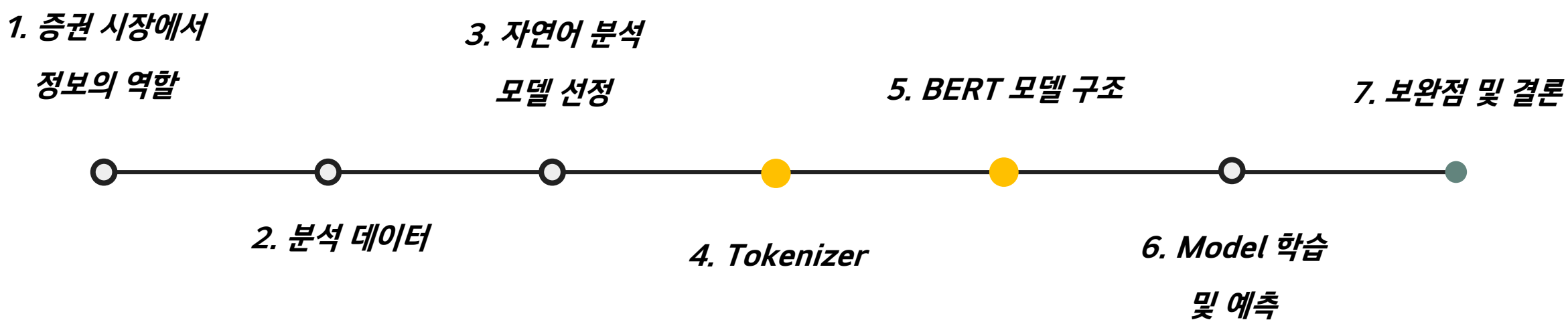
NH투자증권

"데이터, 문화가 되다 : League1"

AI야, 진짜 뉴스를 찾아줘

얼그레이
신승은 유채연

목차


1. 증권 시장에서 정보의 역할
2. 분석 데이터
3. 자연어 분석 모델 선정
4. Tokenizer
5. BERT 모델 구조
6. Model 학습 및 예측
7. 보완점 및 결론

1. 증권시장에서 정보의 역할

1) 신속하게 시장에 반영된다.

- 기업에 대한 이슈는 주가와 거래량에 즉시 반영됨.
  - ex. 애플카-기아차 협력 관련
- 해당 이슈의 사실 여부는 투자자 개인으로서는 바로 인식하기 어려움.

[단독]'애플카' 기아가 맡기로 가닥..美조지아공장 협력 거점

기아 중장기전략 플랜S, 애플카 사업에 적합 판단
조지아공장서 애플카 생산할 듯

등록 2021-01-19 오후 5:02:43
수정 2021-01-19 오후 9:14:51

가 가

<2021-01-19 기사>

	01/19	01/20	변동
주가(종가, ₩)	83,400	87,600	5.04% 증가
거래량(M)	13.37	26.33	96.85% 증가

<표 1>
<01/19~20 기아차(000270) 주가 및 거래량 변동 >

<2021-01-19~20 기아차(000270) 주가 및 거래량 그래프 >

기사출처: https://www.edaily.co.kr/news/read?newsId=04309926628919360&mediaCodeNo=257

1. 증권시장에서 정보의 역할

2) 진위여부가 투자자들의 성과에 크게 영향을 준다.

- 투자자가 알고 있는 정보가 허위라면, 투자자에게 손실을 줄 수 있음.
- 투자자 즉, 기업 외 사람들은 사내 정보를 전부 알 수 없어 공개되는 정보에 의지하여 투자함.
- 분식회계나 부실 회계 등 대중에게 공개되는 정보들이 허위일 때 사법 조치가 이루어지는 이유이기도 함.
- 외부로 공개된 정보들이 진실이라 판단한 투자자들이 해당 정보가 거짓으로 드러났을 때의 입는 타격은 클 수 밖에 없음.

투자자들의 이익 증대를 위해서는 빠르고 정확한 정보 수집이 필요하다.

주의할 점
신속성과 정확성은 서로 Trade-off 관계에 있다.
이 둘의 균형을 잘 이루는 것이 필요하다.

2. 분석 데이터

1) Train Data

- 2020.01.01 ~ 2020.06.30 까지의 사회, 경제 분야 뉴스 Data
- Columns: n_id, date, title, content, ord, info
- 독립변수로는 'title'과 'content'를, 종속 변수로는 'info' 사용

Data Dimension : (118745, 6)

n_id	date	title	content	ord	info
NEWS02580	20200605	[마감]코스닥 기관 678억 순매도	[이데일리 MARKETPOINT]15:32 현재 코스닥 기관 678억 순매도	1	0
NEWS02580	20200605	[마감]코스닥 기관 678억 순매도	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	2	1
NEWS02580	20200605	[마감]코스닥 기관 678억 순매도	하이스탁론, 선취수수료 없는 월 0.4% 최저금리 상품 출시	3	1
NEWS02580	20200605	[마감]코스닥 기관 678억 순매도	종합 경제정보 미디어 이데일리 - 무단전재 & 재배포 금지	4	0
NEWS09727	20200626	롯데·공영 등 7개 TV 홈쇼핑들, 동행세일 동참	전국적인 소비 붐 조성에 기여할 예정	1	0
NEWS09727	20200626	롯데·공영 등 7개 TV 홈쇼핑들, 동행세일 동참	[이데일리 권오석 기자] 중소벤처기업부(이하 중기부)는 대한민국 동행세일에 7개 TV홈쇼핑사가 홍보와 판매에 동참한다고 26일 밝혔다	2	0
NEWS09727	20200626	롯데·공영 등 7개 TV 홈쇼핑들, 동행세일 동참	대한민국 동행세일은 라이브 커머스, 언택트 콘서트, O2O 행사 연계 등 비대면이라는 새로운 형태의 소비촉진 행사다	3	0

<표 2>
Train data의 일부

2. 분석 데이터

2) Test Data

- 2020.01.01~2020.06.30 사회, 경제 분야 뉴스 Data
- Columns: n_id, date, title, content, ord, id

Data Dimension: (142565, 6)

n_id	date	title	content	ord	id
NEWS06731	20200115	`제2 이국종 막자`…정부, 외상센터 손익계산서 샅샅이 본다	이국종 센터장에 대한 아주대 의료원장 폭언 논란 확대- 2018년 아주대에 행정지도복지부 "관리감독 철저히"- 외상센터 손익 연구용역수익성 따지는 홀대 없도록- 늘어난 외상센터 덕에 외상 사망률 2년새 3120%로[이데일리 함정선 기자] 보건당국이 권역외상센터에 대한 관리감독을 강화하기로 했다	1	NEWS06731_1
NEWS06731	20200115	`제2 이국종 막자`…정부, 외상센터 손익계산서 샅샅이 본다	아울러 권역외상센터가 이를 운영하는 병원에 손해만 끼치고 있는지를 제대로 따지기 위해 손익계산 관련 연구용역도 진행한다	2	NEWS06731_2
NEWS06731	20200115	`제2 이국종 막자`…정부, 외상센터 손익계산서 샅샅이 본다	이국종 아주대병원 중증외상센터장이 유희석 아주대 의료원장으로부터 폭언을 당하는 녹음 파일이 공개되며 권역외상센터에 대한 점검이 필요하다는 목소리가 커지자 당국이 대응에 나선 것.15일 관가에 따르면 보건복지부는 이국종 센터장과 유희석 원장 간 갈등이 권역외상센터의 구조적인 문제로 보는 것을 경계하면서도 병원이 외상센터에 병상을 지원하지 않는 등 문제가 발생하는 것에 대해서는 면밀하게 살펴보겠다는 입장을 세웠다	3	NEWS06731_3
NEWS06731	20200115	`제2 이국종 막자`…정부, 외상센터 손익계산서 샅샅이 본다	아주대병원이 외상환자가 사용할 병실을 지원하지 않았고 닥터헬기에 대해서도 병원 수뇌부가 못마땅하게 여겼다는 의혹도 불거졌기 때문이다	4	NEWS06731_4

<표 3>
Test data의 일부

목표

Train Data의 뉴스 데이터들을 학습하여, Test Data의 뉴스 데이터들이 진실인지 거짓인지를 빠르고 정확하게 분류하는 것

3. 자연어 분석 모델 선정

1) 자연어 처리 (NLP) 방법

자연어 처리에는 대표적으로 다음과 같은 방법들이 사용됨.

- 단어의 등장 순서는 고려하지 않고, 빈도수 기반의 단어 표현 방법인 Bag of Words를 이용한 **TF-IDF**
- 단어를 밀집 벡터의 형태로 표현하는 방법을 워드 임베딩 기법을 이용한 **Word2Vec, GloVe**
- 순방향 언어 모델과 역방향 언어 모델을 따로 학습시킨 후에, 이렇게 사전 학습된 언어 모델로부터 임베딩 값을 얻는 **ELMo**

이 외에도 RNN을 이용한 모델 등 여러 방법들이 존재함. 다만 우리는 **BERT 모델**에 집중함.

2) BERT (Bidirectional Encoder Representations from Transformers)

- Google AI에서 개발한 Transformer 기반의 기계 학습 기술.
- 입력에서 단어의 15%를 숨기고 딥 양방향 Transformer encoder를 통해 전체 시퀀스를 실행한 다음 마스크 된 단어만 예측하는 접근법을 가짐.
- '큰 텍스트 코퍼스(Wikipedia)'를 이용하여 범용 목적의 '언어 이해'(language understanding)' 모델을 훈련시키는 것이 목적.
- 기존의 seq2seq의 구조인 인코더-디코더를 따르면서도, 논문의 이름처럼 어텐션(Attention)만으로 구현한 모델.
- 이 모델은 RNN을 사용하지 않고, 인코더-디코더 구조를 설계하였음에도 성능이 RNN보다 우수함.

3. 자연어 분석 모델 선정

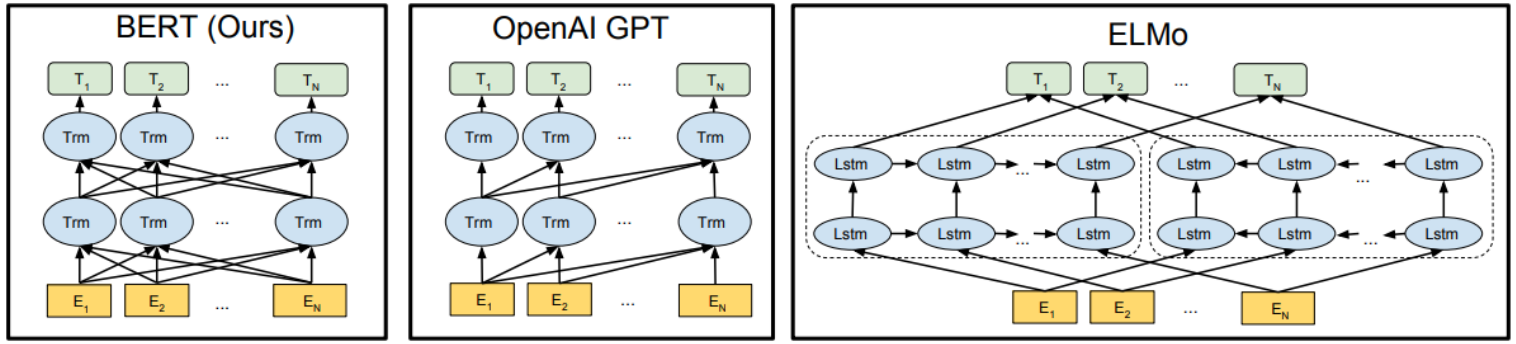
2) BERT (Bidirectional Encoder Representations from Transformers)

- 주요 특징

1. Bidirectional model

- BERT는 하나의 bidirectional model을 사용하여 좌우 context를 동시에 보는 구조

  (다른 모델: OpenAI GPT: left-to-right Transformer / ELMo: 독립적인 left-to-right, right-to-left 모델을 결합해서 사용)



2. 엄청난 크기의 모델과 사전 학습 데이터 (발표된 논문 기준)

- Parameter 수: Bert-base: 110M, Bert-large: 340M
- 사전 학습 데이터: BooksCorpus(800M 단어), English Wikipedia (2,500M 단어)

3. 자연어 분석 모델 선정

2) BERT (Bidirectional Encoder Representations from Transformers)

- 대표적인 모델 : BERT-Base, Multilingual Cased
  - 구성: 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
  - 한국어 포함되어 있음
  - 그러나, 한국어의 불규칙한 언어 변화의 특성을 잘 반영하지 못함.

BERT의 성능을 높이기 위해서는, 분석하려는 데이터에 맞는 리소스들이 사전 훈련에 사용 되어야 함.

방대한 양의 한국어 말뭉치를 사용하고 우리가 사용할 데이터와 같은 뉴스 데이터를 학습한
한국어 KoBERT 모델을 사용함.

데이터	문장 수	단어 수
한국 위키피디아	5M	54M
한국어 뉴스	20M	270M

<표 > KoBERT pre-train 시 사용한 문장 및 단어 수

3. 자연어 분석 모델 선정

3) KoBERT: Korean BERT pre-trained cased

- 특징
  - SKT Brain 팀이 개발, 한국어로 다루는 BERT 모델 중 가장 많이 알려져있음
  - 한글 위키피디아, 뉴스 텍스트 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모 말뭉치 학습

- 분석 모델로 선정한 이유
  - 공개된 데이터인 네이버 영화리뷰 데이터(NSMC)를 KoBERT 모델을 사용하여 감정 분석을 실행 하였을 때 다른 모델보다 우수한 정확도를 보여줌.
  - 이를 진짜, 가짜 뉴스 구별 분석, 즉 진위 여부로 바꿔 적용하여도 좋은 결과를 나타낼 것이라 예상함.

Model	Accuracy
Bert base multilingual cased	0.875
KoBERT	0.901
KoGPT2	0.899

<표 >
Naver Sentiment Movie Corpus를 이용하여 훈련한 결과

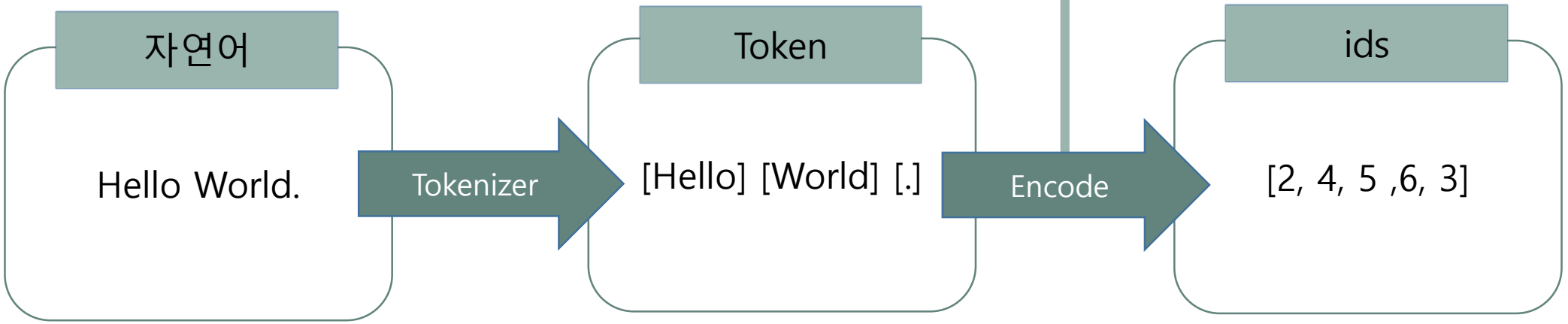
4. Tokenizer

1) Tokenizer의 필요성

- 모든 기계어의 알고리즘은 자연어를 바로 처리할 수 없고 숫자만 이해함.
- 주어진 문장을 특정 단위로 쪼갠 뒤 Vocab의 Value으로 넘버링 함.
- Vocab에 없는 단어(OOV)의 경우 [UNK]로, 문장들을 구분하기 위해 맨 앞과 맨 뒤는 각각 [CLS], [SEP] 토큰으로 나타냄.
- 모든 단어를 음절로 나눈다면 OOV의 비율이 낮아지겠지만 단어의 뜻을 온전히 전달하기 힘듦.
- 밑의 도표는 이러한 과정을 대략적으로 요약하여 보여줌.

<vocab>

Tokens	ids
[CLS]	2
[SEP]	3
Hello	4
World	5
.	6



4. Tokenizer

2) 한국어 Tokenizer 특징

한국어는 영어와는 달리 띄어쓰기만으로는 토큰화를 하기엔 어려움이 있음.

한국어가 영어와는 다른 형태를 가지는 언어인 교착어이기 때문임.

예시) 문장 : 아버지가 책을 읽었다

자립 형태소 : 아버지, 책 / 의존 형태소 : -가, -을, 읽-, -었, -다

한국어에서 영어에서의 단어 토큰화와 유사한 형태를 얻으려면 형태소 토큰화를 수행해야함.

Vocab size 큰 경우, 단어 단위로, **Vocab size 작은 경우, 음절 단위**로 tokenize되는 경향이 있음.

☞ **분석에 필요한 적절한 단어들을 가진 Vocab을 이용한 Tokenizer를 이용해야 함.**

<주요 한국어 Tokenizer 의 예시와 Vocab 단어 수>

	Mulitlingual BERT	KorBERT	KoBERT	KR-BERT
냉장고	냉#장#고	냉#장#고	냉#장#고	냉#장#고
츕다	[UNK]	츕#다	츕#다	츕#다
뱃사람	[UNK]	뱃#사람	뱃#사람	뱃#사람
Vocab 단어 수	**119,547**	**30,797**	**8,002**	**16,424**

4. Tokenizer

3) Tokenizer Vocab 추가

- 단어 수가 많은 모델(KRBERT(snunlp) : 16,424개)의 tokenizer로 나눈 단어들의 상위 500개를 추출하고, 그 중 중복을 제외한 약 70개의 단어를 추가함으로써 우리가 사용한 KOBERT Vocab의 부족함을 보완하려고 노력함.
- 주어진 content의 내용이 최신 뉴스의 정치, 경제, 사회면에 대한 것이므로 최대한 이와 같은 도메인의 단어들이 추가한다면 앞에서 설명한 사전 외 단어(OOV)를 줄일 수 있을 것이라고 생각함.

<추가된 단어 목록>

추천 주를 **무료** 최저 대장 것으로 가능 **상승** 당장 위해 밝혔다 연결 다시 만들어 핵심 **신용** 긴급 모집 통해 최종 내고 준다 **소형** 역대 받으 낸다 영웅 이용하여 도 **급등** 전 여기 역사를 최대 **금지** 말했다 재배 위한 추가 상반기 **확산** 바로 이어 이날 **예상 감염** 등을 이후 **테마** 예정 따르면 따라 **마스크** 대해 제공 확대 **바이러스** 경우 **개인** 있어 이용 관계자 가장 **공략** 해당 집중 오전 출시 **해외** 모두 있다고 주요 **AI** ....

→ 최근 언급되는 코로나 관련 단어들과 주식 관련 단어들이 더해졌음을 알 수 있음.

4. Tokenizer

3) Tokenizer Vocab 추가

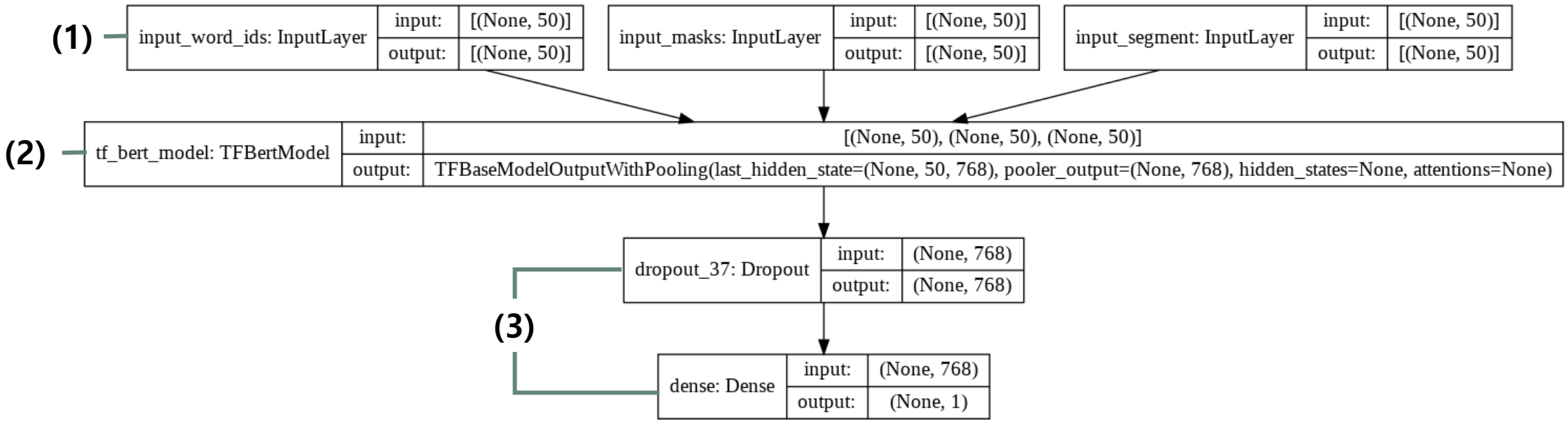
<단어 추가 전/ 후 tokenizer 변화 예시>

원본	추가 전	추가 후
똑똑해진 소비자..한국도 이젠 **소형**차 시대	['_', '똑', '똑', '해진', '소비자', '한국', '도', '이', '젠', **'소', '형'**, '차', '시대']	['_', '똑', '똑', '해진', '소비자', '한국', '도', '이', '젠', **'소형'**, '_차', '시대']
2008년 금융위기 이후 가장 큰 **상승**세다	['_2008', '년', '금융', '위기', '이', '후', '가', '장', '큰', **'상', '승'**, '세', '다']	['_2008', '년', '금융', '위기', '이후', '가장', '_큰', **'상승'**, '_세', '다']
[이데일리 박태진 기자] 신종 코로나 **바이러스** 감염증(코로나19) 팬데믹(세계적 대유행)에 국내 제약·바이오주가 **급등**세를 보이고 있다	['_이데일리', '박', '태', '진', '기자', '신', '종', '코', '로', '나', **'바이', '러스'**, '감', '염', '증', '코', '로', '나', '19', '팬', '데', '믹', '세계', '적', '대', '유', '행', '에', '국내', '제약', '바이오', '주가', **'급', '등'**, '세를', '보이', '고', '있다']	['_이데일리', '박', '태', '진', '기자', '신', '종', '코', '로', '나', **'바이러스'**, '감염', '_증', '코', '로', '나', '19', '팬', '데', '믹', '세계', '적', '대', '유', '행', '에', '국내', '제약', '바이오', '주가', **'급등'**, '_세', '를', '보이', '고', '있다']

→ 전보다 문장에서 단어가 의미하는 바를 포함하도록 나눠진 것을 확인할 수 있음.

5. BERT 모델 구조

1) 학습 모델 전체 구조



- 모델 구성

(1) 앞서 생성한 3가지의 Input data(Token ids, segment ids, mask ids)

(2) KoBERT: TF(TensorFlow)BertModel 이용하여 구현

- 구성: Embedding 층, Encoder 층, Classifier 층

(3) Dropout(p=0.5) 및 Dense layer 추가(활성화 함수: sigmoid)

5. BERT 모델 구조

(1) Input Data

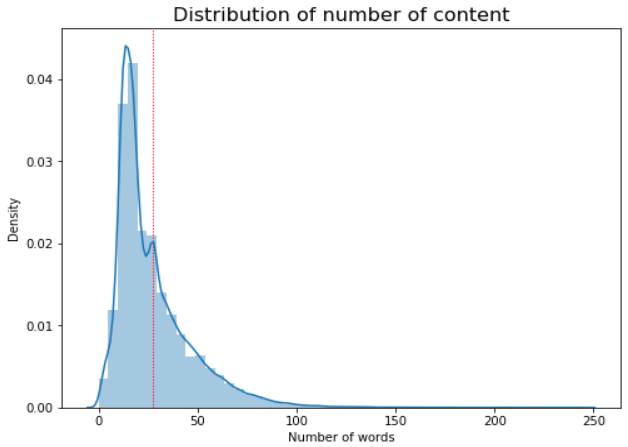
- 본문 내용으로만 진실과 거짓을 판단함.
- 제목과 본문 내용을 같이 넣는다면 이 둘의 내용의 괴리감이 진위여부를 판단하는데 더 도움을 줄 것이라고 생각함.
- 제목은 같지만, 본문에 따라서 진짜 뉴스와 가짜 뉴스가 분류됨을 밑의 표에서 확인 할 수 있음.
- 실제로 우리가 결과를 제출했을 때, Title+Content가 Content로만 학습시킨 모델보다 accuracy가 더 향상되었음.

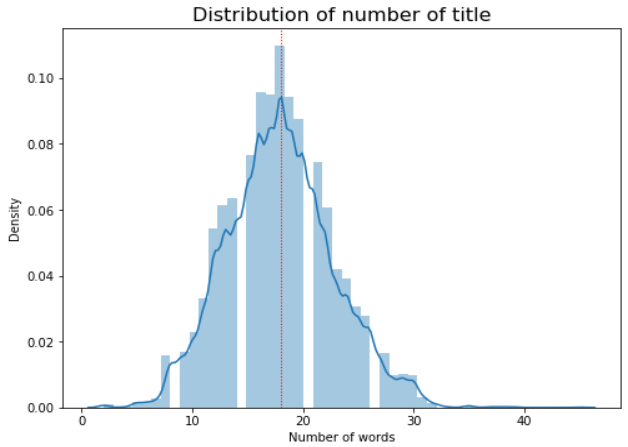
Title	Content	info
[마감]코스닥 기관 678억 순매도	[이데일리 MARKETPOINT]15:32 현재 코스닥 기관 678억 순매도	0
[마감]코스닥 기관 678억 순매도	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1
중견련, '21대 국회'에 "중견기업 육성정책 개선" 촉구	이어 특히 제21대 국회는 정쟁의 국회가 아닌 경제를 살리는 국회, 국민을 위한 국회가 돼야 할 것이라며 여야 구분 없이 위기 극복과 대한민국 재도약을 위한 경제 활성화 입법에 매진해 주기를 간절히 희망한다고 덧붙였다	0
중견련, '21대 국회'에 "중견기업 육성정책 개선" 촉구	"실적기반" 저가에 매집해야 할 8월 급등유망주 TOP 5 전격공개	1

5. BERT 모델 구조

(1) Input Data

- 기본 전처리
  1) 토큰화 하는데 의미 없는 기호들이 영향을 주지 않게 하기 위하여 특수문자를 모두 제거함
  2) 한국어는 띄어쓰기가 문장과 단어 의미에 큰 영향을 주지 않기 때문에 모두 줄여 줌.
- 길이 결정





- Tokenizer로 나눈 Title의 길이의 평균은 17.9개, Content의 길이의 평균은 27.4개임
- 길이를 Title 20개, Content 30개로 두 개를 이어 총 input 길이를 모두 **50**으로 설정함.

5. BERT 모델 구조

(1) Input Data

- 3개의 Input Data가 필요함

• **Token ids** : 토큰화 이후 Encode한 행렬, 주어진 길이보다 짧다면 Truncation하고, 길다면 Padding 과정을 거침.

• **Segment ids** : 두 문장을 이용했다면, 앞 문장과 뒷문장을 구분함.

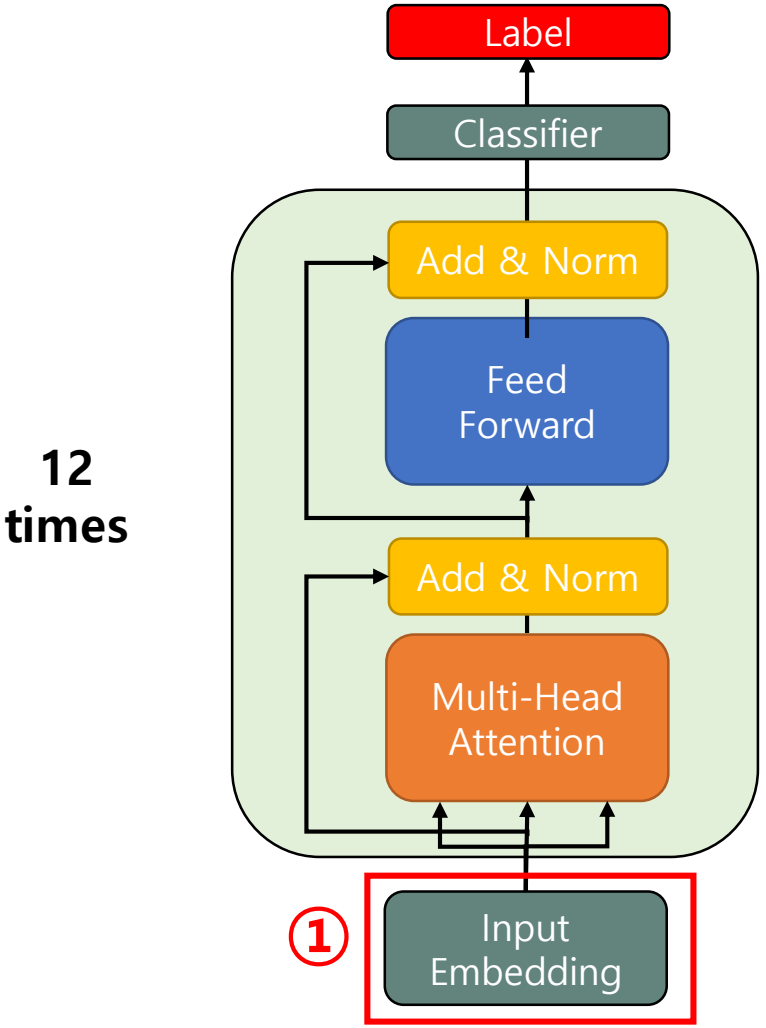
• **Mask ids** : 실제 위치에 Token 값이 Padding 값이 아닌, 실제 값이 있는 지를 나타냄.

<Input Data 예시>

원본	[2020 금융투자대상]`마블로 통한다`..KB증권 MTS로 한큐에 / 업계 최초로 마블에 카카오페이 인증 서비스를 도입한 것도 눈에 띈다
전처리	2020금융투자대상마블로통한다KB증권MTS로한큐에 / 업계최초로마블에카카오페이인증서비스를도입한것도눈에띈다
Token Ids	[2 554 127 5554 7645 5816 6141 6438 6079 7636 7831 308 7317 316 353 6079 7828 7563 6896 3 2 3261 7458 7446 6079 6141 6438 6896 7495 7495 6964 7712 7096 7119 7316 6555 6116 5859 7138 7828 5398 5859 5745 6896 5999 5782 3 1 1 1]
Mask Ids	[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0]
Segment Ids	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

5. BERT 모델 구조

(2) KoBERT 구조

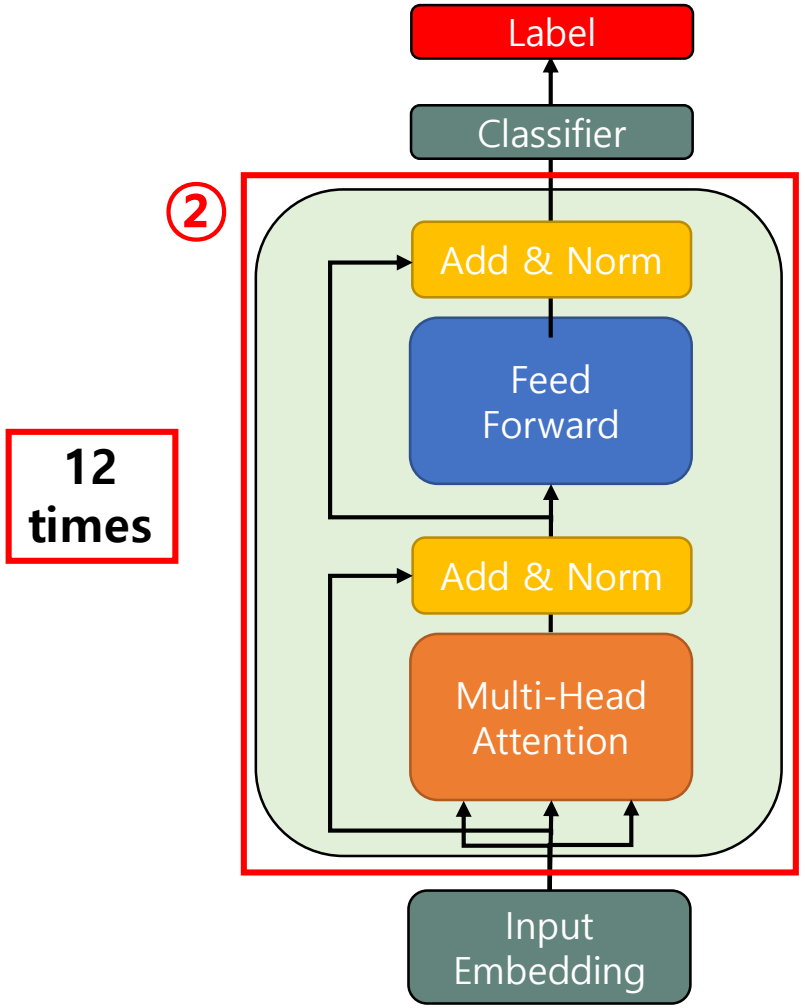


① Embedding 층

- 앞서 만든 세 가지 Input (token ids, segment ids, mask ids)을 합산하여 사용함

5. BERT 모델 구조

(2) KoBERT 구조

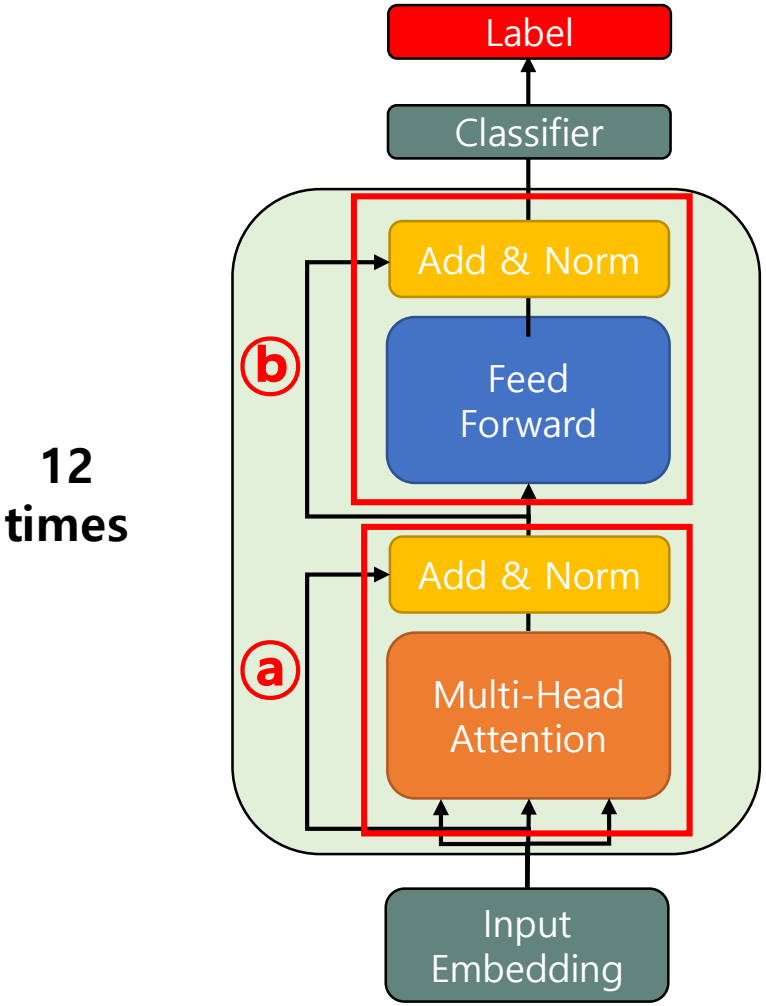


② Encoder 층

- 본격적으로 학습하는 층
- KoBERT는 Bert-base 모델이기 때문에 총 12개의 Encoder 블럭을 가짐. 따라서 한번 학습할 때마다 12번의 Encoder가 학습됨.
- 크게 Attention, Feed Foward, Output 층으로 구성
- 이전 출력 값을 현재의 입력 값으로 하는 RNN과 유사한 특징을 지님

5. BERT 모델 구조

(2) KoBERT 구조



② Encoder 층

a. Attention

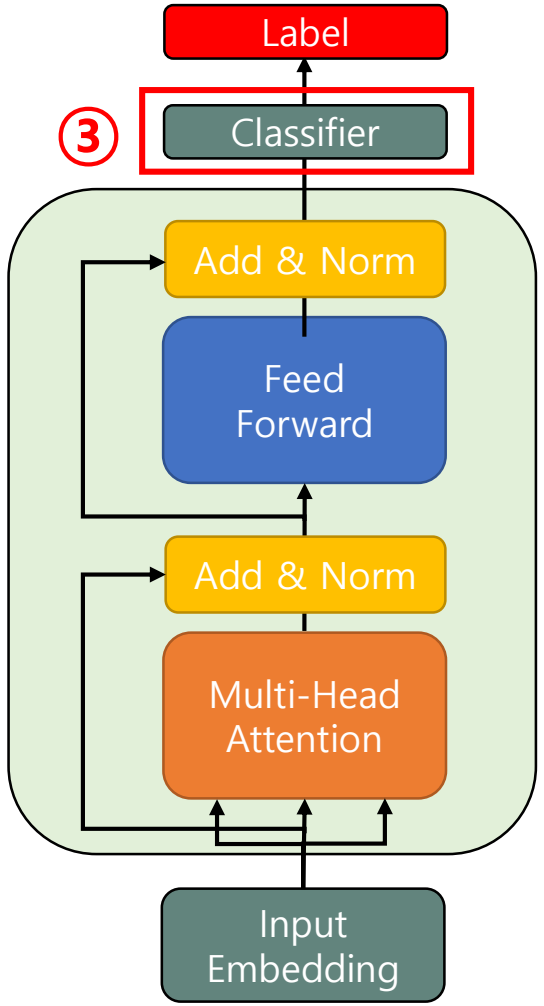
- 입력된 여러 hidden state에 대해서 중점적으로 집중(attention) 해서 봐야 할 벡터를 소프트맥스(softmax) 로 점수를 매긴 후에, 각각을 hidden state에 곱해줌.
- 이런 방식으로 학습을 반복
- 추론 시 padding된 토큰에 대해서는 마스킹 처리를 하고, 이 토큰에 패널티를 부과해 어텐션 점수를 받지 못하도록 구현 ➡ 중요한 값에 "집중"

b. Feed Forward

- 앞서 생성한 어텐션 값을 통과 시키는 층
- 두 개의 Linear Transformations로 구성
- 활성화 함수로는 'tanh' 사용

5. BERT 모델 구조

(2) KoBERT 구조

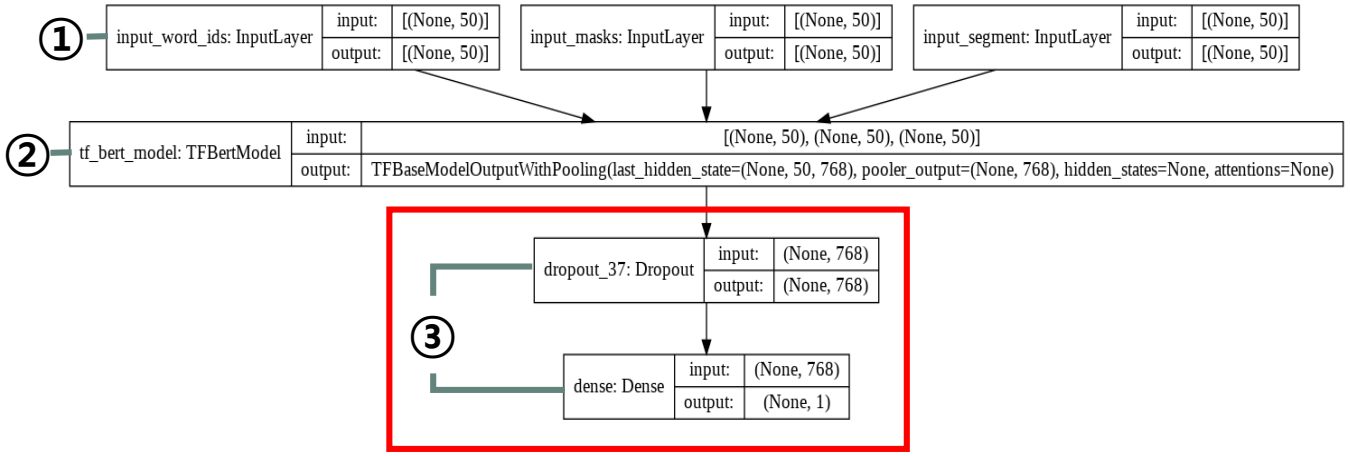


③ Classifier 층

- 마지막으로 결과를 도출해주는 층
- Linear Transformation으로 구성, 768개의 input을 받아 sequence output, pooled output을 내보낸다.

5. BERT 모델 구조

(3) Layer 추가



- 앞에서 받은 두 개의 output 중 pooled output을 이용함.
- KoBERT 구조 외부에 모델 성능의 향상을 위해 Layer 추가
- Dropout(p=0.5)
- Dense
  - 활성화 함수는 'sigmoid'
  - kernel_initializer: 표준편차가 0.02인 Truncated Normal 분포 생성(미세 조정)

6. Model 학습 및 예측

1) 모델 summary 및 학습 정보

- 학습 시킨 parameter 수: 92,244,481개
- Batch size: 32, 1 epoch 당 훈련 steps: 3340
- 설정한 epochs는 30이나, 'Earlystopping' 옵션 사용하여 최종적으로 13번 학습시킴
- 학습 시간: 1 epoch 당 약 20분
- Optimizer: "RectifiedAdam" 사용
  - 옵션: lr=5.0e-5, total_steps = 1000000, warmup_proportion=0.1, min_lr=1e-5, epsilon=1e-8, clipnorm=1.0

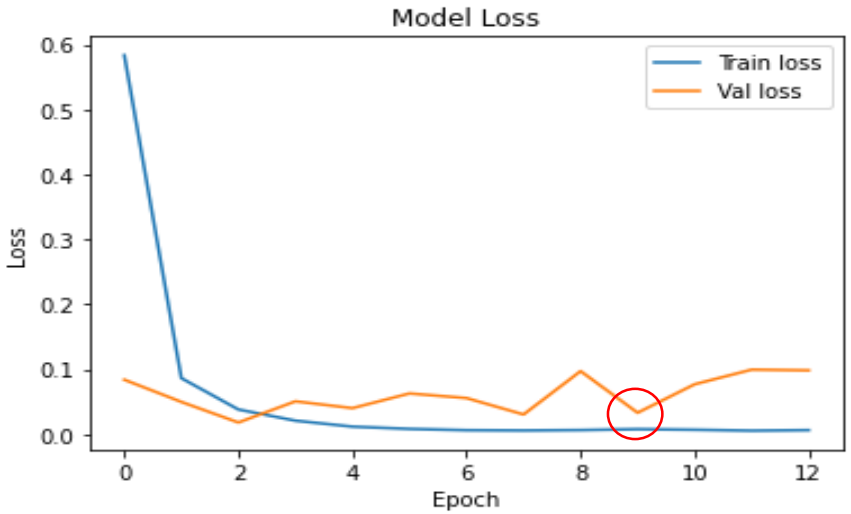
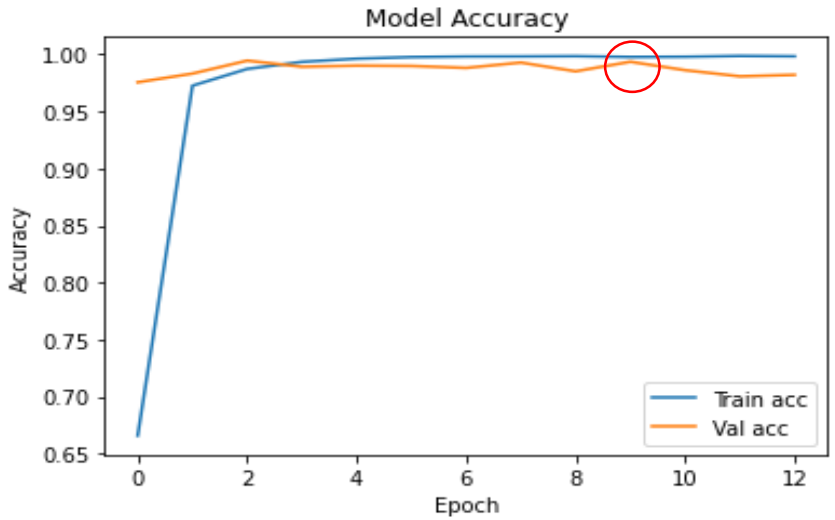
```
Model: "model"
_________________________________________________________________________________________________
Layer (type)                 Output Shape              Param #     Connected to
=================================================================================================
input_word_ids (InputLayer)  [(None, 50)]              0
_________________________________________________________________________________________________
input_masks (InputLayer)     [(None, 50)]              0
_________________________________________________________________________________________________
input_segment (InputLayer)   [(None, 50)]              0
_________________________________________________________________________________________________
tf_bert_model (TFBertModel)  TFBaseModelOutputWit 92243712    input_word_ids[0][0]
                                                              input_masks[0][0]
                                                              input_segment[0][0]
_________________________________________________________________________________________________
dropout_37 (Dropout)         (None, 768)               0           tf_bert_model[0][1]
_________________________________________________________________________________________________
dense (Dense)                (None, 1)                 769         dropout_37[0][0]
=================================================================================================
Total params: 92,244,481
Trainable params: 92,244,481
Non-trainable params: 0
_________________________________________________________________________________________________
```

<Model Summary>

6. Model 학습 및 예측

2) 학습 결과

- 학습을 시킬수록 accuracy가 증가하고, loss가 감소하는 양상을 보임
- 하지만 학습 과정에서 결과가 오락가락하여, validation loss와 accuracy가 가장 좋았던 10번째 학습 모델의 가중치를 사용하여 예측하고자 함 (validation loss = 0.0329, validation_accuracy: 0.9936)
- 예측 시 소요된 시간이 약 8분으로, 방대한 크기의 데이터와 모델을 다루는데 비해 효율적임.
- 최종 Test Accuracy : **0.98214**

학습 시 Loss의 추이	학습 시 Accuracy 추이
	

7. 보완점 및 결론

1) 보완점 - Tokenizer 측면

- 사전 학습된 단어들에 의해 결과가 많은 영향을 받음
- 따라서, 증시에 관련된 주제 혹은 기사 등으로 추가적인 학습을 필요함
- 특히 '코로나 19 바이러스'나 증시 상황에 영향을 줄 수 있는 새로운 주제, 단어들이 등장할 때마다 사전 학습 모델을 업데이트 시킨다면 더 좋은 결과를 얻을 수 있을 것이라 기대함

2) 보완점 - Model Training 측면

- GPU 환경을 사용해야 빠른 학습이 가능한데, Google Colab에서는 사용량에 한계가 있어 긴 시간 동안의 지속적인 학습이 어려웠음
- Epoch이 증가할수록 과적합이 일어나지 않고 학습이 잘 되었으므로, Epoch 수를 더 할당하여 학습을 시켰다면 더 좋은 결과를 얻을 수 있을 것이라 기대함

7. 보완점 및 결론

3) 결론

- ✓ 증권 시장에서 정보는 신속하게 시장에 반영되며, 그 진위여부가 투자자들의 성과에 크게 영향을 주므로 빠르고 정확한 정보 수집이 필요함. 따라서, 진짜 뉴스와 가짜 뉴스를 구분하는 알고리즘이 필요함.
- ✓ 모든 기계어 알고리즘은 자연어를 인식하지 못하므로, 일련의 규칙에 따라 문장을 나누고 각각 나누어진 부분에 숫자를 부여하는 Tokenizer가 필요함. 이때, 분석 데이터에 적절한 단어들을 가진 Vocab을 이용한 Tokenizer를 적용해야함.
- ✓ 여러 자연어 학습 모델 중, 하나의 bidirectional model을 사용하여 좌우 문맥를 동시에 보는 구조를 가진 BERT 모델을 사용하였고, 그 중에서도 한국어 문장을 다루는데 좋은 성능을 보인 KoBERT를 학습에 사용함.
- ✓ 학습 결과, Test accuracy는 0.982를 얻었고, 예측 시간은 약 8분이 소요 되어 빠르고 정확한 알고리즘을 잘 구축했다고 판단됨.

감사합니다