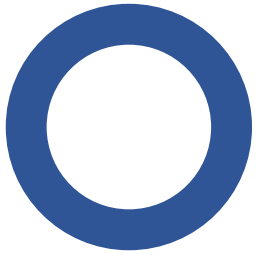
제주ICT 이노베이션 스퀘어
AI 데이터 분석 심화 코스

Project Name : 쇼핑몰 리뷰 평점 분류

TEAM : 강승웅, 김창명, 민병국, 정선우

CONTENTS

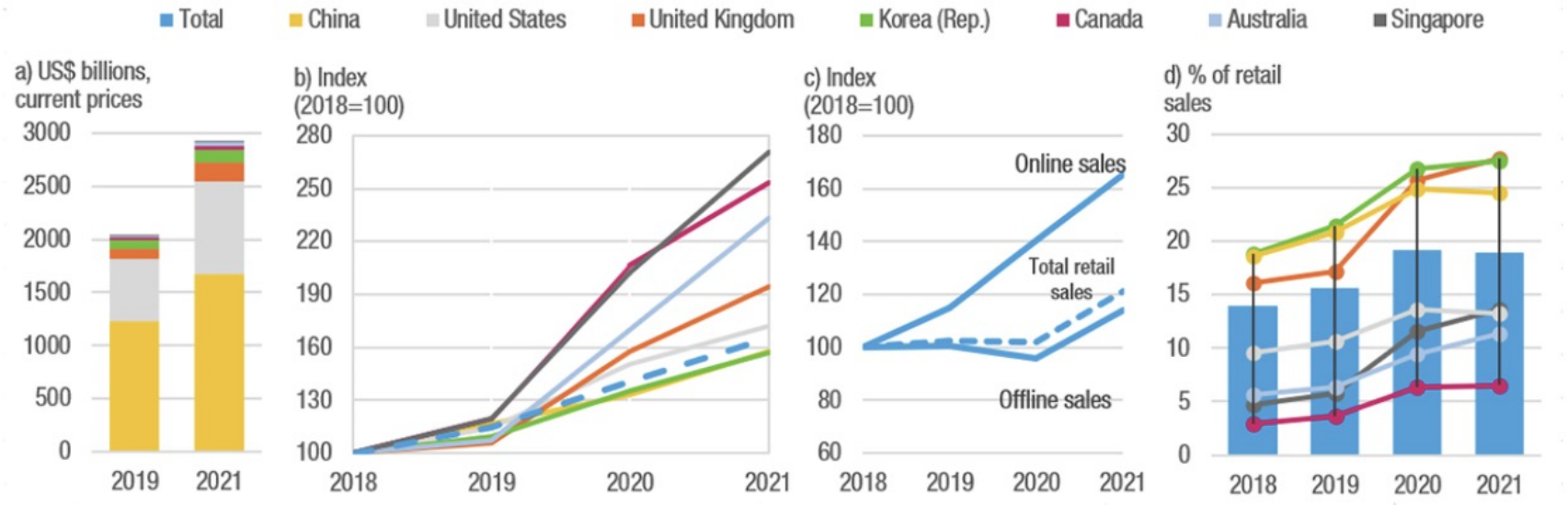


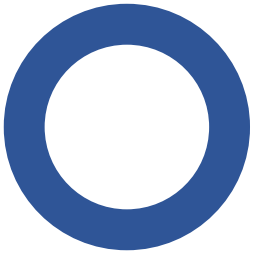
Background

Background
주제선정

Covid-19 이후 e-commerce 시장의 성장에 따라 리뷰 데이터의 중요성도 높아지고 있습니다.
이에 따라 우리 팀에서는 리뷰데이터와 평점을 분석하여 예측하는 주제를 선정하였습니다.

Figure 2. Online retail sales, seven countries, 2018-2021
Value (US$ billions, current prices), Indices (2018=100) and percentage of retail sales

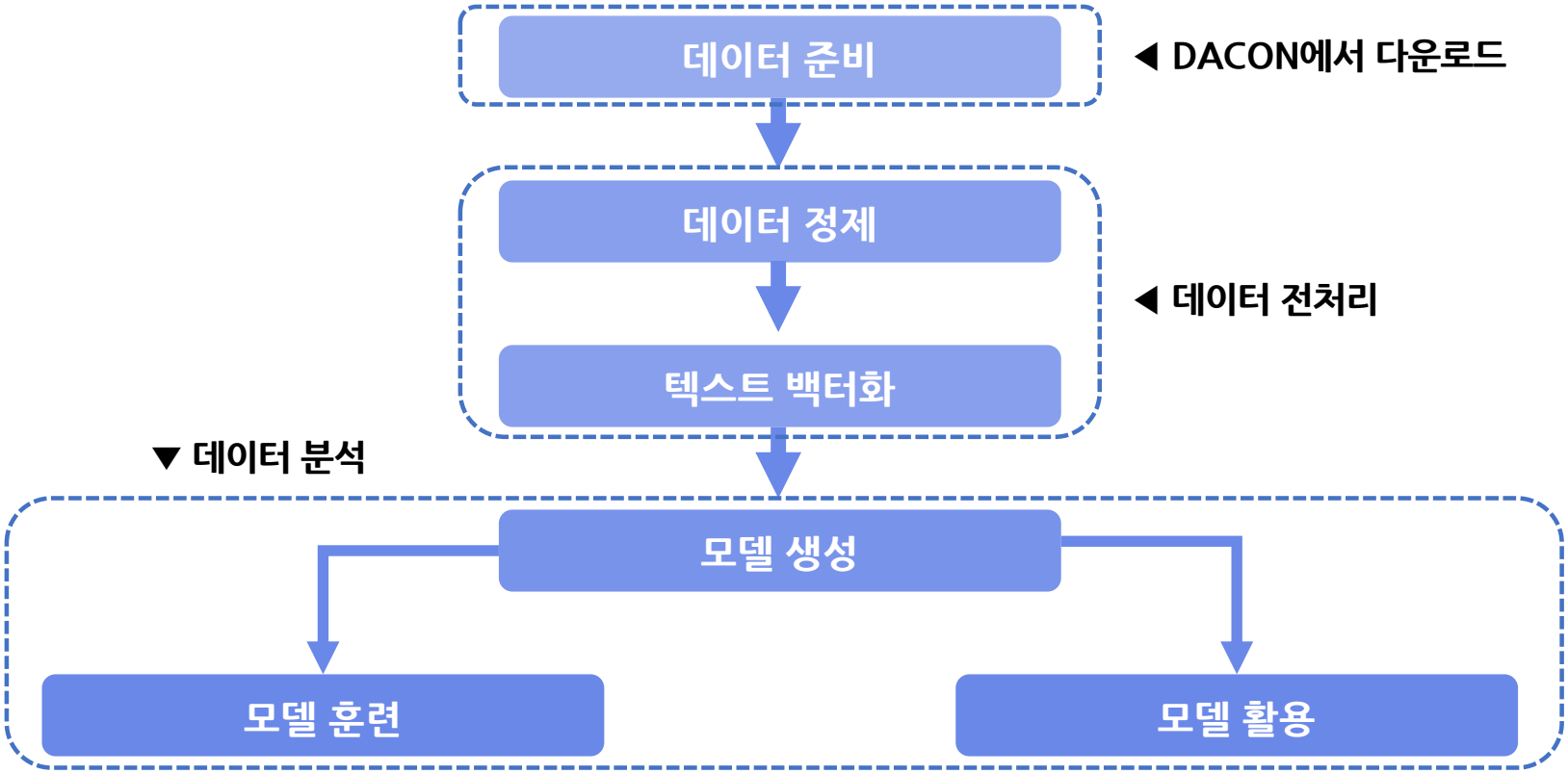


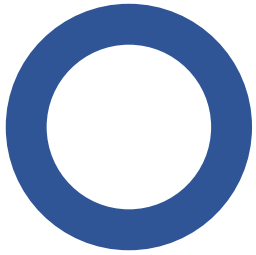


Analysis Step

Analysis Step

분석 과정





Progress

Progress

- 데이터 정제

BEFORE

왠만해선 리뷰 안쓰는데요, 너무 까실까실해서 살이 찔려요. 그리고 군데군데 대나무가 갈라져서 뜯겨 지네요. 반품포장 힘들어 그냥 한해만 대충 쓸까봐요 에휴~~

AFTER

웬만해선 리뷰 안 쓰는데요 너무 까슬까슬해서 살이 찔려요 그리고 군데군데 대나무가 갈라져서 뜯겨지네요 반품 포장 힘들어 그냥 한 해만 대충 쓸까 봐요 어휴

1. 텍스트에 포함되어 있는 특수문자 제거
2. 자음, 모음만 쓴 것 제거
3. 500자 이상 리뷰 제거
   - hanspell 사용 시 500자 이상맞춤법 교정 불가
4. 맞춤법 교정 - 네이버 기반 hanspell 사용

Progress

- 데이터 토큰화

사용해본 토큰화 라이브러리

1) MeCab - 리뷰 특성상 다른 텍스트 문서에 비해 길이가 짧은 편인데, MeCab은 단어의 원형을 추출해주기 때문에 같은 뜻을 가졌지만 형태가 다른 단어들을 하나로 count하여 데이터가 분산되는 것을 방지할 수 있었다. 하지만 토큰화의 정확도가 떨어짐

2) Kobert 토큰화 라이브러리를 사용함

KoBERT(Korean Bidirectional Encoder Representation from Transformer) 란?

BERT는 약 33억 개의 단어로 Transformer의 인코더만을 Bidirectional하게 사용도록 pretrain 되어 있는 기계번역 모델이다.
대규모 데이터를 MLM 방식으로 학습시킨 pretrained model로, 다양한 문제 해결이 가능하며 특히 classfication에 강력하다고 알려져 있다.
따라서 좋은 알고리즘을 갖고 있는 BERT 모델을 한국어에도 잘 활용할 수 있도록 만들어진 것이 KoBERT이다.
KoBERT는 SKTBrain에서 공개한 기계번역 모델인데, BERT를 기반으로 하는 대화엔진 개발을 위해 만들어졌다.
한국어 위키에서 5백만개의 문장과 54백만개의 단어를 학습시킨 모델이다.

Progress

- 모델생성 & 모델훈련

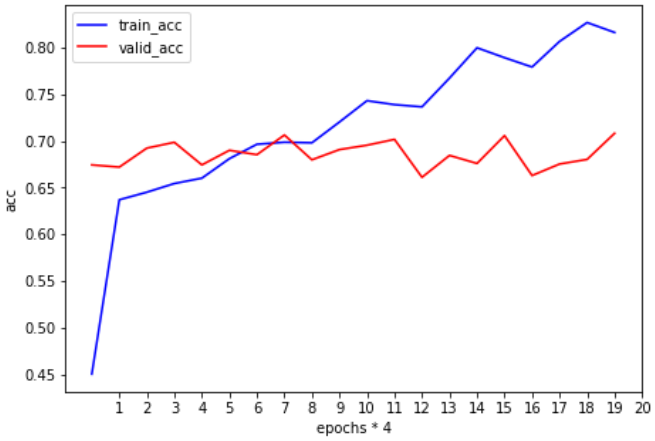
모델생성

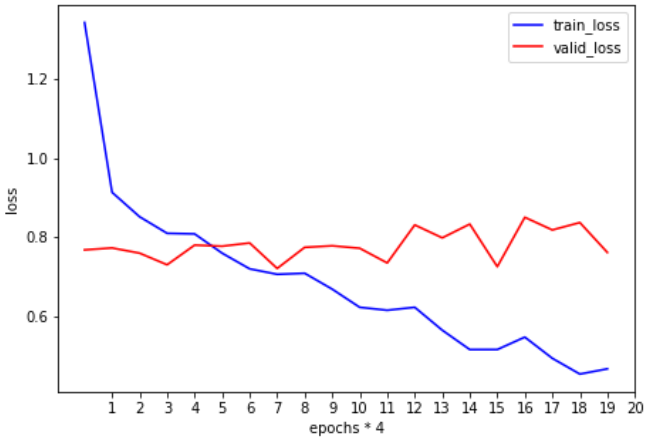
- X(독립변수): 리뷰 데이터
- Y(종속변수): 평점
- 사용한 Model : BERT를 project 목적에 따라 output layer를 추가 및 재조정(multi-classification)

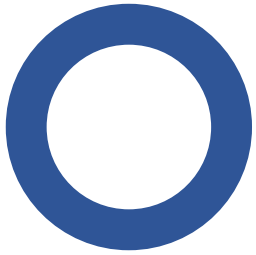
※ KoBert 깃허브에 있는 '네이버 영화평 이중분류 예시 코드 ' 바탕으로 작성

모델훈련

- epoch : 5
- train accuracy : 0.807
- test accuracy : 0.682







Application

Application
모델활용

모델활용

1) Test 데이터 불러오기(리뷰 데이터만 있고 평점 column없음)

	id	reviews
0	0	채소가 약간 시들어 있어요
1	1	발톱 두껍고 단단한 분들 써도 소용없어요 이 테이프 물렁거리고 힘이없어서 들어 올리...
2	2	부들부들 좋네요 입어보고 시원하면 또 살게요
3	3	이런 1. 8 골드 주라니깐 파란개 오네 회사전화걸어도 받지도 않고 머하자는거임?
4	4	검수도 없이 보내구 불량 배송비 5000원 청구하네요 완전별로 별하나도 아까워요
...	...	...
24995	24995	사용해보니 좋아요~^^
24996	24996	저렴한가격에. 질좋고. 핏좋고. 너무. 이쁘게. 입고다녀요..
24997	24997	세트상품이라고 써있어서 그런줄 알고 구매했더니 단품이었네요 낚인 느낌도 들고 그러네...
24998	24998	역시 로네펠트!! 좋아요.
24999	24999	데싱 디바 써보고 갠찮아서 비슷 한줄 알았더니 완전 별로예요——3000언 더주고 디...

25000 rows × 2 columns

◀ test.csv

Application
모델활용

모델활용

2) Train data와 동일한 전처리 및 벡터화 적용

3) 평점 예측값 생성 및 저장

4) 데이콘에 제출하여 정확도 확인 ▷ Baseline에 비해 확연히 좋은 결과 확인

Baseline
0.60712

WINNER 1% 4% 10%

전체 랭킹 ›

#	팀	팀 멤버	점수	제출수	등록일
28	SunooJ	Su	0.66384	2	3시간 전
1	롤케잌	롤케	0.71008	35	8분 전
2	yasuo		0.70912	39	2시간 전
3	minyeamer	mi	0.70736	29	2시간 전
4	나요한		0.7056	13	하루 전
5	709		0.70472	6	3일 전

Application
모델활용

모델활용(+추가작업)

hanspell 전처리 데이터

+ ELECTRA 모델 적용

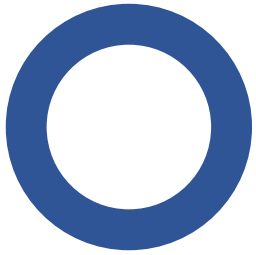
학습 측정치

- epochs=5, max_len=64, hanspell 전처리 데이터 ==> **제출 acc 0.6685**

Step	Training Loss	Validation Loss	Acc	F1	Precision	Recall
200	0.874800	0.755506	0.675670	0.635766	0.608009	0.675670
400	0.714600	0.741090	0.696879	0.659495	0.651146	0.696879
600	0.674700	0.730055	0.698679	0.650947	0.660954	0.698679
800	0.587200	0.785325	0.684874	0.670020	0.661441	0.684874
1000	0.561900	0.824220	0.679872	0.667984	0.661331	0.679872
1200	0.484300	0.845129	0.681673	0.669266	0.662164	0.681673
1400	0.418500	0.915961	0.669468	0.662132	0.656407	0.669468

- epochs=5 ==> **제출 acc 0.6591**

Step	Training Loss	Validation Loss	Acc	F1	Precision	Recall
200	0.172100	1.634539	0.617247	0.618473	0.625467	0.617247
400	0.159200	1.654786	0.650860	0.643516	0.637916	0.650860
600	0.125500	1.630619	0.631853	0.627805	0.624225	0.631853
800	0.086000	1.870700	0.641657	0.635299	0.631096	0.641657
1000	0.074500	2.012760	0.642857	0.637065	0.633939	0.642857
1200	0.055200	2.078029	0.650660	0.640178	0.637454	0.650660
1400	0.039500	2.198593	0.644058	0.640254	0.638873	0.644058



Conclusion

Conclusion
결론

- 리뷰 데이터 특성상 문법 오류, 신조어 등의 문제로 데이터 정제를 하고 학습을 해야 할지 맞춤법 교정을 하지 않고 학습을 해야 할지에 대한 고민을 많이 했지만, 학습의 정확도 차원에서는 데이터 정제 작업이 큰 도움이 되었고 실행 속도 차원에서는 정제를 하지 않는 것이 훨씬 나았다.
- 맞춤법 교정을 하지 않은 모델의 데이콘 점수 0.65 -> 맞춤법 교정을 한 모델의 데이콘 점수 0.66
- 실행 속도가 약 2시간이 차이가 나지만 점수는 0.01 올렸다.
- 단순히 리뷰 내용만이 아닌 문장의 길이 등 다른 변수를 입력하면 더 좋은 결과가 나올 것 같다.