BERT를 활용한 산업분류 모델 개발.

서울대학교 빅데이터 핀테크 프로그램

[통계 데이터 사이언스]

손 민 아

DATE. 2022.03.27

CONTENTS.

01
주제 선정 배경 및
분석 목표

02
자연어 처리
- BERT 모델

03
분석 결과

04
결론 및 향후과제

01_주제 선정 배경 및 분석 목표

02_자연어 처리-BERT 모델

03_분석 결과

04_결론 및 향후과제

01

주제 선정 배경 및 분석 목표

01
주제 선정 배경

01. 통계청 통계개발원, AI 활용 '산업분류 자동화' 가능성 제시

- 통계개발원이 인공지능을 응용한 통계분류 혁신 연구를 통해 산업분류를 자동화할 수 있는 가능성 제시.
- .
- 올해 3월 '데이터 리서치 브리프 ' 6호에 '인공지능 기반 산업분류 자동화 연구' 결과 보고서가 실렸음.
- 연구진은 자연어 처리기술과 머신러닝 기반 분류 알고리즘을 응용해 해당 연구를 진행했는데
  이는 국가정책 수립/평가의 근거가 되는 핵심통계의 정확성과 시의성을 대폭 증진시킬 수 있음.
- 통계개발원장의 말에 따르면 AI 기반 통계분류 자동화 시스템이 구축되면 300여개 국내 법률에서 준용하는 산업, 직업. 질병, 사인분류 활용통계 등이 보다 정확하고 시의적으로 생산될 수 있고 범부처적으로 통계에 기반한 증거기반정책이 강화될 것 암시.

출처:https://www.news1.kr/articles/?4603071

01
주제 선정 배경

02. 자연어 처리 기반 분류 모델 선행 연구

연구 주제	저자	활용 기법
데이터 리서치 브리프_vol6(2021)	통계개발원	CBoWFCNN, 색인 DB
딥러닝 기반 한국 표준 산업분류 자동분류 모델 비교(2020)	임정우, 문현석, 이찬희, 우찬균, 임희석 (고려대학교)	CNN-LSTM
자연어처리와 기계학습을 이용한 요구사항 분석기술 비교 연구(2020)	조병선	다양한 머신러닝 모델 및 CNN

현재 대부분의 산업분류 모델로 활용한 기법으로 CCN, LSTM 등이 차지하고 있으며
BERT 기법을 활용한 분석에 대해서는 아직 연구가 더 필요한 것으로 보인다.

01
데이터 소개

- 데이터 총 100만개
- 출처: 통계청 통계데이터 인공지능 활용대회
- 대분류 : 총 19개 (한국표준산업분류 기준)

RAW 데이터

2. 모델개발용자료

AI_id|digit_1|digit_2|digit_3|text_obj|text_mthd|text_deal
id_0000001|S|95|952|카센터에서|자동차부분정비|타이어오일교환
id_0000002|G|47|472|상점내에서|일반인을 대상으로|채소.과일판매
id_0000003|G|46|467|절단하여사업체에도매|공업용고무를가지고|합성고무도매
id_0000004|G|47|475|영업점에서|일반소비자에게|열쇠잠금장치
id_0000005|Q|87|872|어린이집|보호자의 위탁을 받아|취학전아동보육

활용 데이터 예시

	AI_id	digit_1	digit_2	digit_3	text_obj	text_mthd	text_deal
0	id_0000001	S	95	952	카센터에서	자동차부분정비	타이어오일교환
1	id_0000002	G	47	472	상점내에서	일반인을 대상으로	채소.과일판매
2	id_0000003	G	46	467	절단하여사업체에도매	공업용고무를가지고	합성고무도매
3	id_0000004	G	47	475	영업점에서	일반소비자에게	열쇠잠금장치
4	id_0000005	Q	87	872	어린이집	보호자의 위탁을 받아	취학전아동보육
...	...	...	...	...	...	...	...
999995	id_0999996	C	13	134	제품입고	워싱	청바지워싱
999996	id_0999997	F	42	424	현장에서	고객의요청에의해	실내인테리어
999997	id_0999998	G	47	474	영업점에서	일반소비자에게	여성의류 판매
999998	id_0999999	P	85	856	사업장에서	일반고객을대상으로	필라테스
999999	id_1000000	I	56	561	사업장에서	접객시설을 갖추고	한식(미역구)판매

1000000 rows × 9 columns

AI_id	digit_1	digit_2	digit_3	text_obj	test_mthd	test_deal
id	대분류	중분류	소분류	무엇을 가지고 (원재료, 영업장소 등)	어떤 방법으로 (주요 영업, 생산 활동)	생산·제공하였는가 (최종 재화, 용역)
1	S	95	952	카센터에서	자동차부분정비	타이어오일교환

참고 데이터

한국표준산업분류(10차)_국문

개정 분류체계 (제10차 기준)	
A	농업, 임업 및 어업(01~03)
B	광업(05~08)
C	제조업(10~34)
D	전기, 가스, 증기 및 공기 조절 공급업(35)
E	수도, 하수 및 폐기물 처리, 원료 재생업(36~39)
F	건설업(41~42)
G	도매 및 소매업(45~47)
H	운수 및 창고업(49~52)
I	숙박 및 음식점업(55~56)
J	정보통신업(58~63)
K	금융 및 보험업(64~66)
L	부동산업(68)
M	전문, 과학 및 기술 서비스업(70~73)
N	사업시설 관리, 사업 지원 및 임대 서비스업(74~76)
O	공공 행정, 국방 및 사회보장 행정(84)
P	교육 서비스업(85)
Q	보건업 및 사회복지 서비스업(86~87)
R	예술, 스포츠 및 여가관련 서비스업(90~91)
S	협회 및 단체, 수리 및 기타 개인 서비스업(94~96)
T	가구 내 고용활동 및 달리 분류되지 않은 자가 소비 생산활동(97~98)

01
분석 목표

목표

- 통계 데이터의 산업분류 자동화 모델 발판이 될 수 있는 자연어 처리 기반의 딥러닝 분류 모델 개발
- 현재 자연어 처리의 가장 최신 기술인 Transformer의 BERT 모델 활용
- 산업 대분류 예측에 초점을 맞추어 통계개발원 데이터로 보다 좋은 성능의 다중분류 모델 개발을 목표
- 성능 지표: 다중분류 문제에 적합한 acurracy와 micro-average f1-score
- 더 나아가 데이터를 입력하면 해당 사업이 어떤 산업분류에 속하는지 알려주는 프로그램 개발

모델 적용 예시

입력 데이터

text_obj	test_mthd	test_deal
무엇을 가지고 (원재료, 영업장소 등)	어떤 방법으로 (주요 영업, 생산 활동)	생산·제공하였는가 (최종 재화, 용역)
카센터에서	자동차부분정비	타이어오일교환

출력 데이터

digit_1
대분류
S

딥러닝 산업분류 모델
Transformer KoBERT 자연어 처리
Accuracy & micro-average f1-score
산업분류 알려주는 프로그램

01_주제 선정 배경 및 분석 목표

02_자연어 처리-BERT 모델

03_분석 결과

04_결론 및 향후과제

02

자연어 처리

- BERT 모델

02
전처리 과정

Label encoding → 새로운 'text' 변수 생성 (text_obj+text_mthd +text_deal 데이터 이어붙임.) → 원본 데이터에는 결측치가 존재했지만 text 붙이고 나니 결측치 없음 확인. → Dataframe -> list 로 데이터 타입 전환 확인. → 8:2 로 Train, test dataset split

원본데이터 결측치

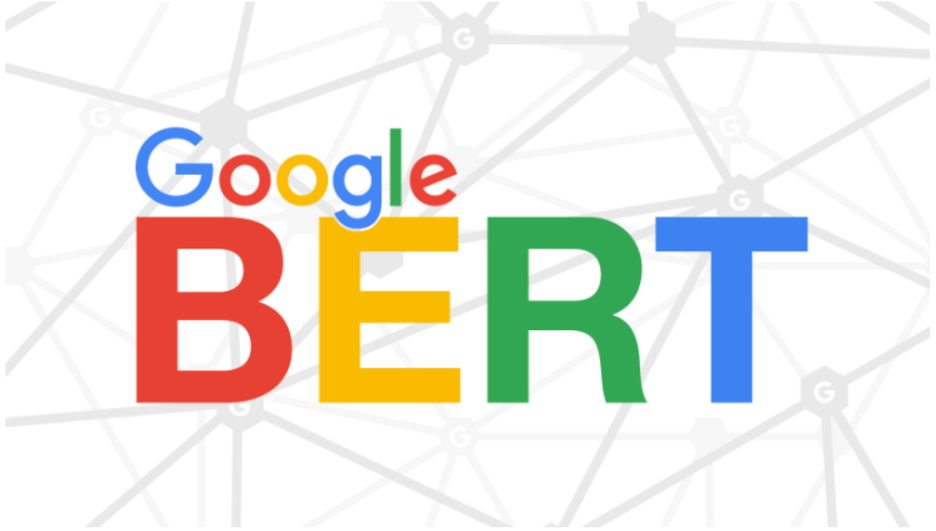
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 7 columns):
 #   Column     Non-Null Count    Dtype
---  ------     --------------    -----
 0   AI_id      1000000 non-null  object
 1   digit_1    1000000 non-null  object
 2   digit_2    1000000 non-null  int64
 3   digit_3    1000000 non-null  int64
 4   text_obj   983323 non-null   object
 5   text_mthd  956381 non-null   object
 6   text_deal  932348 non-null   object
dtypes: int64(2), object(5)
memory usage: 53.4+ MB
```

Label encoding +Text변수 생성 결과

labels	text
18	카센터에서 자동차부분정비 타이어오일교환
6	상점내에서 일반인을 대상으로 채소.과일판매
6	절단하여사업체에도매 공업용고무를가지고 합성고무도매
6	영업점에서 일반소비자에게 열쇠잠금장치
16	어린이집 보호자의 위탁을 받아 취학전아동보육
...	...

02
자연어 처리 - BERT 모델 활용

BERT :
Pre-training of Deep Bidirectional Transformers for Language Understanding



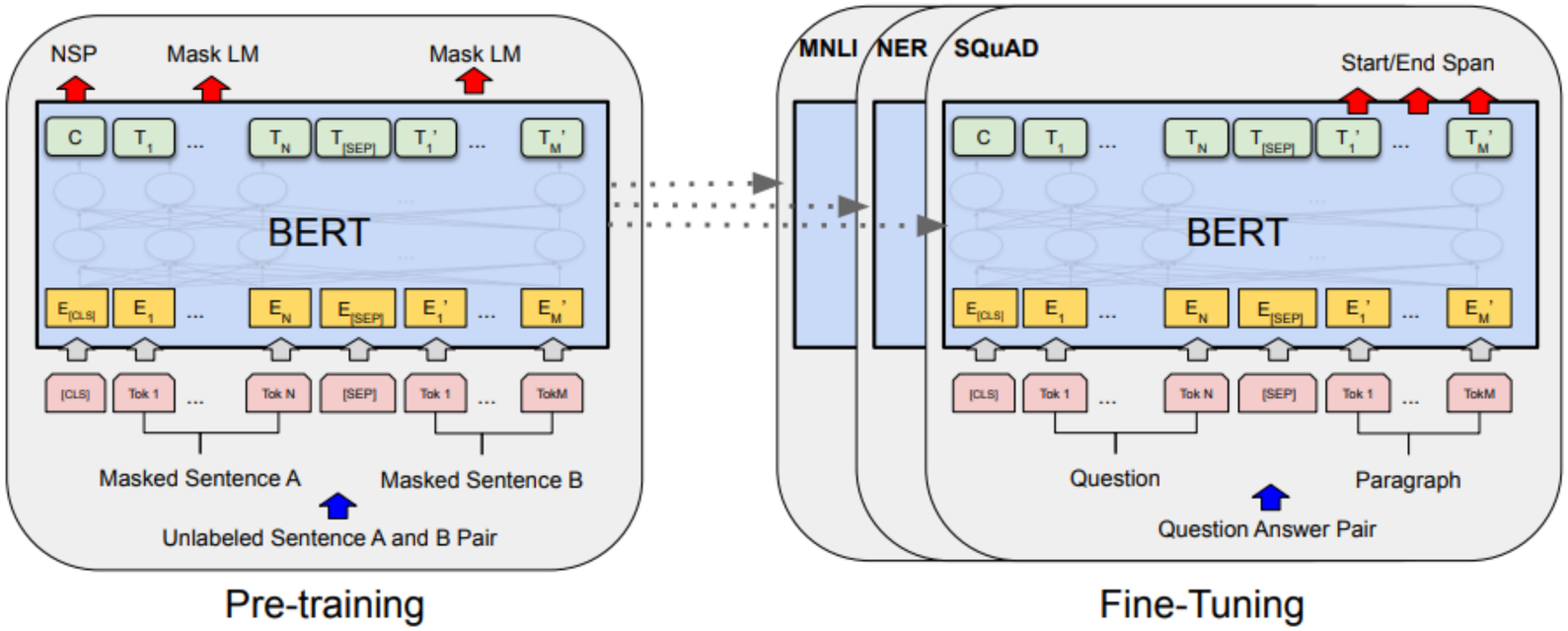
- 구글에서 개발한 NLP(자연어처리) 사전 훈련 기술이며, 특정 분야에 국한된 기술이 아니라 모든 자연어 처리 분야에서 좋은 성능을 내는 범용 Language Model.
- 11개 이상의 자연어처리 과제에서 BERT가 최첨단 성능을 발휘한다고 하지만 그 이유는 잘 알려져 있지 않지만 BERT는 지금까지 자연어처리에 활용하였던 앙상블 모델보다 더 좋은 성능을 내고 있어서 많은 관심을 받고 있는 언어모델.

출처: https://ebbnflow.tistory.com/151 [삶은 확률의 구름]

02
자연어 처리 - BERT 모델 활용

BERT의 분석 과정:

1. pre-training
2. Fine-tuning



출처: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding(2019)-Jacob Devlin(Google AI Language)

02
자연어 처리 - BERT 모델 활용

BERT의 Embedding 과정

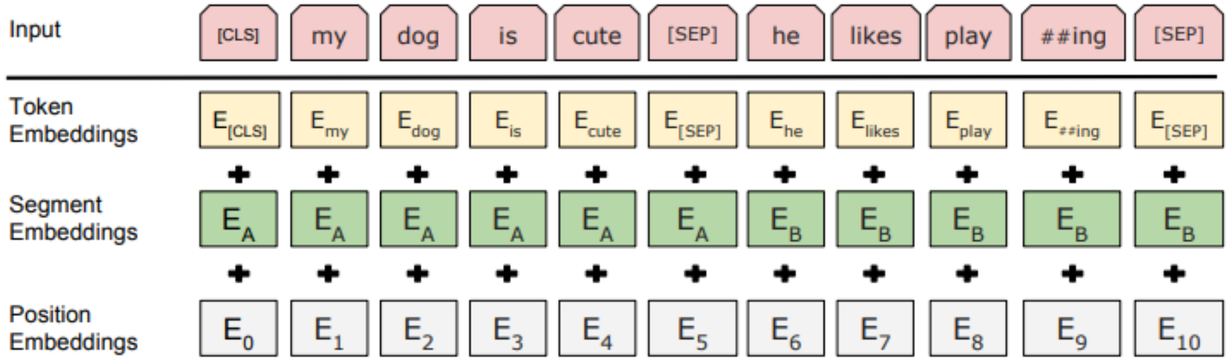


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

출처: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding(2019)-Jacob Devlin(Google AI Language)

01_주제 선정 배경 및 분석 목표

02_자연어 처리-BERT 모델

03_분석 결과

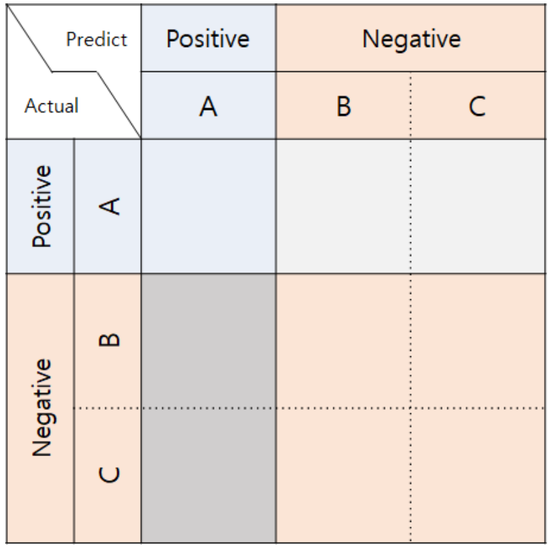
04_결론 및 향후과제

03

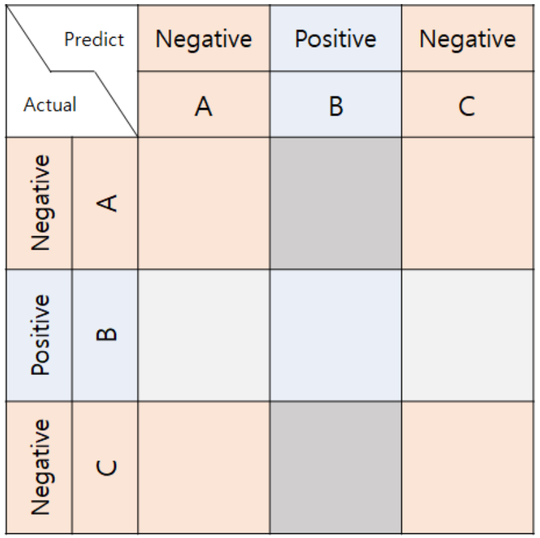
분석 결과

03
다중분류

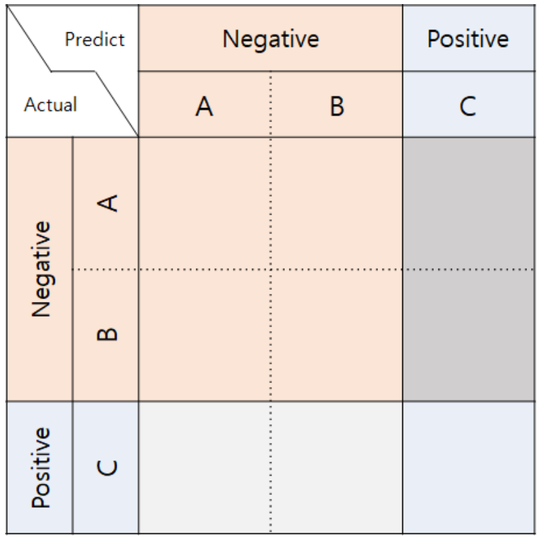
Multiclass Classification



A에 대한 이진분류



B에 대한 이진분류



C에 대한 이진분류

출처: https://rython.tistory.com/14

03
다중분류 모델의 성능 검증

성능 지표

F1-score

선행연구를 통해 accuracy 외에 micro-average f1-score 이 적절한 성능 지표가 된다고 판단.

Micro-average

$$\mathrm{TP}_{\text{total}} = \sum_{i=1}^{c} \mathrm{TP}_i$$

Macro-average

$$\text{Sensitivity} = \sum_{i=1}^{c} p(i)\text{Sensitivity}_i = \sum_{i=1}^{c} p(i)\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}$$

weighted-average

$$\text{Sensitivity} = \frac{1}{c}\sum_{i=1}^{c} \text{Sensitivity}_i = \frac{1}{c}\sum_{i=1}^{c} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}$$

- 분류 문제에서 단순히 accuracy만으로 성능을 판단하기 어렵다.
- 허나 다중 분류, 즉 multiclass classification 문제에서는 TP,FN,FP,TN을 바로 정의할 수 없다.
- class별로 TP, FN, FP, TN 값을 따로 정의하게 되고 그 값들을 이용해서 계산하는 방법에 따라 micro average 와 macro average 지표 두 가지 활용.
- 따라서 본 연구에서는 다중분류 모델의 성능지표로 accuracy와 f1-score 선택.

Sklearn에서 제공해주는 다중분류를 위한 성능 지표 f1_score

average : *{'micro', 'macro', 'samples','weighted', 'binary'} or None, default='binary'*

This parameter is required for multiclass/multilabel targets. If `None`, the scores for each class are returned. Otherwise, this determines the type of averaging performed on the data:

`'binary'`:
Only report results for the class specified by `pos_label`. This is applicable only if targets (`y_{true,pred}`) are binary.

`'micro'`:
Calculate metrics globally by counting the total true positives, false negatives and false positives.

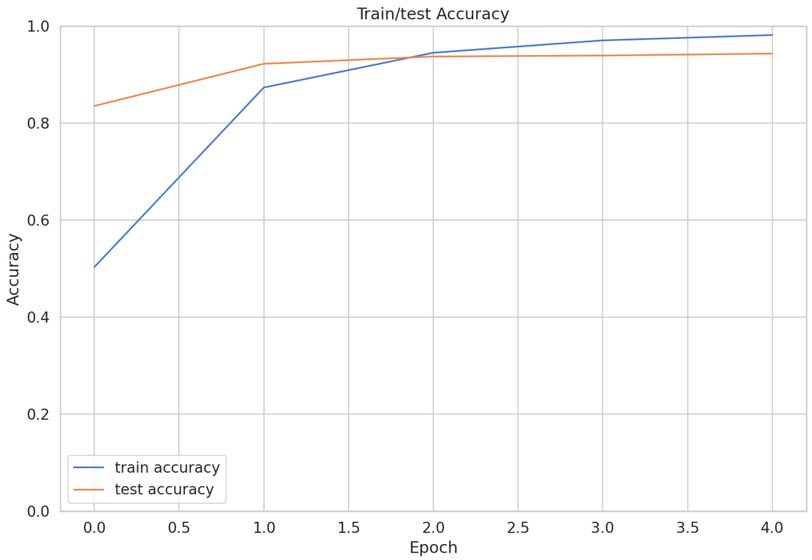
`'macro'`:
Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.

`'weighted'`:
Calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters 'macro' to account for label imbalance; it can result in an F-score that is not between precision and recall.

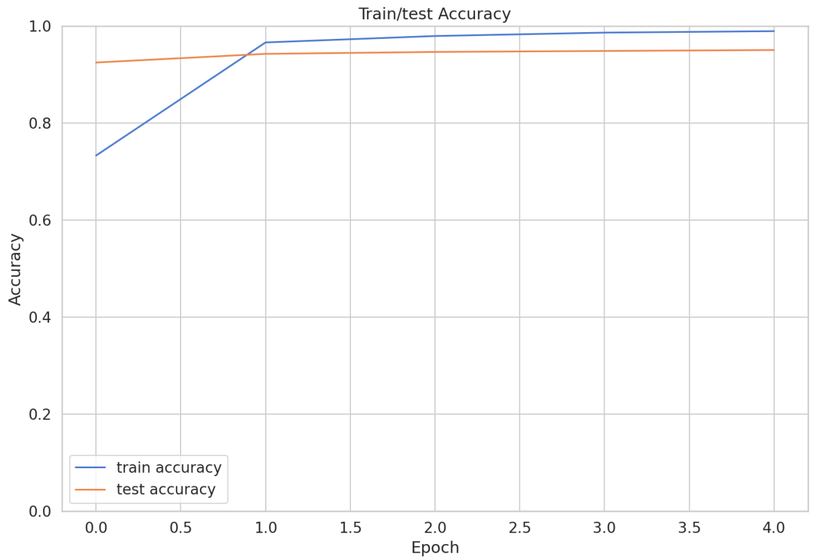
출처: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

03
분석 결과

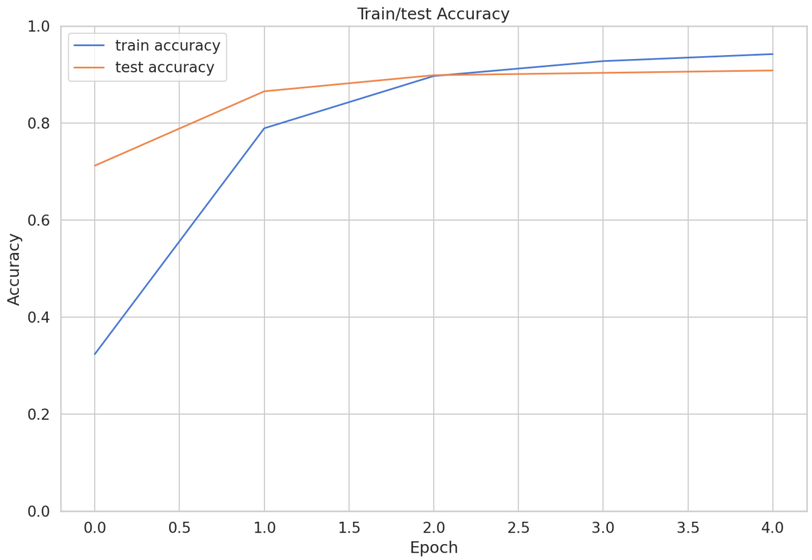
Batch size별
Train accuracy와 test accuracy 시각화



Batchsize = 8



Batchsize = 16



Batchsize = 32

03
분석 결과

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"(2019)의 저자 Jacob Devlin은 fine tuning 시

- Batch-size = 16, 32
- epoch = 2,3,5
- Learning rate = 5e-5,4e-5,2e-5

의 하이퍼파라미터 값을 추천했다.

본 연구에서는 batch-size = 16, learning rate = 2e-5, epoch =5 가 최적의 하이퍼파라미터 set 이라고 판단 후 BERT 모델에 적용하였다.

Fine-tuning결과

Hyperparameter	Accuracy	F1-score (micro average)
Batch-size = 32 Learning rate = 2e-5 Epoch = 5	0.908	0.909
Batch-size = 16 Learning rate = 2e-5 Epoch = 5	0.958	0.958
Batch-size = 8 Learning rate = 2e-5 Epoch = 5	0.943	0.943

정확도: 95.8%
F1-score : 95.8%

- BERT 모델의 Fine-tuning 과정에서 learning rate, batchsize, epoch 위주로 tunin을 해본 결과 batch size 튜닝 시 가장 눈에 띄는 변화를 관찰함.
- Epoch과 learning rate 를 고정 후, batch size만 8, 16, 32 로 조정하면서 더 자세한 변화를 관찰해본 결과, **batch size 가 16일 때, 해당 모델이 가장 좋은 성능을 보여줌.**

03
분석 결과

다중 분류 결과

대분류 예측 결과 비교

text	labels	label_answer
개인택시로 일반인을 대상으로 승객운송서비스	7	7
주점에서 접객시설을 갖추고 주류 판매	8	8
취재.교정하여 간행물업체에서 주간지발행	9	12
의류임가공 제조 청바지	2	2
영업장에서 접객시설을갖추고 탁주	8	8
공장에서 제조 용접형강	2	2
개인택시로 일반인을 대상으로 승객운송서비스	7	7
교회에서 기독교계통의종교활동 종교서비스	18	18
매장에서 일반소비자에게 판매(셔츠, 청바지)	6	6
교회에서 종교인 대상으로 종교활동	18	18
음식점에서 접객시설갖추고 짜장면	8	8
개인택시로 일반인을 대상으로 승객운송서비스	7	7

랜덤하게 선택한 sample의 예측 결과 정확도가 비교적 높다는 것을 알 수 있음.

대분류 예측 결과를 출력해주는 프로그램 결과

```
>> 해당 사업은 예술, 스포츠 및 여가 관련 서비스업 산업분류에 속합니다.
>> 해당 사업은 숙박 및 음식점업 산업분류에 속합니다.
>> 해당 사업은 사업시설 관리, 사업지원 및 임대 서비스업 산업분류에 속합니다.
>> 해당 사업은 협회 및 단체, 수리 및 기타 개인 서비스업 산업분류에 속합니다.
>> 해당 사업은 숙박 및 음식점업 산업분류에 속합니다.
>> 해당 사업은 숙박 및 음식점업 산업분류에 속합니다.
>> 해당 사업은 교육 서비스업 산업분류에 속합니다.
>> 해당 사업은 예술, 스포츠 및 여가 관련 서비스업 산업분류에 속합니다.
>> 해당 사업은 운수 및 창고업 산업분류에 속합니다.
>> 해당 사업은 협회 및 단체, 수리 및 기타 개인 서비스업 산업분류에 속합니다.
>> 해당 사업은 운수 및 창고업 산업분류에 속합니다.
>> 해당 사업은 교육 서비스업 산업분류에 속합니다.
>> 해당 사업은 운수 및 창고업 산업분류에 속합니다.
>> 해당 사업은 숙박 및 음식점업 산업분류에 속합니다.
>> 해당 사업은 숙박 및 음식점업 산업분류에 속합니다.
>> 해당 사업은 운수 및 창고업 산업분류에 속합니다.
```

원하는 출력 프로그램 만들기 성공!

01_주제 선정 배경 및 분석 목표

02_자연어 처리-BERT 모델

03_분석 결과

04_결론 및 향후과제

04

결론 및 향후과제

04
결론

결론

- **본 연구는 통계개발원에서 제공한 데이터로 통계 데이터의 산업분류 자동화 모델의 발판이 될 수 있는 BERT 다중분류 모델로 보다 좋은 분류 성능을 이끌어냈다는 것에 의의가 있다.**
- 본 연구에서는 산업분류 대분류 코드 예측에 초점을 맞추어 통계개발원 데이터로 보다 좋은 성능의 모델을 개발함
- 현재 자연어 처리의 가장 최신 기술인 Transformer의 BERT 모델을 활용함
- 데이터를 입력하면 해당 사업이 어떤 산업분류에 속하는지 알려주는 프로그램 개발함
- 다중분류 문제에 적합한 acurracy와 micro-average f1-score를 기준으로 좋은 성능 결과를 도출함 accuracy = 95.8%, f1-score= 95.8%
- 인적자원/일을 하기 위해 할애되는 시간과 비용 절약 가능
- 해당 모델을 향후 산업분류 분야뿐만 아니라 다양한 자연어 기반의 다중분류 문제에 적용가능

04
향후과제

향후 과제

- 더 좋은 컴퓨팅 하드웨어로 중,소분류 예측까지 분석을 진행
- 더 자세한 오류분석 진행
- 활용한 데이터가 불균형 데이터인데 oversampling으로 balanced data를 만들어 분석을 진행한 결과 굉장히 좋지 않은 성능 관찰
  따라서, 불균형 데이터 연구가 좀 더 필요함
- Validation set 까지 나눠서 6:2:2 비율로 분석을 시도해보았으나 좋지 않은 성능 결과로 train과 test 데이터셋만 나눠서 분석 진행함
  이 부분도 좀 더 연구가 필요함
- 어떤 산업에 적용할 수 있는지 더 고민해볼만 함

손 민 아
통계 데이터 사이언스
BERT를 활용한 산업분류 모델 개발