DACON
DATA TO VALUE

AI야, 진짜 뉴스를 찾아줘! AI 경진대회

2021.01.29

닉네임이제일어려워 팀

NH투자증권

목 차

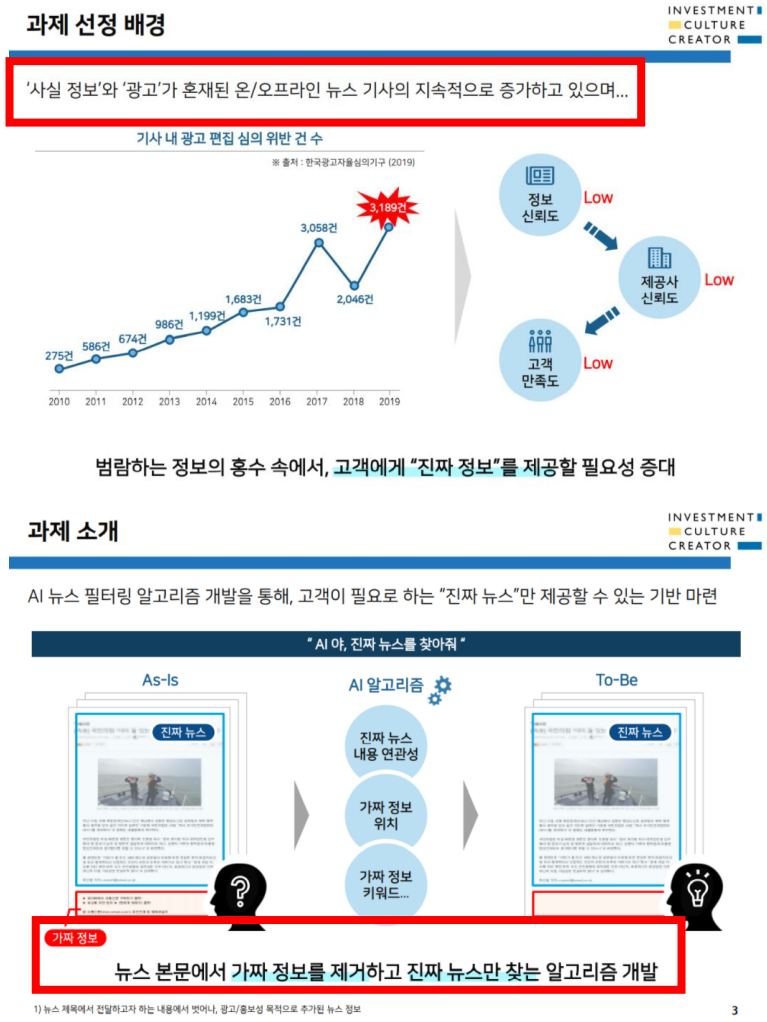
DACON
DATA TO VALUE

NH투자증권

EDA

DACON
DATA TO VALUE

1. 경진대회 주제 분석



출처 : 공개된 데이터 명세
https://dacon.io/competitions/official/235658/talkboard/401869?page=1&dtype=recent&ptype=pub

가짜 뉴스?

```
train_csv.query('info == "1"').head(20)
```

	content	ord	info
	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	2	1
	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	3	1
	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	13	1
	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	14	1
	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	24	1
	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	25	1
	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	13	1
	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	14	1
	온라인결제 관련주가 코로나19 사태의 최대 수혜주라는 평가가 나왔다. 언택트 소비 ...	1	1
	한편, 스탁론에 대한 관심이 날로 높아지고 있다. 모처럼 잡은 투자기회를 놓치지 않...	2	1
	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	3	1
	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	4	1
	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	4	1
	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	5	1
	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	19	1
	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	20	1
	고수의 관심종목 무료 카톡방 바로가기	1	1
	미국의 자율주행 및 전기차 기업인 테슬라가 역사상 신고가를 달성하면서 전기차에 대한...	2	1
	다음에 바로 갈 2차전지주 받아보기	3	1
	하지만 대부분의 투자자들은 삼성SDI, LG화학, 일진머티리얼즈 등이 이미 선제적이...	4	1

가짜 뉴스 = 광고성 문구

NH투자증권

DACON
DATA TO VALUE

2. 데이터 EDA

데이터의 사용 목적 : 가짜뉴스 판별

컬럼 사용 O

제목은 뉴스 내용을 요약한 정보

판별할 기사 데이터

진짜 뉴스, 가짜 뉴스를 담은 정보

NO.	컬럼명	컬럼 설명	예시
1	date	뉴스 발행 날짜	20200626
2	n_id	뉴스 Index 번호	NEWS09727
3	ord	뉴스 내용 순서	5
4	title	뉴스 제목	롯데·공영 등 7개 TV 홈쇼핑들, 동행세일 동참
5	content	뉴스 내용	이번 동행세일에서는 롯데·공영·CJ·현대 ……
6	info	진짜뉴스 유무	0

컬럼 사용 X

뉴스 발행 날짜와 관련이 없어야 함

단일 문장으로 뉴스 판별이 가능

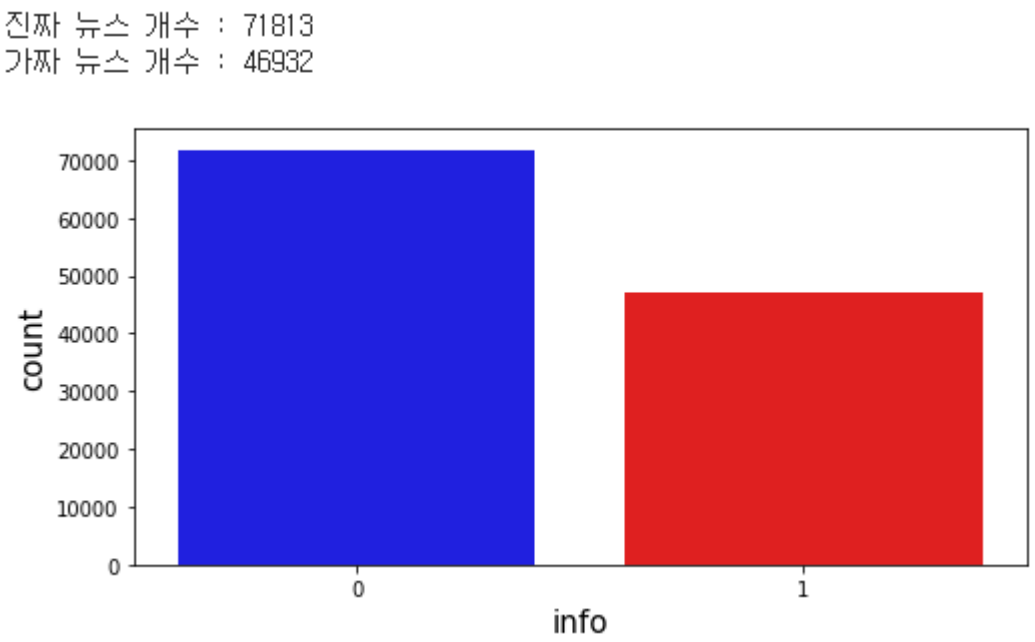
박영선 장관은 이번 동행세일 행사에 TV홈쇼핑사의 동참을 통해 내수 활성화에 한발짝...	0
한편 앞서 중기부에 따르면 지난 25일 글로벌 쇼트 비디오 앱 틱톡(Tiktok)의...	0
"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1

NH투자증권

DACON
DATA TO VALUE

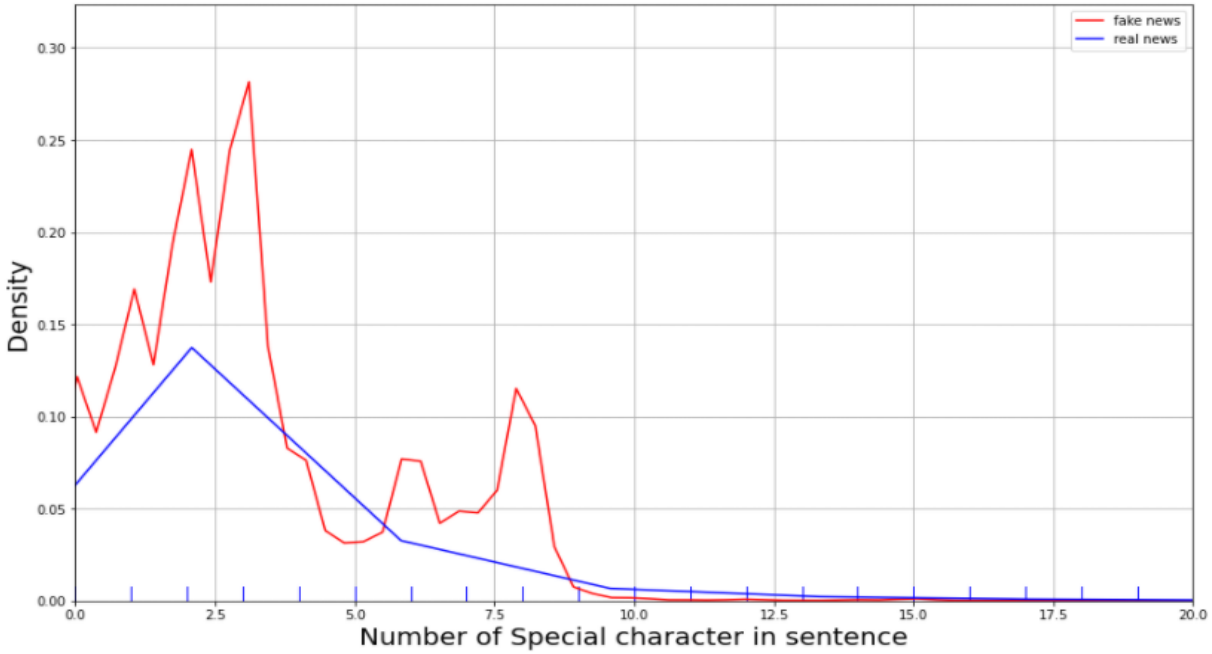
3. 데이터 시각화

진짜 가짜 개수



진짜 뉴스와 가짜 뉴스의 비율 확인

뉴스 내용에 특수문자가 포함되는 비율



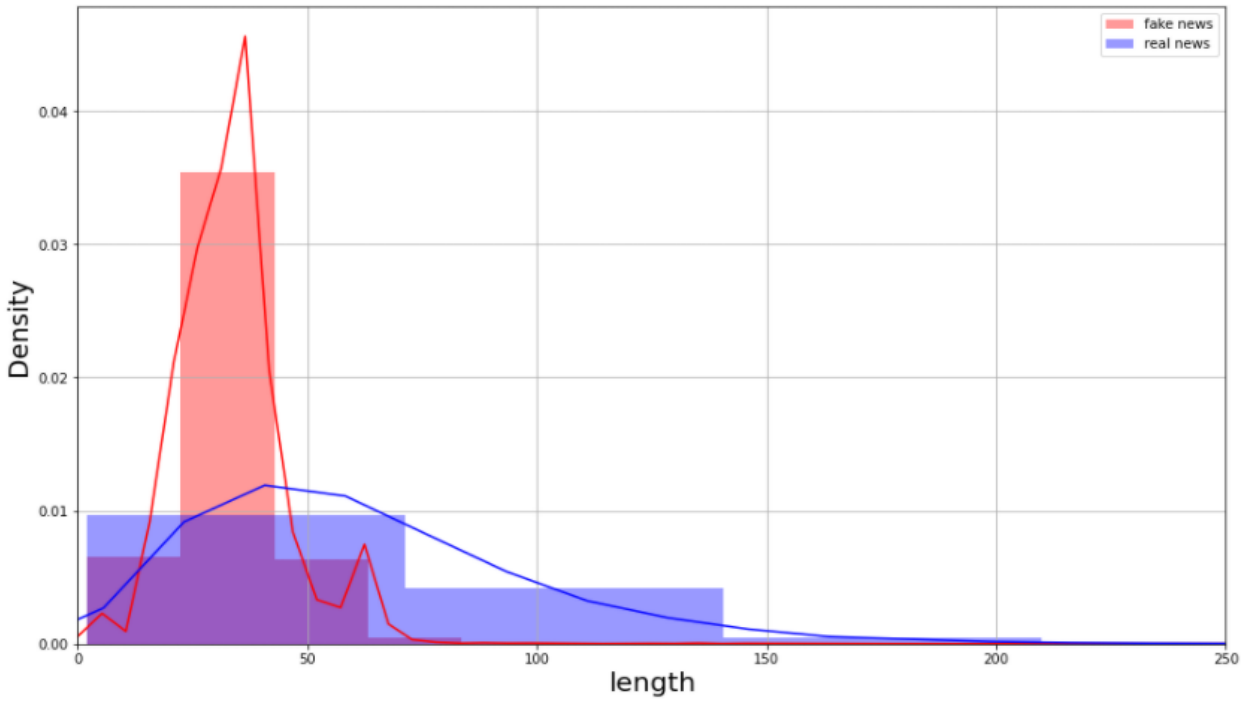
가짜 뉴스가 특수문자를 더 많이 포함

NH투자증권

EDA

3. 데이터 시각화

진짜 가짜 문장 길이 비교



광고성 문구의 길이는 0 ~ 50자 이내에 집중적으로 분포

제목과 가짜뉴스 데이터

	title	content	info
1	[마감]코스닥 기관 678억 순매도	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
2	[마감]코스닥 기관 678억 순매도	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
16	롯데·공영 등 7개 TV 홈쇼핑들, 동행세일 동참	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
17	롯데·공영 등 7개 TV 홈쇼핑들, 동행세일 동참	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
42	13년만에 늦깎이 개발 '양주 회천' 봄볕 드나	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
43	13년만에 늦깎이 개발 '양주 회천' 봄볕 드나	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
57	BMW코리아, 온라인 한정판 `M340i 퍼스트 에디션` 출시	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
58	BMW코리아, 온라인 한정판 `M340i 퍼스트 에디션` 출시	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
60	온라인 결제株, 코로나19 사태로 최대 수혜를?	온라인결제 관련주가 코로나19 사태의 최대 수혜주라는 평가가 나왔다. 언택트 소비 ...	1
61	온라인 결제株, 코로나19 사태로 최대 수혜를?	한편, 스탁론에 대한 관심이 날로 높아지고 있다. 모처럼 잡은 투자기회를 놓치지 않...	1
62	온라인 결제株, 코로나19 사태로 최대 수혜를?	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
63	온라인 결제株, 코로나19 사태로 최대 수혜를?	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
68	거래소 "이에스브이 실질심사 대상 결정 조사기간 연장"	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
69	거래소 "이에스브이 실질심사 대상 결정 조사기간 연장"	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
89	백화점의 역설...코로나에도 전남도민들은 소비 늘렸다	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
90	백화점의 역설...코로나에도 전남도민들은 소비 늘렸다	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
92	▶ 2차전지 향후 전망과 종목	고수의 관심종목 무료 카톡방 바로가기	1
93	▶ 2차전지 향후 전망과 종목	미국의 자율주행 및 전기차 기업인 테슬라가 역사상 신고가를 달성하면서 전기차에 대한...	1
94	▶ 2차전지 향후 전망과 종목	다음에 바로 갈 2차전지주 받아보기	1
95	▶ 2차전지 향후 전망과 종목	하지만 대부분의 투자자들은 삼성SDI, LG화학, 일진머티리얼즈 등이 이미 선제적이...	1

제목과 가짜 뉴스는 **한 눈에 구별**이 가능

문제 해결을 위한 접근

DACON
DATA TO VALUE

1. 데이터 전처리 – 조사 사용 여부

- 한국어 처리 Task에서는 품사를 분리 후 조사를 제외하고 사용하는 경우가 많음
- 그러나 뉴스 내용을 분석한 결과 조사도 가짜 뉴스를 구분하는 특징이 될 수 있다고 판단

➡ 뉴스 내용에 대한
조사 제거 전처리 진행 X

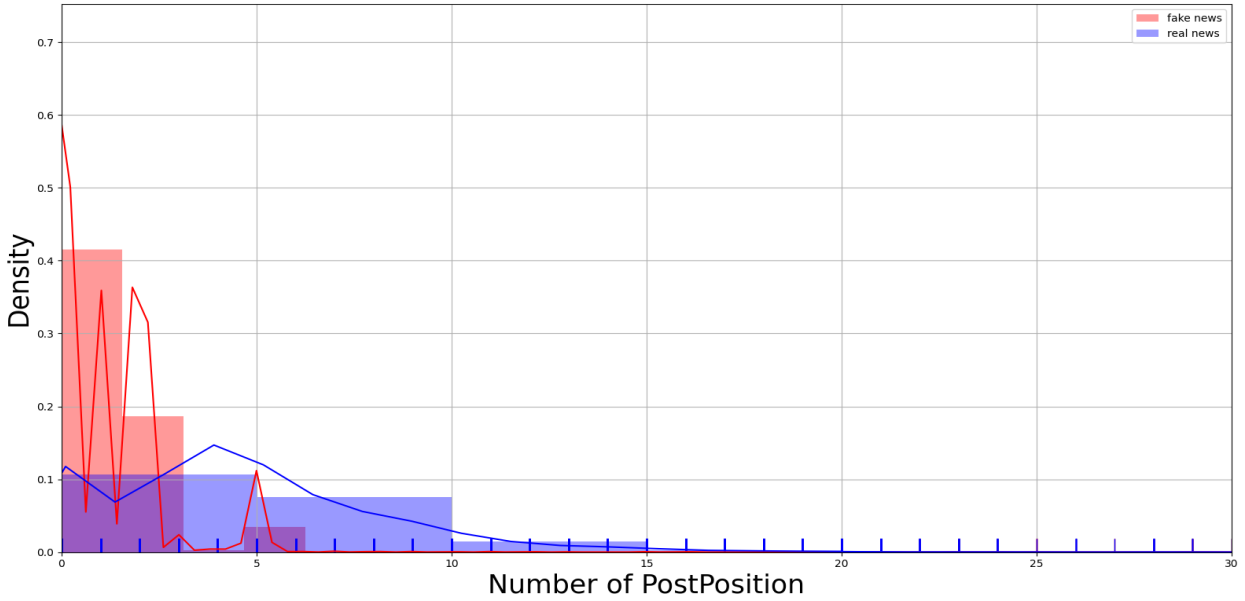
진짜 뉴스

	content	info
0	[이데일리 MARKETPOINT]15:32 현재 코스닥 기관 678억 순매도	0
3	종합 경제정보 미디어 이데일리 - 무단전재 & 재배포 금지	0
4	전국적인 소비 붐 조성에 기여할 예정	0
5	[이데일리 권오석 기자] 중소벤처기업부(이하 중기부)는 대한민국 동행세일에 7개 T...	0
6	대한민국 동행세일은 라이브 커머스, 언택트 콘서트, O2O 행사 연계 등 비대면이라...	0
...	...	...
118696	양수금액은 89억4565만원이며 이는 총자산대비 11.54%, 자기자본대비 13.8...	0
118719	[헤럴드경제=증권부] 모나리자는 사업다각화를 위해 위생용품 제조판매업체인 중원 주식...	0
118720	[헤럴드경제=증권부] 모나리자는 사업다각화를 위해 위생용품 제조판매업체인 중원 주식...	0
118721	양수금액은 89억4565만원이며 이는 총자산대비 11.54%, 자기자본대비 13.8...	0
118722	양수금액은 89억4565만원이며 이는 총자산대비 11.54%, 자기자본대비 13.8...	0

가짜 뉴스

	content	info
1	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
2	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
16	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
17	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	1
42	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
...	...	...
118740	미 FDA 임상3상 허가 임박. 묻고 따블로 갈 바이오 황제주.	1
118741	똑똑해진 소비자..한국도 이젠 소형차 시대	1
118742	똑똑해진 소비자..한국도 이젠 소형차 시대	1
118743	2020년 한국 TV 2대중 1대 인터넷 연결된다	1
118744	2020년 한국 TV 2대중 1대 인터넷 연결된다	1

뉴스 내용에서 조사의 개수



➡
- ✓ 가짜 뉴스 조사의 개수는 대부분 0 ~ 5자 사이에 위치
- ✓ 진짜 뉴스 조사의 개수는 0 ~ 15자 사이에 고르게 분포

NH투자증권

문제 해결을 위한 접근

DACON
DATA TO VALUE

1. 데이터 전처리 – 형태소 단위 Tokenizer

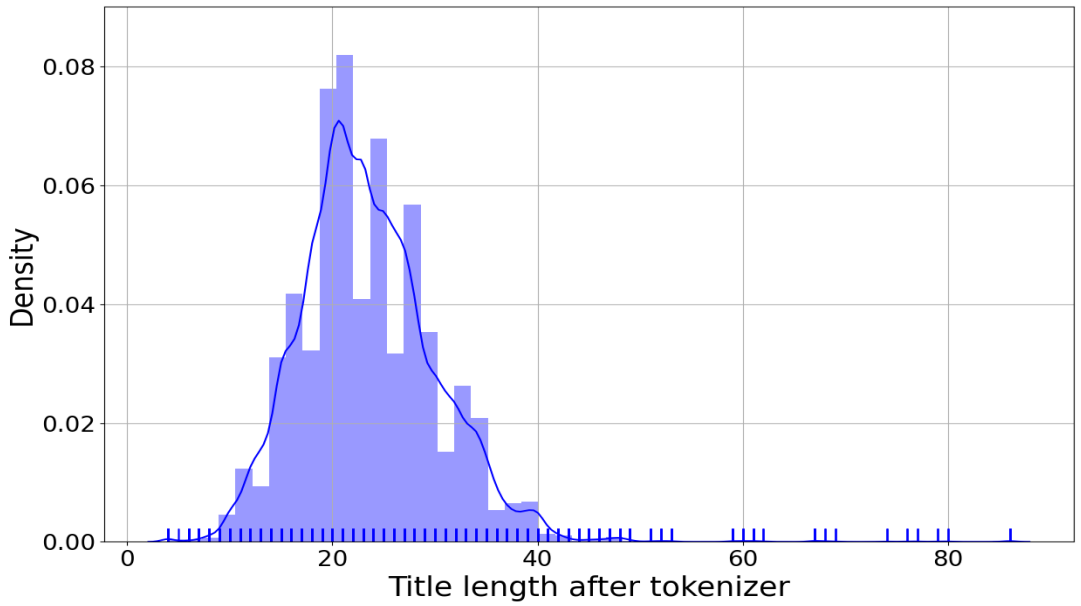
- 뉴스의 제목과 내용을 각각 형태소 단위로 Tokenize 진행
- 토큰화 된 Sentence의 길이는 제목이 0~40자, 내용은 0~80자에 대부분 위치
- 광고성 문구는 Sentence의 첫 부분만 확인해도 대부분 한 눈에 구분이 가능한 특징을 가짐

✓ 형태소 단위 Tokenize 진행 후의 문장의 길이를 Input length로 선언

✓ 제목 + 내용의 길이는 120자에 대부분 위치

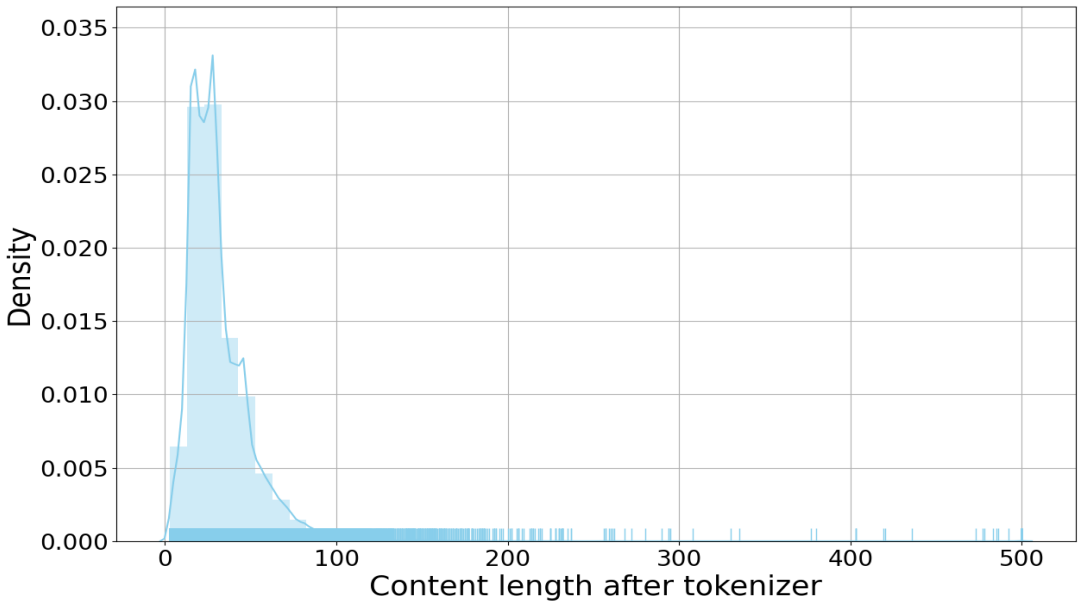
✓ Input length의 길이를 128자로 지정

Tokenizer 이후 Title의 길이



대부분 0 ~ 40자 이내에 위치

Tokenizer 이후 Content의 길이



대부분 0 ~ 80자 이내에 위치

NH투자증권

문제 해결을 위한 접근

2. 모델 소개

Model name : Google BERT

- Transformer Model Encoder 기반
- 다양한 자연어 처리(NLP) Task에서 가장 좋은 성능을 보임
- 최근 연구되는 NLP Model은 대부분 BERT Model 기반
- Pre-training Model을 Fine-Tuning 해서 다양한 NLP Task에서 사용 가능

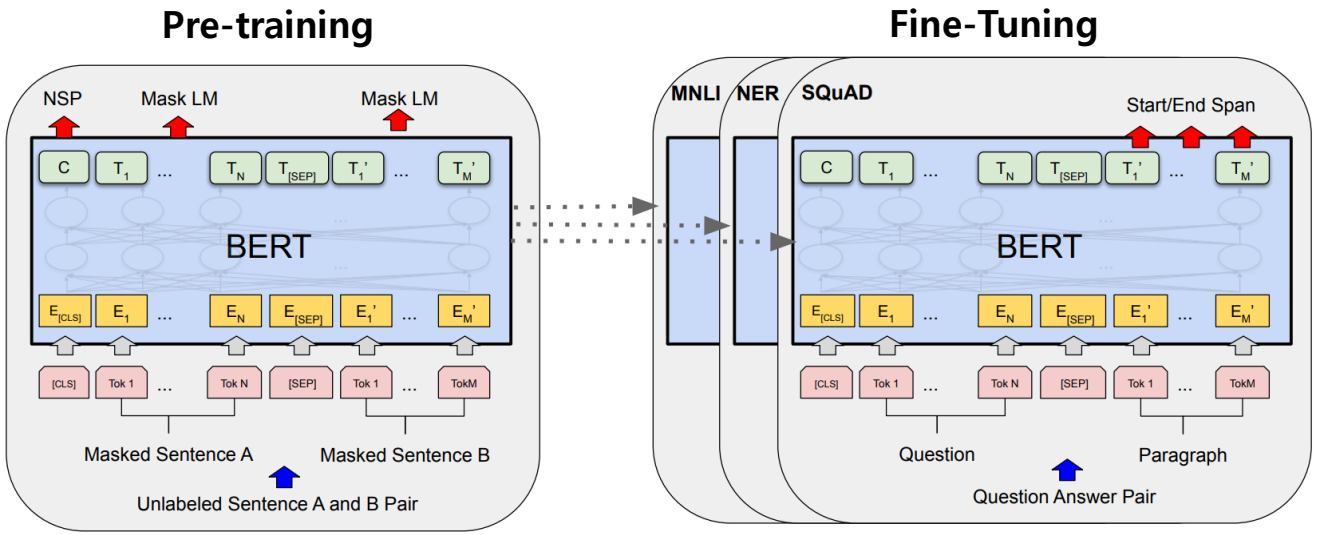


그림. BERT의 Pre-training, Fine-Tuning의 Architecture [1]

그림 설명

- Pre-training과 Fine-Tuning 모두 동일한 Architecture를 사용
- Pre-trained model parameters를 사용하여 Fine-Tuning 진행
- Pre-trained model에 Layer를 추가하여 Fine-Tuning 진행
- Fine-Tuning 진행 시, model의 모든 parameter가 미세 조정됨

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.

문제 해결을 위한 접근

DACON
DATA TO VALUE

2. 모델 소개

1) BERT Model의 사용 이유?

- BERT는 문맥에 따라 동음이의어를 판단할 수 있어 문장의 이해도가 높음
- 사전 훈련된 Pre-trained language model을 사용함으로써
  목적에 따라 Fine-tuning 진행 시 적은 Epoch로도 좋은 성능을 보임

2) KoBERT : 한국어 BERT Model

- Bert Model을 한국어로 훈련시킨 Model
- Bert-Multilingual Model에 비해
  한국어 NLP task 성능이 뛰어남
- SK T-Brain에서 제공한
  한글 위키 + 뉴스 텍스트 기반으로 훈련시킨
  한국어 Pre-trained Bert Model

❖ **BERT Model과 KoBERT Model 성능 비교**
❖ **주제 : 네이버 영화 리뷰 감정 분석 데이터를 이용한 감정 분류**

Model	Model 설명	Accuracy
BERT Model (multilingual cased)	다양한 언어를 지원하는 BERT Model	0.875
KoBERT	**한국어로 훈련시킨 BERT Model**	**0.901**

➢ 데이터 출처 : https://github.com/SKTBrain/KoBERT
➢ 네이버 영화 리뷰 감정 분석 데이터 : 영화 리뷰 데이터를 긍정 또는 부정으로 분류하는 목적으로 만든 데이터셋

NH투자증권

문제 해결을 위한 접근

DACON
DATA TO VALUE

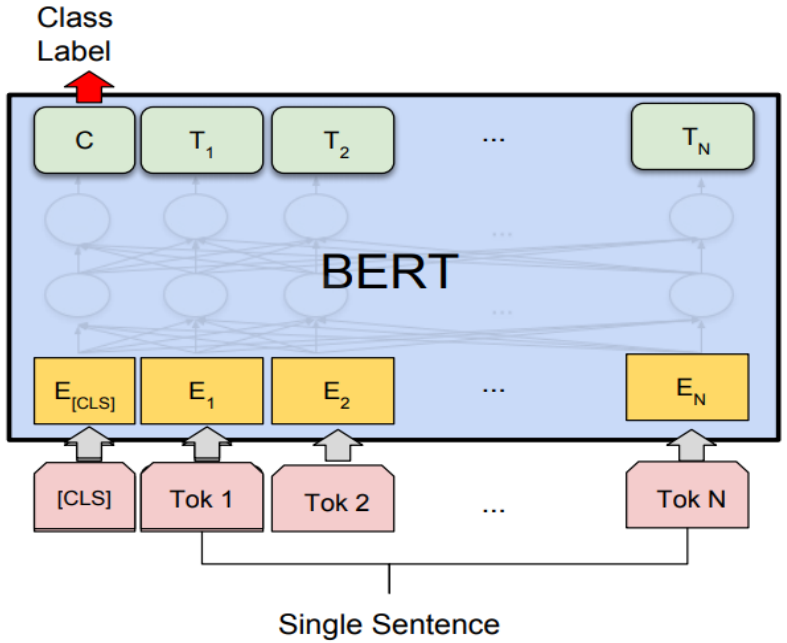
3. Method of KoBERT Fine-tuning

Method : Pre-trained KoBERT Model을 두 가지 방식의 Fine-tuning 진행

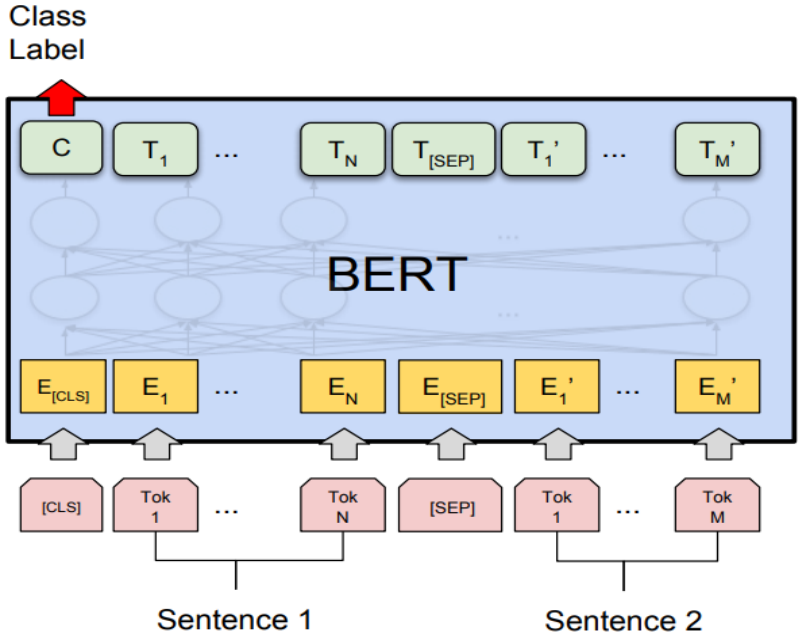
1) 주어진 단일 Sentence가 진짜 뉴스 또는 가짜 뉴스인지 구분하는 Single Sentence Classification Task 수행

2) 두 문장 A, B가 주어질 때 B가 A의 뒤에 오는 것이 적절한지 구분하는 Sentence Pair Classification Task 수행

1) Single Sentence Classification Tasks [1]



2) Sentence Pair Classification Tasks [1]



NH투자증권

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.

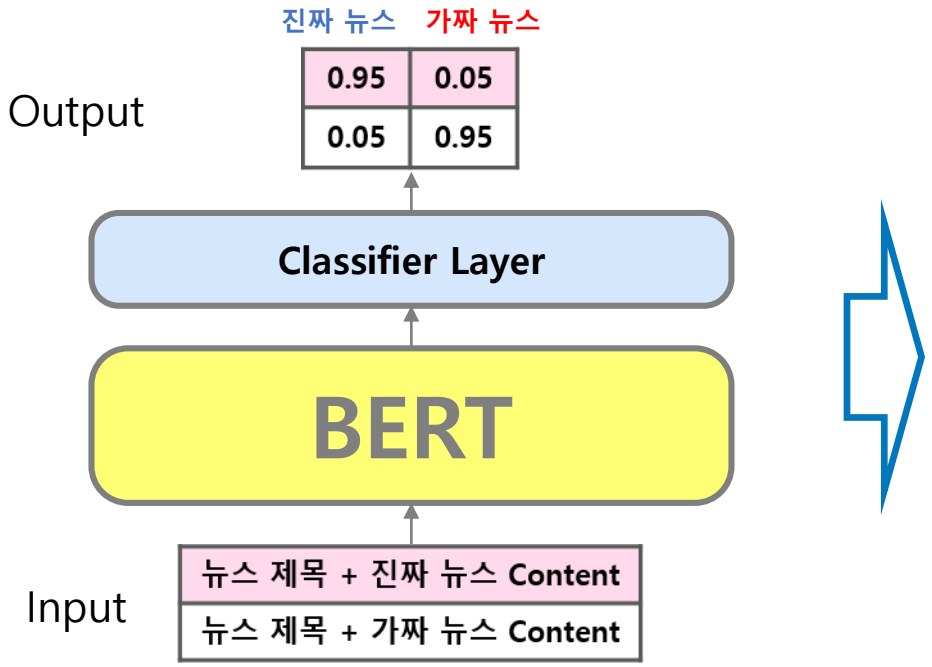
실 험

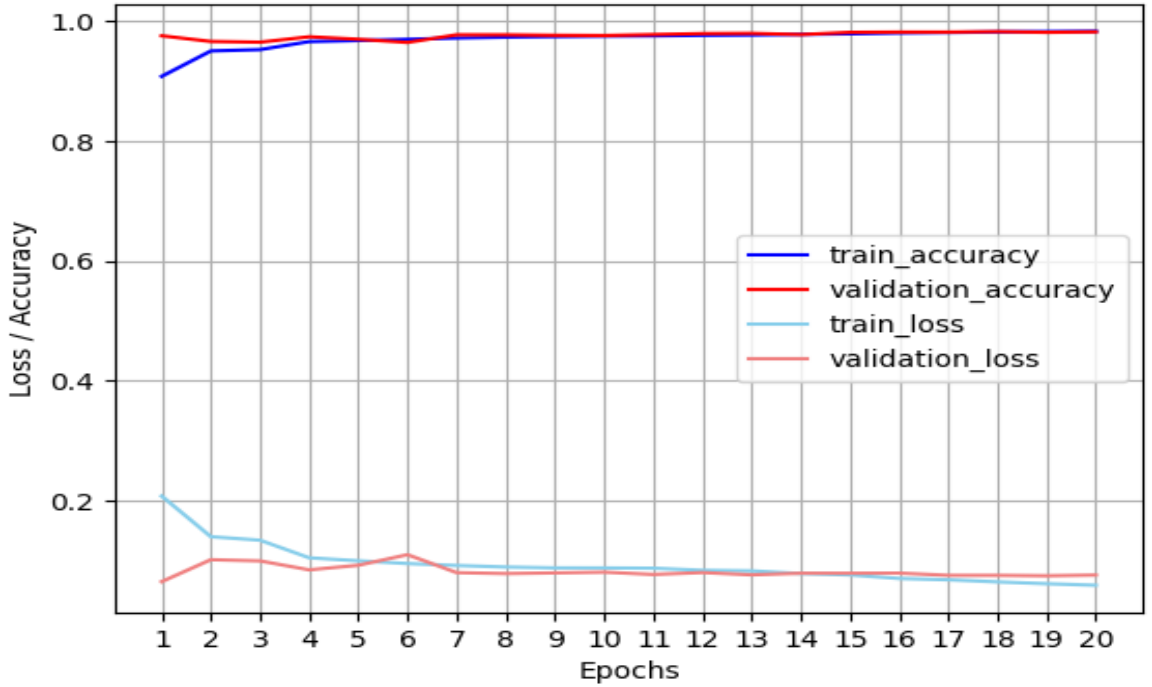
1. Fine-tuning (Single Sentence Classification Tasks)

1-1. Classification (input = 뉴스 제목 + 내용)

- Input으로 뉴스 **제목 + 내용**을 사용하여 Fine-tuning 진행

Train only 1 Epoch
- Validation Accuracy = 0.9761
- Validation Loss = 0.0648





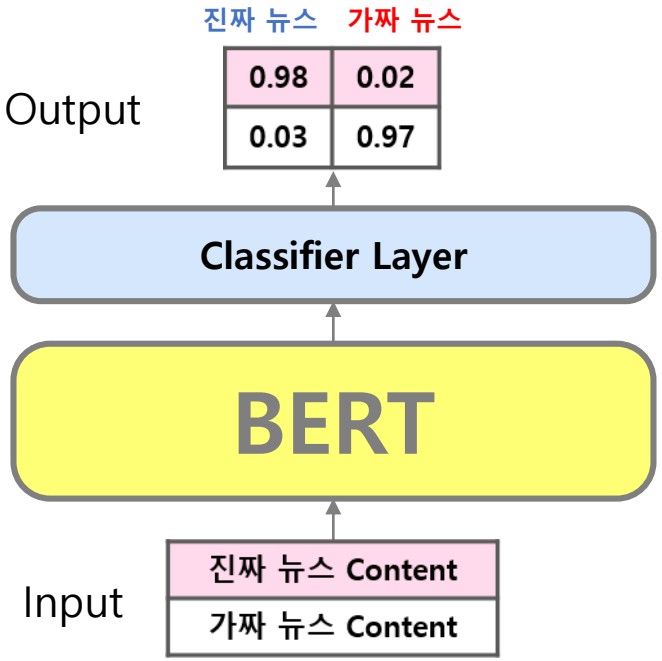
실 험

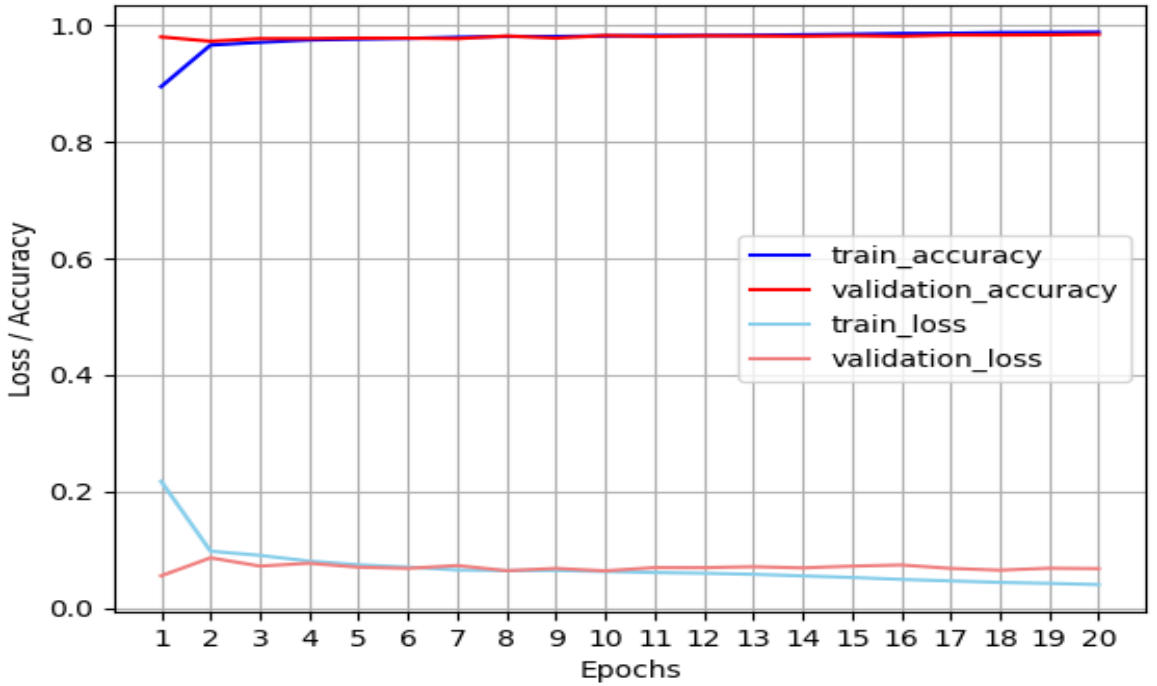
1. Fine-tuning (Single Sentence Classification Tasks)

1-2. Classification (input = 뉴스 내용)

- Input으로 뉴스 **내용**만 사용하여 Fine-tuning 진행

Train only 1 Epoch
- Validation Accuracy = 0.9806
- Validation Loss = 0.0552



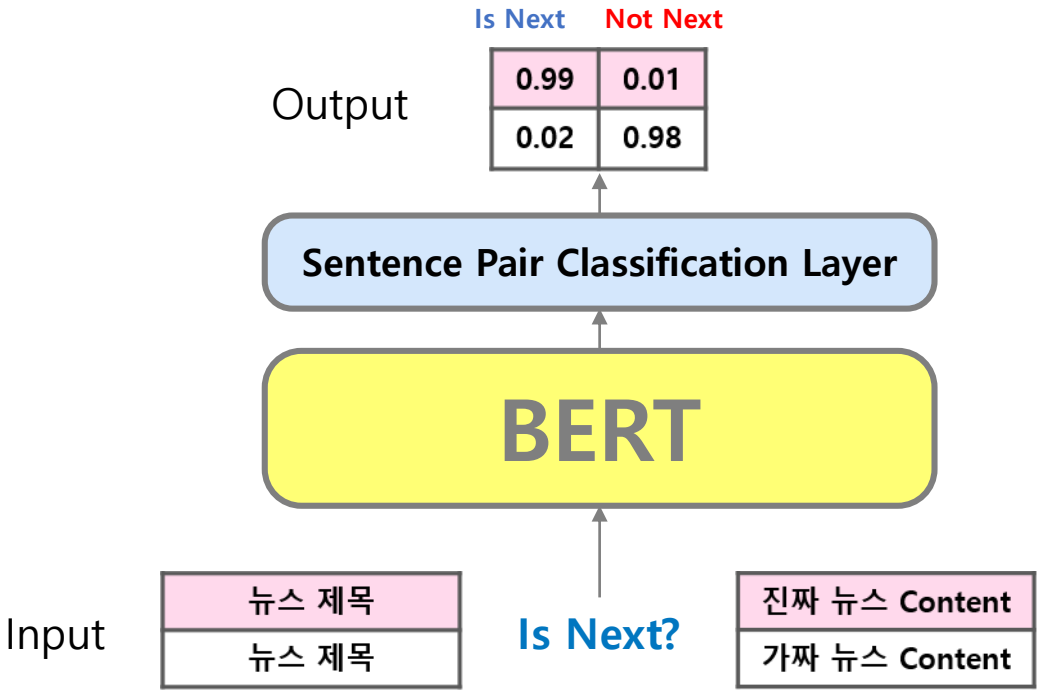


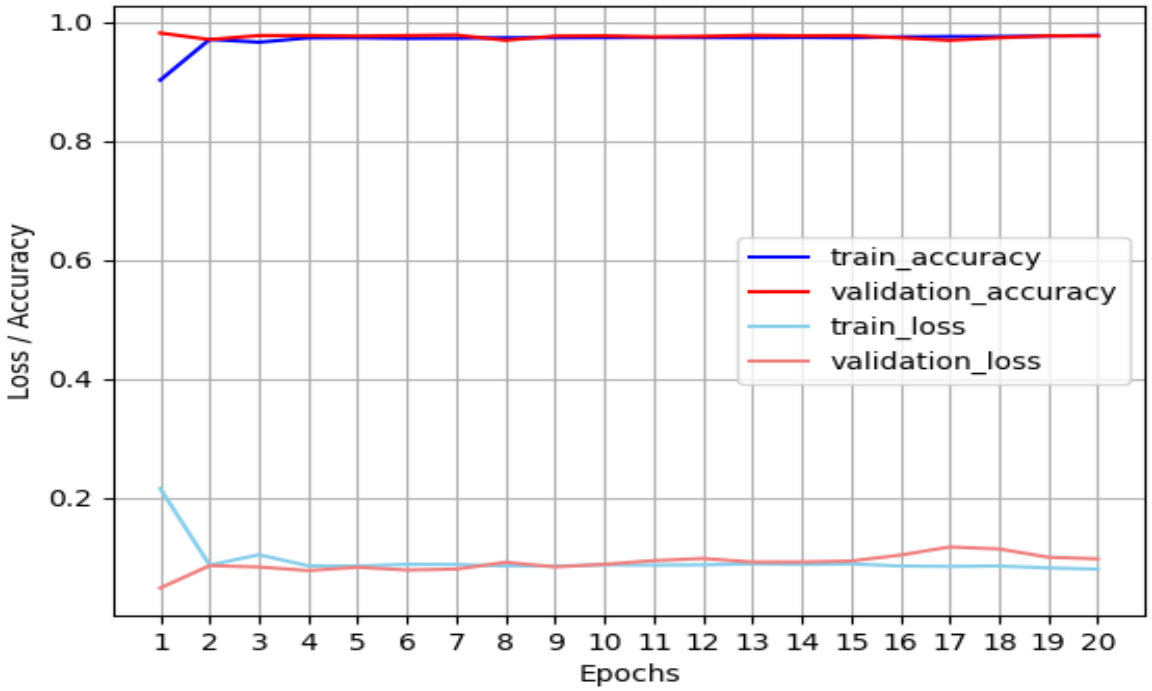
실 험

2. Fine-tuning (Sentence Pair Classification Tasks)

- **첫번째 문장은 뉴스 제목, 두번째 문장은 뉴스 내용**으로 사용

Train only 1 Epoch
- Validation Accuracy = 0.9810
- Validation Loss = 0.0484





DACON
DATA TO VALUE

3. Model 성능 비교 및 결과 분석

성능 비교

- Case 2 Model이 Case 1 Model에 비해 성능 향상
- Case 3 Model이 Case 2 Model에 비해 성능 향상

Case	Model	val loss	val accuracy
1	Single Sentence Classification (input = 뉴스 제목 + 내용)	0.0648	0.9761
2	Single Sentence Classification (input = 뉴스 내용)	0.0552	0.9806
3	**Sentence Pair Classification**	**0.0484**	**0.9810**

결과 분석

✓ **뉴스 제목은 광고성 문구에 해당되지 않아 모델이 진짜 뉴스라고 판단**

1. Case 1, Case 2 결과 분석

- Single Sentence Classifier(뉴스 제목 + 내용)은 내용의 진위 여부에 따라
  1) Title(진짜 뉴스) + 진짜 뉴스 Content -> 진짜 뉴스로 분류
  2) Title(진짜 뉴스) + 가짜 뉴스 Content -> 진짜 뉴스 + 가짜 뉴스가 섞여 Model의 판별 성능에 영향을 미침

  두 가지 경우로 판단

- 결과적으로 Single Classifier Model에서 뉴스 제목은 적합하지 않은 데이터로 분석됨

2. Case 2, Case 3 결과 분석

- Sentence Pair Classification Model은 두 문장 A, B가 주어질 때, B가 A의 뒤에 오는 것이 적합한지의 여부를 예측함
- A는 뉴스 Title, B는 뉴스 Content로 지정하며, 이 경우 내용의 진위 여부에 따라
  1) Title(진짜 뉴스) **IsNext?** 진짜 뉴스 Content
  2) Title(진짜 뉴스) **IsNext?** 가짜 뉴스 Content

  두 가지 경우로 판단

- 결과적으로 Case 3 Model은 뉴스 제목, 뉴스 내용 데이터를 사용하는 것이 적합하며, 다른 Model에 비해 좋은 성능을 보임

NH투자증권

실 험

DACON
DATA TO VALUE

4. Input 길이에 따른 성능 비교 실험

성능 비교

- 공통적으로 Input length가 길수록 수행시간이 늘어남
- 뉴스 Title이 사용되는 경우 Input length가 32이면 성능이 하락함
- Case 1 ~ 3의 경우 Input length가 길수록 성능이 향상됨
- Case 4 ~ 9의 경우 성능의 큰 차이가 없음
- Sentence Pair Classification Model이 다른 모델에 비해 성능이 대부분 우수함

Case	Model	Input length	val loss	val acc
1	Single Sentence Classification (input = 뉴스 제목 + 내용)	32	0.2142	0.9120
2		64	0.0756	0.9749
3		**128**	**0.0648**	**0.9761**
4	Single Sentence Classification (input = 뉴스 내용)	**32**	**0.0515**	**0.9800**
5		64	0.0631	0.9795
6		128	0.0552	0.9806
7	Sentence Pair Classification	32	0.0603	0.9789
8		**64**	**0.0422**	**0.9848**
9		128	0.0484	0.9810

결과 분석

- 형태소 단위 Tokenizer를 적용한 이후
  뉴스 Title, Content의 길이는 각각 0 ~ 40, 0 ~ 80자에 위치하고 있음
- Case 1 ~ 3의 경우 Input length가 짧을수록 Title 문장이 모두 사용되고,
  Content 문장에서 사용될 토큰이 줄어들 확률이 높아지며 예측에 어려움이 발생하며
  Input length가 길어질수록 Content 문장의 토큰이 많이 사용되며, 성능이 향상됨

Case	Model	Input length	Test Set 수행 시간	GPU
8	**Sentence Pair Classification**	**64**	**3분 8초**	**NVIDIA TITAN RTX**
9	Sentence Pair Classification	128	6분 2초	NVIDIA TITAN RTX

- Case 4 ~ 6의 경우 Content의 길이가 대부분 0 ~ 80자에 위치하고 있으며,
  Sentence의 앞부분만 확인해도 진짜 뉴스와 구분되는 광고성 문구의 특성 때문에 Input length에 관계없이 3가지 경우 모두 비슷한 성능을 보임
- Case 7의 경우 Case 1과 같이 Content 문장에서 사용될 토큰이 줄어들 확률이 높아져 성능 하락
- **Case 8 ~ 9의 경우 성능의 큰 차이는 없지만, Case 8의 성능과 수행 시간을 고려하여 최종 Model로 선택**

NH투자증권

대회 종료 이후 추가 실험

DACON
DATA TO VALUE

1. 성능 개선 실험

Weight decay?

- Regularization의 대표적인 기법이며 Generalization을 개선하여 Model의 일반화 성능을 높일 수 있음
- Weight들의 값이 증가하는 것을 제한함으로써, 모델의 복잡도를 감소시키는 기법으로 Overfitting을 방지하는 기법으로 사용됨
- Weight decay는 L2 regularization과 동일하며 L2 penalty로도 부름
- L2 규제는 loss function에 가중치에 대한 L2노름의 제곱을 더하여 사용

❖ L2 노름 : $||w||_2 = \sqrt{\sum_{i=1}^{n} |w_1|^2}$

❖ L2 규제 : $E(w) = E_0(w) + \frac{1}{2}\lambda \sum_{i=1}^{n} |w_i|^2$

loss function + L2 노름의 제곱

실험 조건 및 결과 분석

- 최종적으로 제출한 모델을 기준으로, weight decay를 제거 후 실험 (기존 Model은 $\lambda = 0.01$ 지정)
- 실험 결과 Weight decay를 적용하지 않은 Model이 성능이 우수함
- 광고성 문구는 진짜 뉴스와 같은 구어체가 아니기 때문에 Generalization 작업이 필요하지 않음
- L2 regularization 적용 시 오히려 성능이 하락하는 모습을 보임

Case	Model	Model 구분	Weight decay 적용 여부	val loss	val accuracy
1	KoBERT	최종 제출 Model	O ($\lambda = 0.01$)	0.0422	0.9848
2	**KoBERT**	**추가 실험 Model**	**X**	**0.0153**	**0.9945**

NH투자증권

대회 종료 이후 추가 실험

DACON
DATA TO VALUE

2. 수행 속도 개선 실험

KoELECTRA : BERT Model의 수행 속도와 계산 비용 문제를 개선한 ELECTRA Model 의 한국어 Model

- BERT Model을 바탕으로 GAN의 메커니즘을 적용하여 훈련시킨 Model
- 기존의 BERT 모델에 비해 빠른 수행 속도와 좋은 성능을 보임
- Fine-tuning 과정은 KoBERT Model과 동일하게 사용
- 경진대회 기간이 끝나 TestSet의 Score를 확인하지 못함

Model	Train FLOPs	Params	SQuAD 1.1 dev EM	SQuAD 1.1 dev F1	SQuAD 2.0 dev EM	SQuAD 2.0 dev F1	SQuAD 2.0 test EM	SQuAD 2.0 test F1
BERT-Base	6.4e19 (0.09x)	110M	80.8	88.5	–	–	–	–
BERT	1.9e20 (0.27x)	335M	84.1	90.9	79.0	81.8	80.0	83.0
SpanBERT	7.1e20 (1x)	335M	88.8	94.6	85.7	88.7	85.7	88.7
XLNet-Base	6.6e19 (0.09x)	117M	81.3	–	78.5	–	–	–
XLNet	3.9e21 (5.4x)	360M	**89.7**	**95.1**	87.9	**90.6**	87.9	90.7
RoBERTa-100K	6.4e20 (0.90x)	356M	–	94.0	–	87.7	–	–
RoBERTa-500K	3.2e21 (4.5x)	356M	88.9	94.6	86.5	89.4	86.8	89.8
ALBERT	3.1e22 (44x)	235M	89.3	94.8	87.4	90.2	88.1	90.9
BERT (ours)	7.1e20 (1x)	335M	88.0	93.7	84.7	87.5	–	–
ELECTRA-Base	6.4e19 (0.09x)	110M	84.5	90.8	80.5	83.3	–	–
ELECTRA-400K	7.1e20 (1x)	335M	88.7	94.2	86.9	89.6	–	–
ELECTRA-1.75M	3.1e21 (4.4x)	335M	**89.7**	94.9	**88.0**	**90.6**	**88.7**	**91.4**

➢ BERT Model과 ELECTRA Model의 Parameter와 성능을 비교한 표[2]

실험 결과

Case	Model	Model 구분	Weight decay 적용 여부	Input length	val loss	val accuracy	Test Set 수행 시간	GPU
1	KoBERT	최종 제출 Model	O (λ = 0.01)	64	0.0422	0.9848	3분 8초	NVIDIA TITAN RTX
2	KoBERT	추가 실험 Model	X	64	0.0153	0.9945	3분 8초	NVIDIA TITAN RTX
3	KoELECTRA	추가 실험 Model	O (λ = 0.01)	64	0.0486	0.9861	2분 38초	NVIDIA TITAN RTX
4	**KoELECTRA**	**추가 실험 Model**	**X**	**64**	**0.0122**	**0.9949**	**2분 38초**	**NVIDIA TITAN RTX**

NH투자증권

[2] K. Clark, M. Luong, Q. V. Le, and C. D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In ICLR.

결 론

DACON
DATA TO VALUE

1. 요약

- 제공된 데이터에서 뉴스 title, content, ord 컬럼을 사용
- 광고성 문구의 특성을 고려해 특수 문자와 조사를 제거 하지 않고, 형태소 단위 Tokenize를 적용
- KoBERT Model을 사용함으로써 기존의 BERT Model 보다 한국어 처리 Task에 좋은 성능을 보임
- Input length는 32일 때 대체적으로 낮은 성능을 보임
- Input length가 64와 128일 때 서로 비슷한 성능을 보이나, 수행시간을 고려하여 Input length를 64로 지정
- Classification Model에선 뉴스 Content만을 사용하는 방법이 Title + Content을 사용한 방법보다 성능이 향상됨
- Classification Model에 비해 Sentence Pair Classification Model의 성능이 대체적으로 우수
- 최종 Model은 Input length = 64인 Sentence Pair Classification Model을 사용
- TestSet에 대한 최종 Model의 Score는 0.9883의 성능을 보임
- 대회 종료 이후 weight decay를 사용하지 않는 추가 실험에서 성능이 월등히 향상됨을 확인
- KoBERT Model와 비교할 때 비슷한 성능을 보이지만, 수행 시간이 더 빠른 KoELECTRA Model을 사용 가능함을 확인함

NH투자증권

DACON
DATA TO VALUE

감사합니다 !

NH투자증권