2021 산업혁신 빅데이터 플랫폼 공모전

딥러닝을 활용한 표준산업코드 분류

애쓰는 감자팀

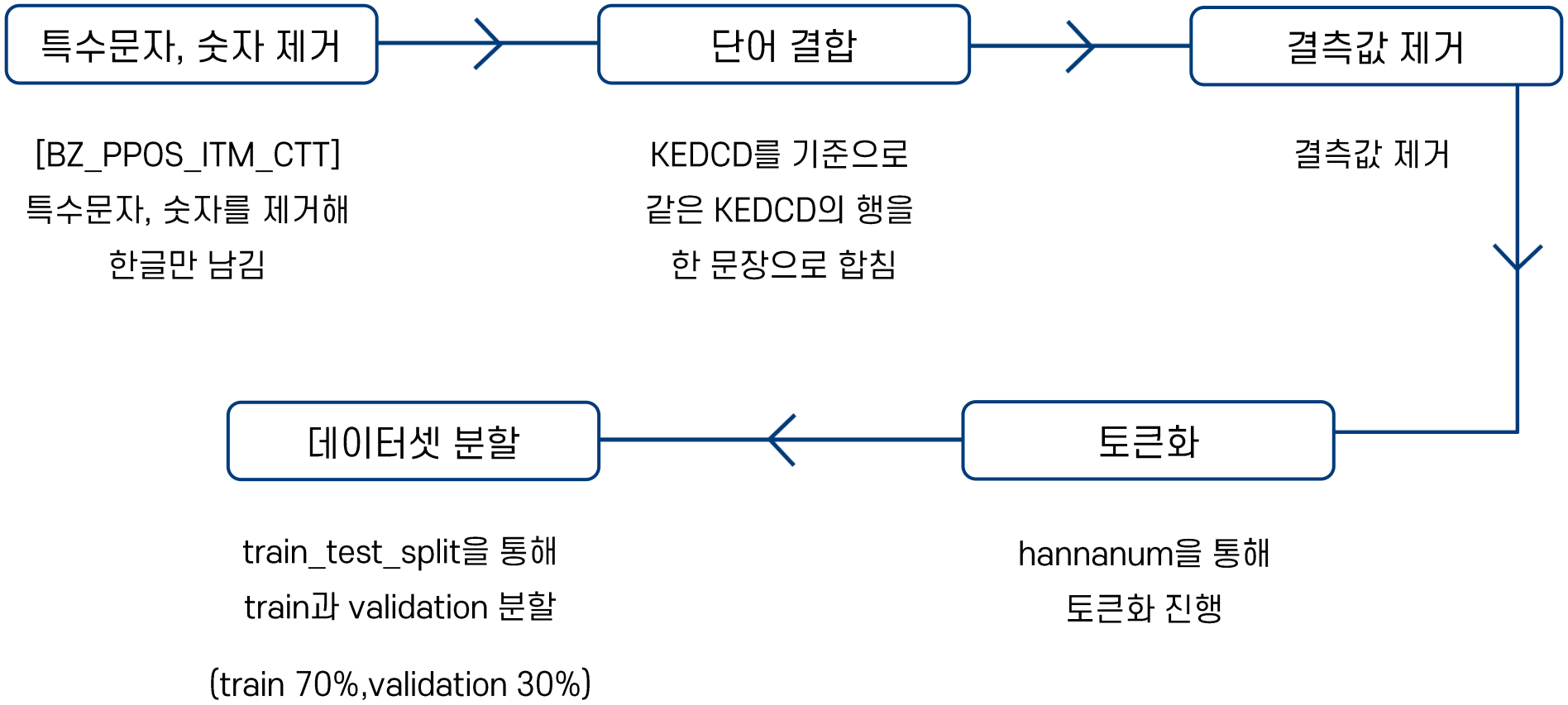
곽민지, 김영민, 유수빈

목차


Ⅰ. 공통 전처리
01. text preprocessing
03. 한국표준 산업 분류 데이터
04. train 파일 생성
Ⅱ.알고리즘
01. KoBERT
02. LSTM
03. 선택과정
Ⅲ.예측 결과
01. 최종 예측 결과 산출방식
02. 최종 예측 결과

전처리

text preprocessing



전처리

한국 표준 산업 분류 데이터

	A	B	C	D	E	F	G	H	I	J
1	개정 분류체계(제10차 기준)									
2	대분류(21)		중분류(77)		소분류(232)		세분류(495)		세세분류(1,196)	
3	코드	항목명	코드	항목명	코드	항목명	코드	항목명	코드	항목명
4	A	농업, 임업 및 어업(01~03)	01	농업	011	작물 재배업	0111	곡물 및 기타 식량작물 재배업	01110	곡물 및 기타 식량작물 재배업
5							0112	채소, 화훼작물 및 종묘 재배업	01121	채소작물 재배업
6									01122	화훼작물 재배업
7									01123	종자 및 묘목 생산업
8							0113	과실, 음료용 및 향신용 작물 재배업	01131	과실작물 재배업
9									01132	음료용 및 향신용 작물 재배업
10							0114	기타 작물 재배업	01140	기타 작물 재배업
11							0115	시설작물 재배업	01151	콩나물 재배업
12									01152	채소, 화훼 및 과실작물 시설 재배업
13									01159	기타 시설작물 재배업
14					012	축산업	0121	소 사육업	01211	젖소 사육업
15									01212	육우 사육업
16							0122	양돈업	01220	양돈업
17							0123	가금류 및 조류 사육업	01231	양계업
18									01239	기타 가금류 및 조류 사육업
19							0129	기타 축산업	01291	말 및 양 사육업
20									01299	그 외 기타 축산업
21					013	작물재배 및 축산 복합농업	0130	작물재배 및 축산 복합농업	01300	작물재배 및 축산 복합농업

“통계청 – 한국 표준 산업 분류표” 데이터 사용

전처리

한국 표준 산업 분류 데이터

대분류 : 대 / 중 / 소 / 세 / 세세분류의 [BZ_PPOS_ITM_CTT]와 [KSIC10_BZC_CD] 추가

중분류 : 중 / 소 / 세 / 세세분류의 [BZ_PPOS_ITM_CTT]와 [KSIC10_BZC_CD] 추가

소분류 : 소 / 세 / 세세분류의 [BZ_PPOS_ITM_CTT]와 [KSIC10_BZC_CD] 추가

세분류 : 세분 / 세세분류의 [BZ_PPOS_ITM_CTT]와 [KSIC10_BZC_CD] 추가

세세분류 : 세세분류의 [BZ_PPOS_ITM_CTT]와 [KSIC10_BZC_CD] 추가

→ **데이터를 만든 후 분류에 맞게 [KSIC10_BZC_CD] 가공**

전처리

한국 표준 산업 분류 데이터

EX) 중분류 데이터 생성

	A	B	C	D	E	F	G	H	I	J
1	개정 분류체계(제10차 기준)									
2	대분류(21)		중분류(77)		소분류(232)		세분류(495)		세세분류(1,196)	
3	코드	항목명	코드	항목명	코드	항목명	코드	항목명	코드	항목명
4	A	농업, 임업 및 어업(01~03)	01	농업	011	작물 재배업	0111	곡물 및 기타 식량작물 재배업	01110	곡물 및 기타 식량작물 재배업
5							0112	채소, 화훼작물 및 종묘 재배업	01121	채소작물 재배업
6									01122	화훼작물 재배업
7									01123	종자 및 묘목 생산업
8							0113	과실, 음료용 및 향신용 작물 재배업	01131	과실작물 재배업
9									01132	음료용 및 향신용 작물 재배업
10							0114	기타 작물 재배업	01140	기타 작물 재배업
11							0115	시설작물 재배업	01151	콩나물 재배업
12									01152	채소, 화훼 및 과실작물 시설 재배업
13									01159	기타 시설작물 재배업
14					012	축산업	0121	소 사육업	01211	젖소 사육업
15									01212	육우 사육업
16							0122	양돈업	01220	양돈업
17							0123	가금류 및 조류 사육업	01231	양계업
18									01239	기타 가금류 및 조류 사육업
19							0129	기타 축산업	01291	말 및 양 사육업
20									01299	그 외 기타 축산업
21					013	작물재배 및 축산 복합농업	0130	작물재배 및 축산 복합농업	01300	작물재배 및 축산 복합농업

	BZ_PPOS_ITM_CTT	mid
21	농업	01
22	임업	02
23	어업	03
24	석탄 원유 및 천연가스 광업	05
25	금속 광업	06
26	비금속광물 광업 연료용 제외	07
27	광업 지원 서비스업	08
28	식료품 제조업	10
29	음료 제조업	11
30	담배 제조업	12
31	섬유제품 제조업 의복 제외	13
32	의복 의복 액세서리 및 모피제품 제조업	14
33	가죽 가방 및 신발 제조업	15
34	목재 및 나무제품 제조업 가구 제외	16
35	펄프 종이 및 종이제품 제조업	17
36	인쇄 및 기록매체 복제업	18
37	코크스 연탄 및 석유정제품 제조업	19
38	화학 물질 및 화학제품 제조업 의약품 제외	20
39	의료용 물질 및 의약품 제조업	21
40	고무 및 플라스틱제품 제조업	22

중분류를 예측한다면, 분류표의 중, 소, 세, 세세분류의 데이터를 가져와서 train 데이터에 추가

전처리

분류별 클래스 개수

A01234

A : 대분류 21개

01 : 중분류 77개

2 : 소분류 232개

3 : 세분류 495개

4 : 세세분류 1,196개

전처리

train 데이터 생성

- LSTM & CNN & Transformer: 분류표를 추가한 train 데이터 사용

	KEDCD	BZ_PPOS_ITM_CTT	KSIC10_BZC_CD	mid	Big	se	sese
0	8078635.0	는 자산유동화에관한법률이하 자산유동화법이라 한다의 규정에 따라 다음 각호의 사업을 ...	K66199	K66	K	6619	66199
1	8020873.0	기계설비제작업 방사선시설 설계 및 공사업 방사선안전관리대행 및 전문기술업 엘...	C25929	C25	C	2592	25929
2	7375653.0	민국 교육의 근본이념에 의거하여 국민생활에 직접 필요한 직업의 지식과 기술을 연마함...	P85212	P85	P	8521	85212
3	5099079.0	토목 건축 공사업 시설물 유지관리업 시설물 안전점검 및 정밀안전진단 용역업 ...	N74100	N74	N	7410	74100
4	8086464.0	건축자재품 제조및 제작업 이형압출성형품 제조가공및 도소매업 사출성형품 제조 가공및 ...	G47912	G47	G	4791	47912
...	...	...	...	...	...	...	...
1081124	70920.0	산업디자인업 컴퓨터그래픽디자인업 웹디자인개발 홍보 영상물 제작 전시 영상물...	J62010	J62	J	6201	62010
1081125	7911282.0	복지기본법 및 동법시행령의 규정에 따라 사내근로복지기금의 관리 운영을 효율적으로 함...	K65139	K65	K	6513	65139
1081126	8377811.0	레스토랑 생기를담아 협동조합은 지역순환과 협동의 원리에 기초한 지역먹거리체계 형성을...	S94990	S94	S	9499	94990
1081127	7904943.0	부동산 컨설팅업 부동산 개발및 사업시행업 주택 및 상가 분양대행업 부동산 임...	L68222	L68	L	6822	68222
1081128	170992.0	실내외 인테리어 건축업 가구제작 및 도소매업 철근콘크리트 공사 및 상하수도 설...	F42121	F42	F	4212	42121

1081129 rows × 7 columns

전처리

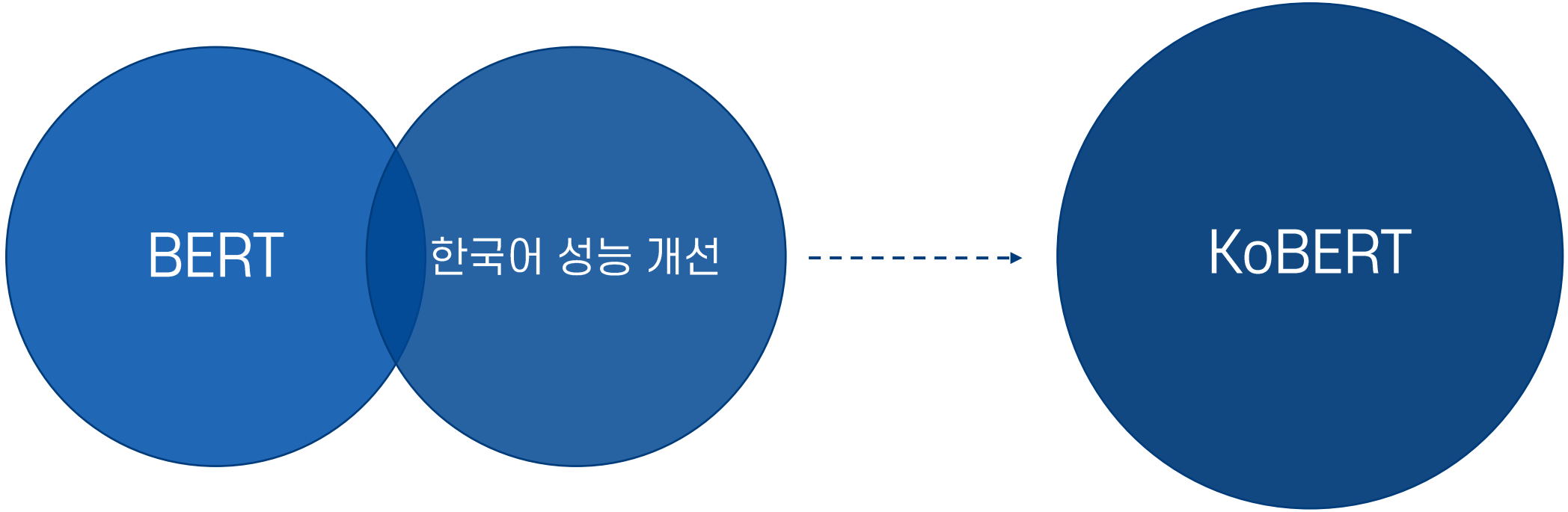
train 데이터 생성

- BERT: 분류표를 추가한 train 데이터의 라벨 인코딩 → 필요한 열 select → text 파일 생성

	BZ_PPOS_ITM_CTT	mid_idx
0	용역 경비업 ...	63
1	사회적협동조합은 자주적 자립적 자치적인 조합활동을 통하여 학부모교원지역주민이 께 만...	71
2	생활용품 도소매업 생활용품 중개업 창고업 창고 임대업 부동산 임대업 위 ...	40
3	기계제작 및 제조 기계류 및 건축자재 도매업 기계류 및 건축자재 소매업 기계...	26
4	부동산 개발 부동산 투자 및 투자컨설팅 부동산 매매 임대업 부동산 분양 및 ...	57

알고리즘

KoBERT



알고리즘

KoBERT

설명

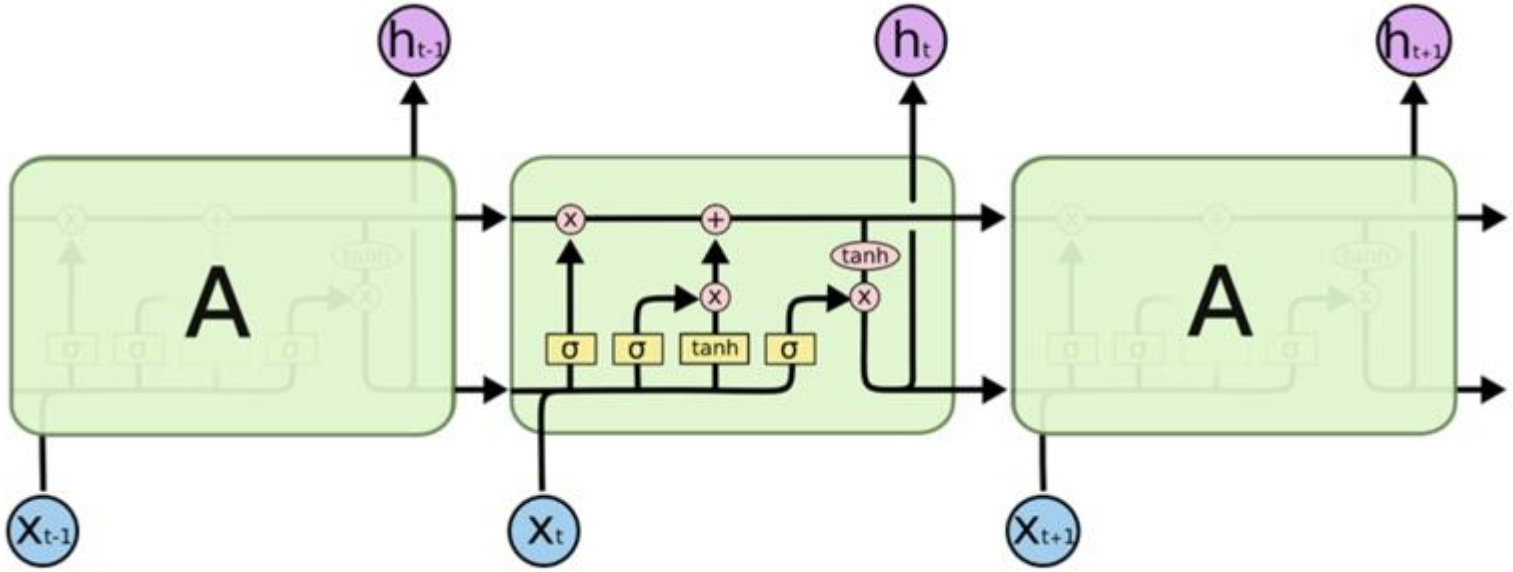
- 수백만 개의 한국어 문장으로 이루어진 대규모 말뭉치(corpus)를 학습
- 한국어의 불규칙한 언어 변화의 특성을 반영하기 위해 데이터 기반 토큰화(Tokenization) 기법 사용
- 대량의 데이터를 빠른 시간에 학습하기 위해 링 리듀스 기반 분산 학습 기술을 사용

전처리

- Tokenizer : Sentencepiece tokenizer
- vocab : KoBERT의 vocab
- KoBERT 모델에 맞게 tranform

알고리즘

LSTM



- 기존 RNN의 위치 차이가 커질 시 문맥 연결이 어렵다는 문제를 개선한 알고리즘
- 싱글 레이어 였던 RNN과 달리 상호작용하는 4개의 레이어 有
- 여러 종류의 게이트가 있어 입력을 선별적으로 허용, 계산결과 선별적 출력

알고리즘

CNN & LSTM

01

Konlpy 모듈의 mecab을 사용하여 토큰화, 불용어 제거

MECAB

02

Tensorflow의 texts_to_sequences() 메서드를 이용해서 단어를 시퀀스의 형태로 변환

Texts_to_sequences

03

Tensorflow의 pad_sequences함수를 사용하여 길이를 동일하게

Pad_sequences

04

sklearn의 OneHotEncoder를 사용하여 y값을 인코딩 및 어레이화

OneHotEncoder

알고리즘

CNN_1 모델 구성

```
model1 = Sequential()
model1.add(Embedding(vocab_size,64, input_length=max_len))
model1.add(Conv1D(64, 5, activation = 'relu', padding = 'same', kernel_regularizer=l2(0.01),
kernel_initializer=tf.keras.initializers.TruncatedNormal(0.02)))
model1.add(Conv1D(128, 5, activation = 'relu', padding = 'same', kernel_regularizer=l2(0.01),
kernel_initializer=tf.keras.initializers.TruncatedNormal(0.02)))
model1.add(GlobalMaxPooling1D())
model1.add(Dense(231, activation = 'softmax', kernel_initializer=tf.keras.initializers.TruncatedNormal(0.02)))
model1.summary()
```

→ Embedding Layer를 활용하여 정수 인코딩이 된 단어들을 입력받아 임베딩 벡터화

→ CNN Layer를 추가하여 특징을 추출

→ GlobalMaxPooling을 적용하여 CNN Layer에서 나온 특징벡터들 중 가장 큰 벡터 택함

알고리즘

CNN_ 2 모델 구성

```
model2 = Sequential()
model2.add(Embedding(vocab_size, 64, input_length = max_len))
model2.add(Conv1D(64, 5, activation = 'relu',kernel_regularizer= l2(0.001)))
model2.add(MaxPooling1D(5))
model2.add(Dropout(.5))
model2.add(Conv1D(64, 5 ,activation = 'relu', kernel_regularizer=  l2(.001)))
model2.add(GlobalMaxPooling1D())
model2.add(BatchNormalization())
model2.add(Dense(231, activation = 'softmax', kernel_regularizer = l2(0.001)))
model2.summary()
```

→ Embedding Layer를 활용하여 정수 인코딩이 된 단어들을 입력받아 임베딩 벡터화

→ CNN Layer를 추가하여 특징을 추출

→ GlobalMaxPooling을 적용하여 CNN Layer에서 나온 특징벡터들 중 가장 큰 벡터 택함

→ Drop out, BatchNormalization을 적용하여 과적합 방지

알고리즘

LSTM _ 1모델 구성

```
model3 = Sequential()
model3.add(Embedding(vocab_size, 64, input_length = max_len))
model3.add(Conv1D(32, 5, activation='relu'))
model3.add(Conv1D(32, 5, activation='relu'))
model3.add(Conv1D(32, 5, activation='relu'))
model3.add(MaxPooling1D(pool_size=4))
model3.add(LSTM(16))
model3.add(Dropout(0.4))
model3.add(BatchNormalization())
model3.add(Dense(231, activation='softmax'))
model3.summary()
```

→ Embedding Layer를 활용하여 정수 인코딩이 된 단어들을 입력받아 임베딩 벡터화

→ CNN Layer를 추가하여 특징을 추출

→ MaxPooling을 적용하여 CNN Layer에서 나온 특징벡터들 중 가장 큰 벡터 택함

→ LSTM Layer를 추가

→ Drop out, BatchNormalization을 적용하여 과적합 방지

알고리즘

LSTM _ 2 모델 구성

```
model4 = Sequential()
model4.add(Embedding(vocab_size, 32, input_length=max_len))
model4.add(Dropout(0.3))
model4.add(Conv1D(32, 5, activation='relu'))
model4.add(Conv1D(32, 5, activation='relu'))
model4.add(MaxPooling1D(pool_size=4))
model4.add(LSTM(32))
model4.add(BatchNormalization())
model4.add(Dense(231, activation='softmax'))
model4.summary()
```

→ Embedding Layer를 활용하여 정수 인코딩이 된 단어들을 입력받아 임베딩 벡터로 만든다.

→ CNN Layer를 추가

→ MaxPooling을 적용하여 CNN Layer에서 나온 특징벡터들 중 가장 큰 벡터 택함

→ LSTM Layer를 추가

→ Drop out, BatchNormalization을 적용하여 과적합 방지

알고리즘

LSTM 선택 과정

	대분류	중분류	소분류	세분류	세세분류
CNN_1	0.806	0.713	0.618	-	-
CNN_2	0.817	0.716	0.609	-	-
LSTM_1	0.818	0.719	0.610	-	-
LSTM_2	0.817	0.729	0.633	0.511	0.288
Transformer	0.716	0.706	0.601	-	-

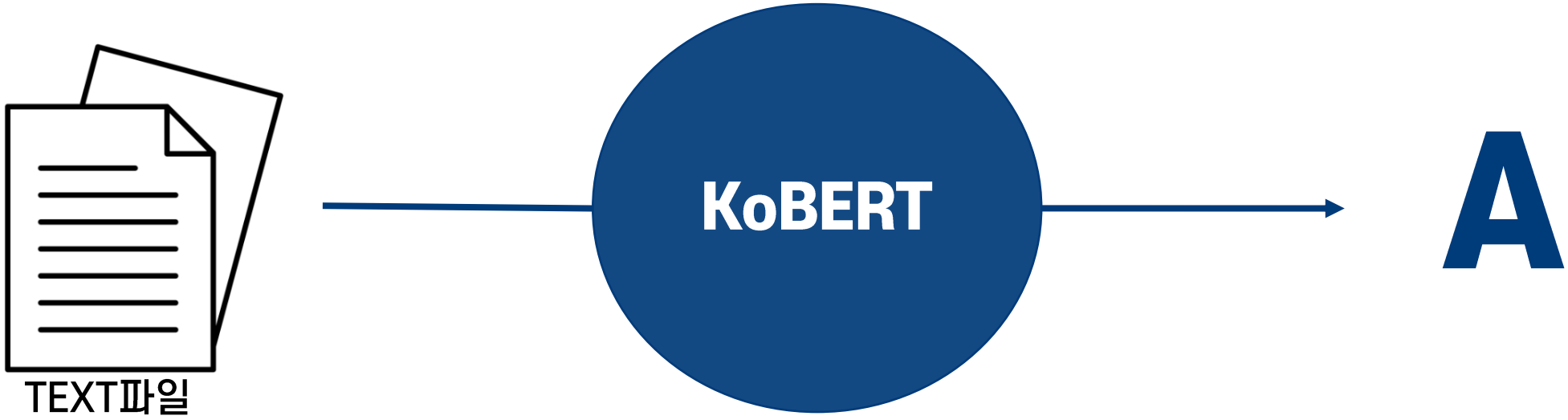
알고리즘

선택 과정

	대분류	중분류	소분류	세분류	세세분류
LSTM_	0.818	0.729	0.633	0.511	0.288
BERT	0.843	0.752	0.661	0.436	0.47

예측 결과

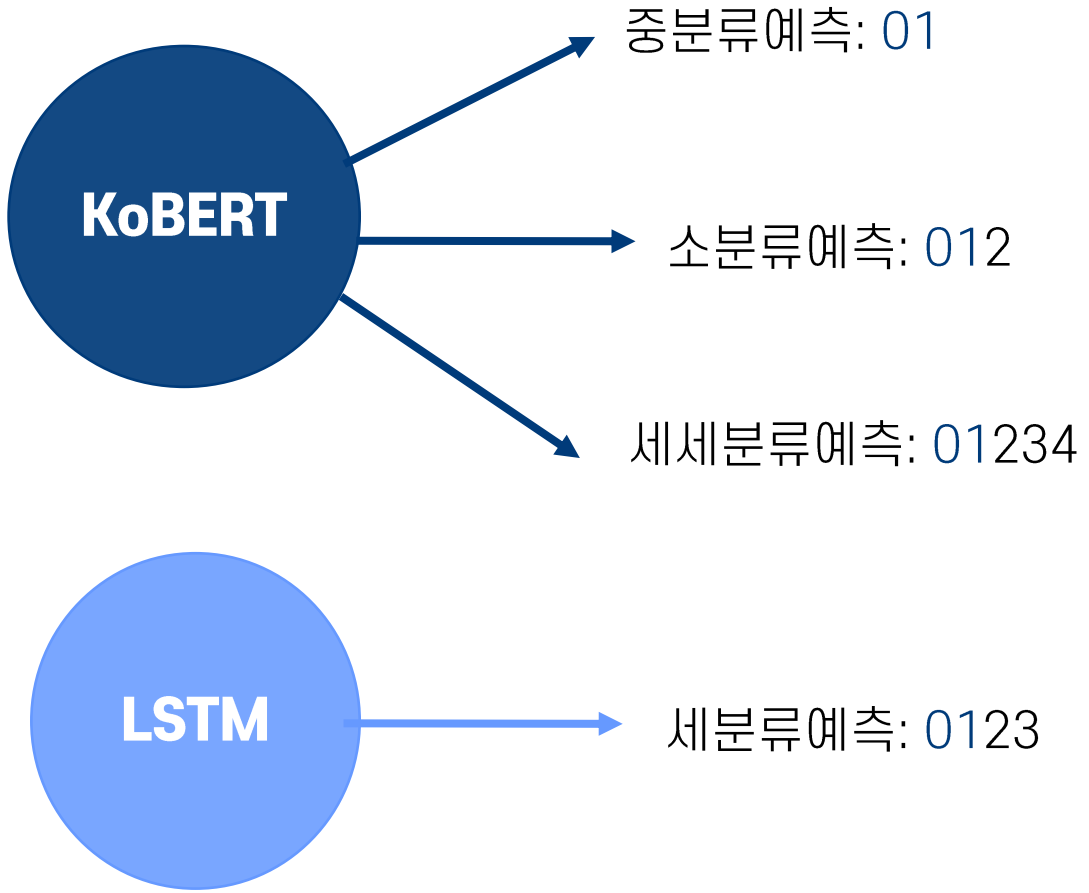
최종 예측 결과 산출 방식 (대분류)



→ 대분류의 경우 BERT 모델이 예측한 예측결과를 선택

예측 결과

최종 예측 결과 산출 방식 (중분류)



01

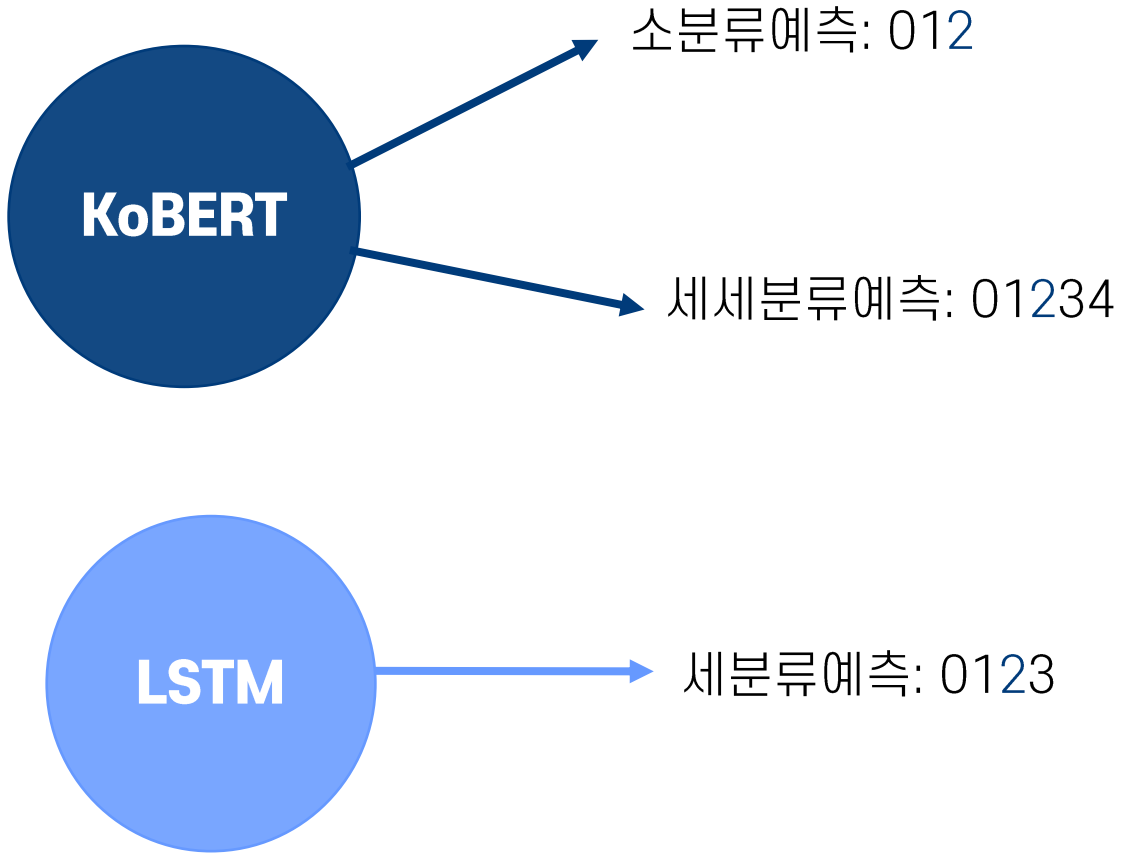
→ 중분류, 소분류, 세세분류용 BERT 모델 / 세분류용 LSTM 모델을 통해 중분류 예측값을 구한다.

예측값이 모두 다를 경우, 2:2의 상황일 경우 중분류용 BERT 모델이 예측한 값을 택하고,

그 이외에는 최빈값을 택한다.

예측 결과

최종 예측 결과 산출 방식 (소분류)



2

→ 소분류, 세세분류용 BERT 모델 / 세분류용 LSTM 모델을 통해 소분류 예측값을 구한다.

예측값이 모두 다를 경우 소분류용 버트 모델이 예측한 값을 택하고, 그 이외에는 최빈값을 택한다.

예측 결과

최종 예측 결과 산출 방식 (세분류, 세세분류)

KoBERT → 세세분류예측: 01234 → 4

LSTM → 세분류예측: 0123 → 3

→ 세분류의 경우 세분류용 LSTM 모델이 예측한 세분류를 사용

→ 세세분류는 세세분류용 KoBert 모델이 예측한 세세분류를 사용

예측 결과

최종 예측 결과 산출 방식

A + 01 + 2 + 3 + 4

= A01234

→ 각 분류별 예측한 값을 합하여 최종결과 산출

예측 결과

최종 예측 결과

	A	B	C
1	KEDCD	예측한 업종코드	
2	7410201	A01152	
3	7864893	A01129	
4	7865236	A01139	
5	7869309	A01132	
6	7869239	A01139	
7	7598892	A01102	
8	7786718	A01102	
9	7874883	A01132	
10	7764674	A01190	
11	7088772	A01192	
12	7526425	A01212	
13	7533023	A01199	
14	7528480	A01190	
15	7470998	A01192	
16	927913	A01121	
17	7448048	A01172	
18	7670759	A01192	
19	7672040	L68132	
20	7111949	A01109	
21	7113727	A01112	
22	7115094	A01399	
23	7097394	A01112	
24	7099178	A01199	
25	7155795	A01112	

Sheet1

.....

	A	B	C
19137	7089203	A01101	
19138	7515988	S94192	
19139	7495912	M70120	
19140	7754832	N71539	
19141	6004077	M72911	
19142	8036242	M70139	
19143	7098613	S94990	
19144	6565745	M71599	
19145	7656706	S96901	
19146	7788999	S94990	
19147	7788889	S94192	
19148	7650436	Q87222	
19149	6667333	S96911	
19150	7592435	C94392	
19151	7735078	S94110	
19152	7745025	S94202	
19153	2141550	S94990	
19154	6644087	N75202	
19155	7839359	S94122	
19156	7548852	S94110	
19157	6746270	K64291	
19158	2056422	S94990	
19159	7666646	O94904	
19160			
19161			

Sheet1