 미해결
(문장과 문장 aug는 (train,val,test)중 같은 곳에 위치해야한다!
- 3.

Bi-LSTM

