응시번호	AI 0138

「통계데이터 인공지능 활용대회」 2차 심사 코드설명자료

1. 구동환경 및 정보

1) 구현에 사용한 프로그래밍 언어 : Python

2) GPU, CPU 등 하드웨어 정보

- Google Colab Pro : Google 리서치팀에서 제공하는 Colaboratory. 호스팅된 Jupyter 노트북 서비스로 GPU를 포함한 컴퓨팅 리소스 사용하는 가상머신

- Colab Pro 버전 사용 시 할당된 GPU 및 메모리 확인 **Colab에서 제공하는 GPU 사양은 매번 달라질 수 있음*

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 460.32.03    Driver Version: 460.32.03    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla P100-PCIE...  Off  | 00000000:00:04.0 Off |                    0 |
| N/A   34C    P0    27W / 250W |      0MiB / 16280MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

[그림1 : 할당된 GPU 정보 예시]

```
1 !head -n 3 /proc/meminfo

MemTotal:       13302920 kB
MemFree:        10241248 kB
MemAvailable:   12108808 kB
```

[그림2 : 할당된 메모리 정보]

3) 학습 및 테스트에 소요되는 시간

• 전처리 후 학습용(모델개발용)데이터를 5개의 Fold로 나눠 검증데이터셋을 각각 달리 하는 다섯 쌍의 학습-검증데이터셋을 구성해 학습 데이터 구성만 다른 5가지의 모델 구성함.

• 모델 1개(batch size=64, epoch=7 기준), 약 9시간의 학습시간과 약 30분의 예측 시간으로 총 9시간 30분이 소요됨

• 5 Fold 모델 소프트 보팅을 통해 최종 예측값을 도출함 전처리와 보팅 포함 총 약 10시간 이내로 소요됨.

2. 알고리즘 개요

0) Intro

• 실습용 자료 내, target값에 해당하는 표준산업분류의 빈도가 imbalance함. 일부 target에 편중됨

• 텍스트 클렌징 뿐 아니라 target 데이터 imbalance 해소를 위한 데이터 증강(Data Augmentation) 시도

• 높은 성능을 보인 사전 학습된 한국어 언어모델 기반의 모델링 및 예측 진행

1) 전처리 방식

• 한국어 중심의 예측을 위해 숫자 제거 및 한글 자모음 삭제

• 일정 규칙이 확인된 특수기호(!~) 보존 / 한자어 및 일부 특수기호(₩ ; ℃) 삭제

• 면적 관련 특수기호(㎠, ㎜, ㎡ 등) 및 영어 관련 특수기호(V, G, P L, T) 대체

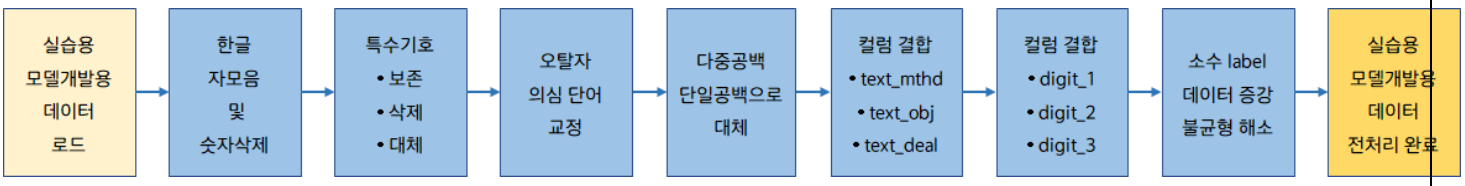
** 영어 관련 특수기호 : 텍스트 상으론 영어로 보이나 Python 코드를 통해 특수기호임을 확인한 문자*

• 오탈자 의심 단어 직접 교정(케? -> 케일 / 숙박시? -> 숙박시설 등)

• text_obj / text_mthd / text_deal 3개의 column을 순서대로 결합

• digit_1 / digit_2 / digit_3 3개의 column 결합해 label column 생성해 예측할 target 데이터 생성

• 학습데이터 100만개 중에서 label 빈도수가 20개 미만인 label의 경우 데이터 개수 3배로 증강



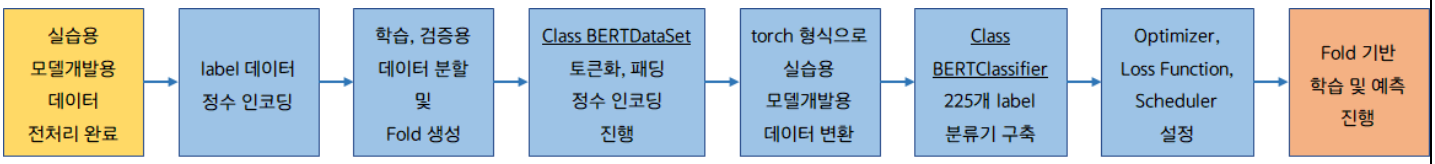
[그림3 : 전처리 도식]

2) 사전학습 언어모델 – KoBERT (출처 : https://github.com/SKTBrain/KoBERT)

• KoBERT는 트랜스포머로 구현된 양방향 자연어 처리 모델인 BERT의 한국어 성능 한계를 극복하기 위해 위키피디아나 뉴스 등을 수집한 수백만 개의 한국어 문장으로 이뤄진 대규모 말뭉치(corpus)를 추가학습한 언어모델

• 본 팀은 SKTBrain팀이 제공하는 KoBERT 모델과 토크나이저를 사용함.

• 사전학습된 KoBERT모델에 모델개발용 데이터를 전처리해 학습하는 fine-tunning하여 예측함



[그림4 : KoBERT 모델링 도식]

• 모델개발용 데이터를 StratifiedKfold를 이용해 학습, 검증용데이터로 분할

• Class BERTDataset을 정의해 전처리된 데이터가 KoBERT 모델의 입력가능 하도록 토큰화, 정수 인코딩, 패딩 진행

• KoBERT 모델의 8000여개 대규모 말뭉치(Corpus)를 기반으로 정수 인코딩

• 데이터의 각 문장을 토큰화하고, 토큰화된 문장의 최대 차원 수(max_seq_len)를 임의로 지정하고, max_seq_len보다 차원의 수가 작다면 1로 패딩해 모든 토큰화된 문장의 차원 통일함

• DataLoader를 이용해 torch 형식의 데이터로 변환

• Class BERTClassifier로 학습용 데이터의 225개의 label(소분류 기준으로 분류된 산업군 개수) 분류기를 정의함

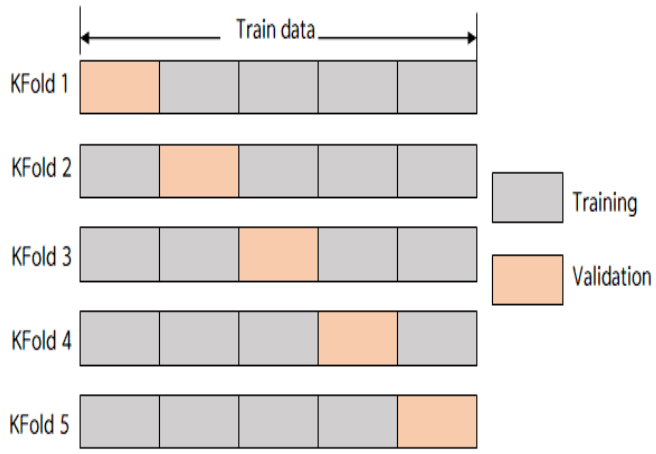
3) 모델 hyper-parameter 정보

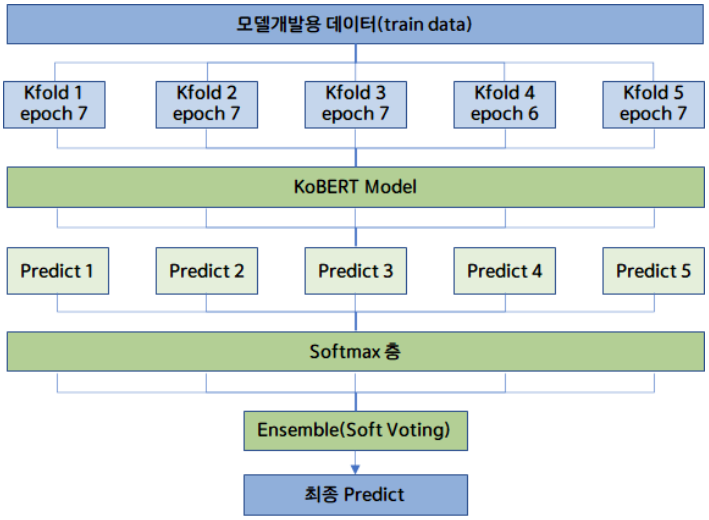
실험을 통해 최적의 파라미터 및 손실함수 설정

Loss Fuction	Drop rate	Batch size	Max seq len	Epoch	Hidden size	Max Grad Norm	Log Interval	Learning rate
Cross Entropy	0.2	64	64	6~7	768	1	200	5e-5

• StratifiedKFold : 특정 label에 편중된 불균형한 분포도를 가진 label(결정 클래스) 데이터 집합을 위한 KFold 방식

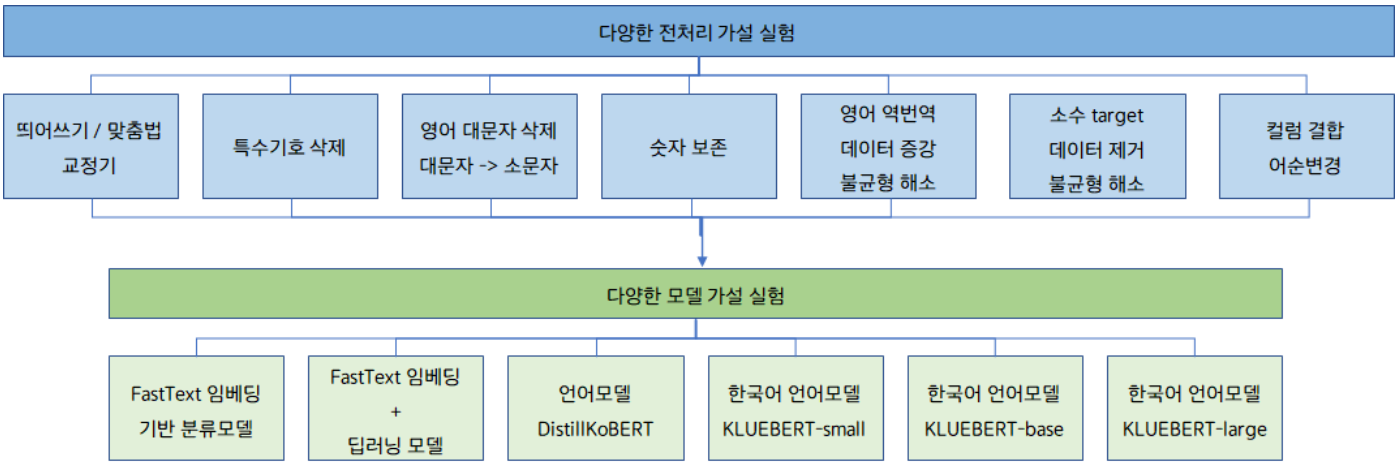
• Ensemble(Soft Voting) : 각 label 결정확률을 더한 평균 중 확률이 가장 높은 label을 최종 보팅 결괏값으로 선정함





4) 기타 코드

다양한 전처리 및 모델 가설을 실험해보았고 성능이 낮아지는 경우 해당 항목을 배제함



신청자	소속/직위/팀명	건국대학교	성 명	김유빈
	휴대전화	010-7164-6794	전자우편	kimyusintwo@gmail.com
제출일	2022년 4월 15일			