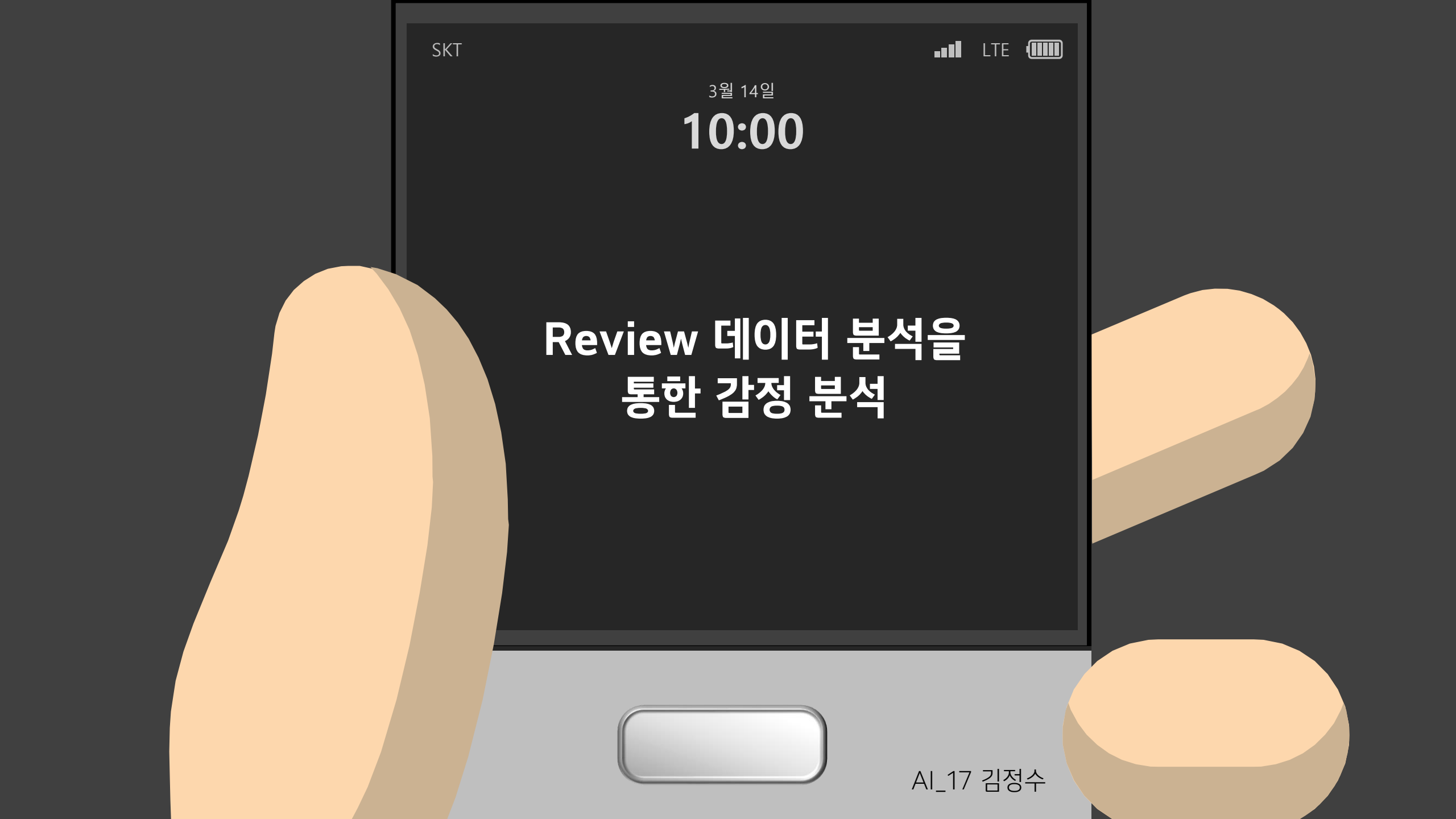

SKT
LTE
3월 14일
10:00
Review 데이터 분석을
통한 감정 분석
AI_17 김정수

CONTENTS

프로젝트 & 데이터 소개

가설 설정

데이터 전처리

딥러닝 모델

모델 평가

한계점, 추후 해결 방안

1. 프로젝트 & 데이터 소개
2. 가설 설정

☑ 프로젝트 소개

현재 MUSINSA 는 MUSINSA 에 입점한 브랜드들의 매출 극대화를 위해 지원 활동에 나서고 있습니다.

요번 프로젝트에서는 단순히 review 별점으로만 상품에 대해서 평가하는 것이 아닌 사용자들이 직접 작성한 review를 자연어처리를 통해 감정 분석해 보고자 합니다. 이를 통해 고객의 의견을 이해하고 제품 개선 및 마케팅 전략 수립에 활용함으로써, 해당 브랜드에게 분석 내용을 공유하여 매출 극대화를 시키기 위함 입니다.

☑ 데이터 소개

- 네이버 쇼핑 홈페이지에서 제품별 review와 별점이 수집된 데이터. – 모델 학습용 데이터
- MUSINSA 홈페이지에서 review 데이터 크롤링 (BeautifulSoup사용)

☑ 가설 설정 🔔

1. MUSINSA review데이터에서 별점으로만 봤을 때 불만족인 경우는 3%이지만, review데이터를 분석해 보면 실제로는 그 이상이다

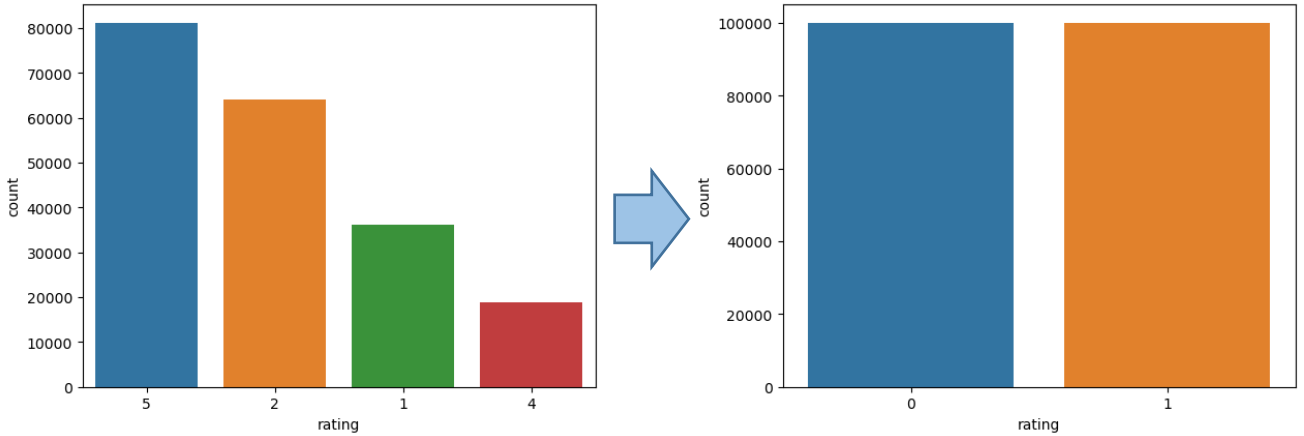
3. 데이터 전처리

☑ 네이버 데이터 전처리

네이버 쇼핑 홈페이지에서 제품별 review와 별점이 수집된 데이터

	rating	review
0	5	배공빠르고 굿
1	2	택배가 엉망이네용 저희집 밑에층에 말도없이 놔두고가고
2	5	아주좋아요 바지 정말 좋아서2개 더 구매했어요 이가격에 대박입니다. 바느질이 조금 ...
3	2	선물용으로 빨리 받아서 전달했어야 하는 상품이었는데 머그컵만 와서 당황했습니다. 전...
4	5	민트색상 예뻐요. 옆 손잡이는 거는 용도로도 사용되네요 ㅎㅎ
...	...	...
199995	2	장마라그런가!!! 달지않아요
199996	5	다이슨 케이스 구매했어요 다이슨 슈퍼소닉 드라이기 케이스 구매했어요가격 괜찮고 배송...
199997	5	로드샵에서 사는것보다 세배 저렴하네요 ㅜㅜ 자주이용할께요
199998	5	넘이쁘고 세련되보이네요~
199999	5	아직 사용해보지도않았고 다른 제품을 써본적이없어서 잘 모르겠지만 ㅎㅎ 배송은 빨랐습니다

200000 rows × 2 columns



- Rating 데이터 타입 정수로 변경.
- 별점이 4,5 점이면 1(만족) 1,2 점이면 0(불만족) 으로 분류
- 중복된 review 텍스트 삭제
- 한글과 공백을 제외하고 제거해주는 정규식 적용

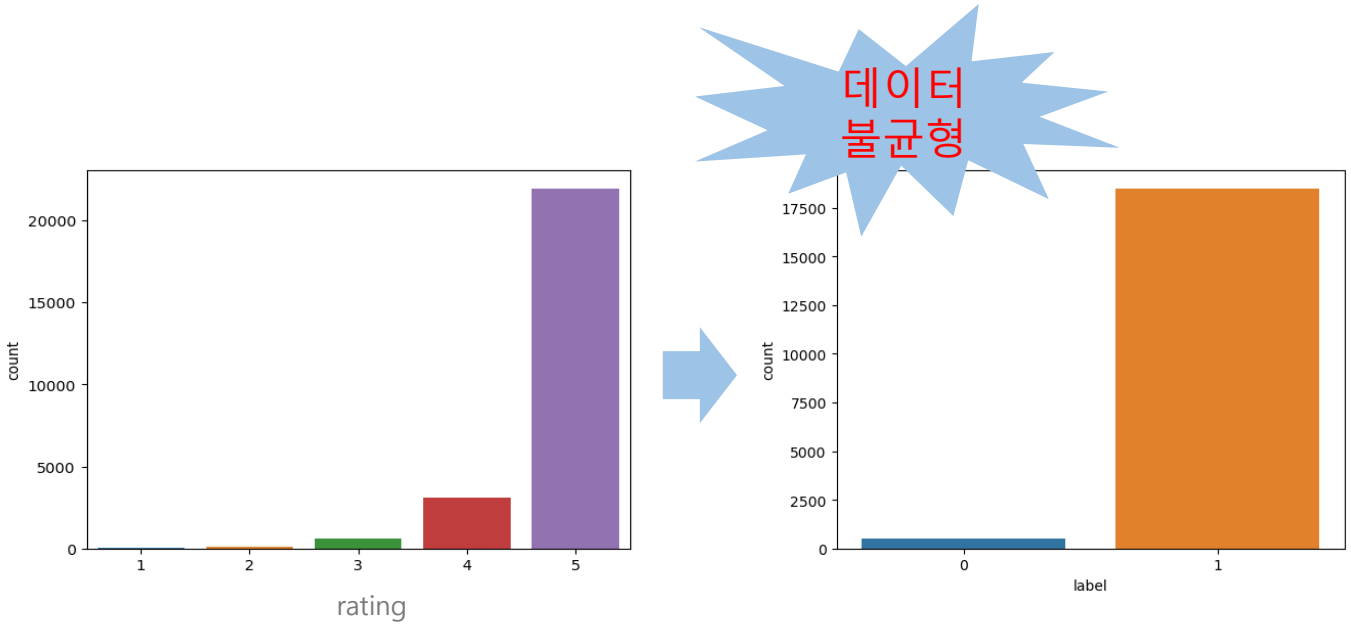
3. 데이터 전처리

MUSINSA 데이터 전처리

MUSINSA 홈페이지에서 review 데이터

	별점	리뷰
0	4	여름에 입을 반팔 좋게 사서 좋아요 많이 입을 것 같아요
1	5	레이어드 용으로 샀는데 길이가 생각보다 조금 더 길어서 애매합니다.ㅠ
2	5	원래 크게 입는 걸 좋아해서 저번에 L사이즈 샀었는데 품 작은 것 같아서 XL샀더니...
3	5	디자인 귀엽고 좋아요 평상시 어떤 옷에도 잘 코디가 되네요
4	5	국민템이긴한데 그만큼 편하게 신기 좋아요 맘에들어요
...	...	...
25780	5	No Review
25781	5	우선 밑단 디테일이 있어서 주위에서 포인트 좋다는 의견을 많이 들었어요 그리고 핏도...
25782	5	이거 사고 계속 이것만 써요 진짜 편하고 좋은 것 같아요
25783	5	색감이 일단 너무너무 이쁘고요. 길이는 딱 요즘 유행하는 바지선길이 까지 오네요! ...
25784	4	정말 가볍네요. 사이즈 때문에 고민 했는데 노트북 넣기에 좋은데 바닥이 조금 앏아서 .

25785 rows × 2 columns



- 별점이 4,5 점이면 1(만족) 1,2,3 점이면 0(불만족) 으로 분류
- review가 Null이거나 No Review인 경우 삭제.
- MUSINSA 데이터의 전처리는 여기까지!

3. 데이터 전처리

	rating	review
0	5	배공빠르고 굿
1	2	택배가 엉망이네용 저희집 밑에층에 말도없이 놔두고가고
2	5	아주좋아요 바지 정말 좋아서2개 더 구매했어요 이가격에 대박입니다. 바느질이 조금 ...
3	2	선물용으로 빨리 받아서 전달했어야 하는 상품이었는데 머그컵만 와서 당황했습니다. 전...
4	5	민트색상 예뻐요. 옆 손잡이는 거는 용도로도 사용되네요 ㅎㅎ
...	...	...
199995	2	장마라그런가!!! 달지않아요
199996	5	다이슨 케이스 구매했어요 다이슨 슈퍼소닉 드라이기 케이스 구매했어요가격 괜찮고 배송...
199997	5	로드샵에서 사는것보다 세배 저렴하네요 ㅜㅜ 자주이용할께요
199998	5	넘이쁘고 세련되보이네요~
199999	5	아직 사용해보지도않았고 다른 제품을 써본적이없어서 잘 모르겠지만 ㅎㅎ 배송은 빨랐습니다

200000 rows × 2 columns

한글 텍스트를 형태소 단위로 분석하는 이유는 한글이 가지는 특성 때문입니다. 한글은 자모음이 결합하여 글자를 이루는 특징을 가지고 있으며, 이로 인해 단어가 띄어쓰기로 구분되는 영어와는 달리, 한글 단어 사이에는 띄어쓰기가 되지 않는 경우가 많습니다. 따라서 한글 텍스트를 형태소 단위로 토큰화 하여 분석하는 것이 자연어 처리 분야에서 일반적으로 사용되는 방법입니다. 이를 통해 단어의 의미와 문맥을 더욱 정확하게 파악할 수 있습니다. 또한, 형태소 분석을 통해 불필요한 정보를 제거하고 중요한 정보에 집중할 수 있으며, 분석 성능을 높일 수 있습니다.

3. 데이터 전처리

한국어를 처리할 때, Konlpy library를 사용합니다.

1) Okt(Open Korea Text)
2) Mecab(메캅)
3) Komoran(코모란),
4) Hannanum(한나눔)
5) Kkma(꼬꼬마)

```
['책상',
 '이',
 '유리',
 '가',
 '올라가다',
 '받치다',
 '없다',
 '힘',
 '을',
 '조금',
 '만',
 '주다',
```

형태소 분석기 중 Okt 를 사용했고, 한글 불용어는 파이썬 라이브러리가 없어서 직접 불용어 사전을 만들어 제거.

```
stopwords = ['하다','이','도','에','너무','요','은','는','가','을','를','그리고',
             '그러나','하지만','있다','없다','아','휴','아이구','아이쿠','아이고',
             '어','나','우리','저희','따라','의해','에','의','가','으로',
             '로','에게','뿐이다','의거하여','근거하여','입각하여','기준으로',
             '예하면','예를 들면','예를 들자면','저','소인','소생','저희','지말고'
             '하지마','하지마라','다른','물론','또한','그리고','비길수 없다',
             '해서는 안된다','불가능하다','무엇','어느','어떤','아래윗','조차',
             '한데','그럼에도 불구하고','여전히','심지어','까지도','조차도',
             '하지 않는다면','않으면','만 못하다','하는 편이 낫다','불완전하다',
             '투자한다','생각한다','입니다','요','ㅎ','ㅎㅎ','ㅎㅎㅎ','ㅠ','ㅠㅠ',
             'ㅠㅠㅠ','ㅜ','ㅜㅜ','ㅜㅜㅜ','네','때','에는','가','각','것','ㅏ']
```

일반적으로 한글 자연어 처리에서는 형태소 분석한 다음에 **토큰화(tokenization)**->패딩(Padding)

4. 딥러닝 모델

MLP

```
model_mlpp = Sequential()
model_mlpp.add(Dense(64, activation='relu', input_shape=(25,)))
model_mlpp.add(Dense(64, activation='relu'))
model_mlpp.add(Dense(1, activation='sigmoid'))
```

LSTM

```
model_lstm = Sequential()
model_lstm.add(Embedding(vocab_size, 100))
model_lstm.add(LSTM(128, dropout=0.3))
model_lstm.add(Dense(1, activation='sigmoid'))
```

CNN

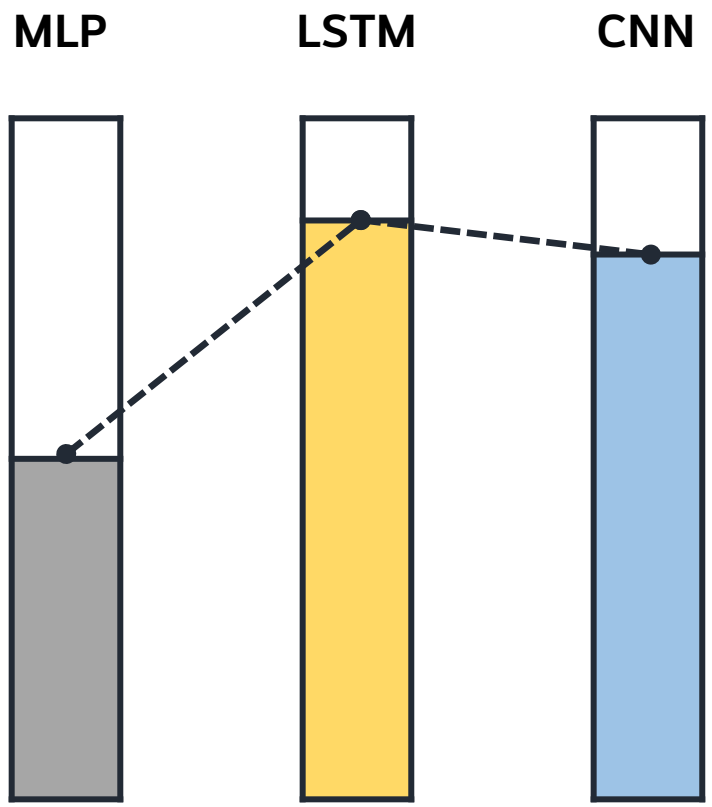
```
model_cnn = Sequential()
model_cnn.add(Embedding(input_dim=vocab_size, output_dim=embedding_dim, input_length=25))
model_cnn.add(Conv1D(filters=64, kernel_size=5, activation='relu'))
model_cnn.add(GlobalMaxPooling1D())
model_cnn.add(Dense(units=32, activation='relu'))
model_cnn.add(Dense(1, activation='sigmoid'))
```

EarlyStopping
validation loss가 4 epoch 연속으로 감소하지 않을 경우 학습을 멈추도록 설정.

모델 컴파일
이진 분류 문제이기 때문에 'binary_crossentropy' 손실함수 사용.
최적화 알고리즘으로 adam 사용.
불균형한 데이터가 아니기 때문에 평가지표로 정확도 사용.

CNN모델의 경우 이미지 뿐만 아니라 텍스트 분석에도 사용.

5. 모델 평가



	MLP	LSTM	CNN
정확도	0.57	0.89	0.87
정밀도	0.56	0.88	0.84

최종 모델 : LSTM

- CNN과 큰 차이는 아니지만, 조금이라도 정확도가 높은 LSTM 모델 선택.

- CNN 모델의 이미지 분류에 많이 사용되지만, 텍스트 분석에도 사용돼서 시도.

- 정확도가 높은 점도 있지만, 불만족인데 만족이라고 판단하는 에러가 적어야 하기 때문에 정밀도가 높은 모델 선택.

5. 모델 평가

보통 별점을 1점 주고 좋은 review를 작성하지 않는다고 생각한다.
별점이 5점인 경우, review 내용이 정말 좋은 지 체크해본다.

1. 허리가 좀 크지만 괜찮아요 좀 얇아서 겨울엔 추울까봐 걱정했는데 오늘 입어보니까 안 춥네요!
✔ 별점 5점 -> 88.75% 확률로 만족인 review

2. 스키장 온 느낌쓰... 코디하기 좀 어렵지만 그래도 존예
✔ 별점 5점 -> 94.54% 확률로 만족인 review

3. 가벼운 선크림 원하시는 분들은 이거 사시면 될 거같아요 근데 냄새가 좀
✔ 별점 5점 -> 66.38% 확률로 만족인 review

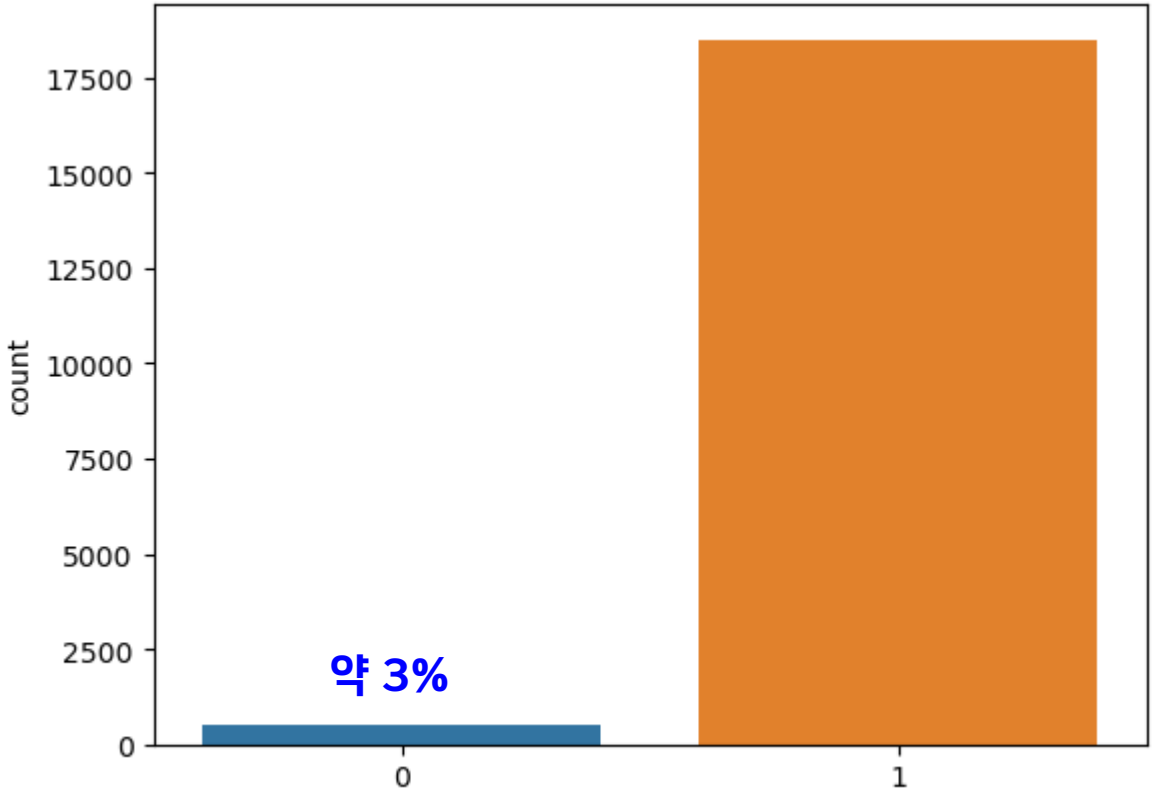
4. 딱 좋아요 근데 한 두번 입으니까 완전 보풀 일어나서 주의하세요
✔ 별점 5점 -> 50.20% 확률로 불만족인 review

LSTM

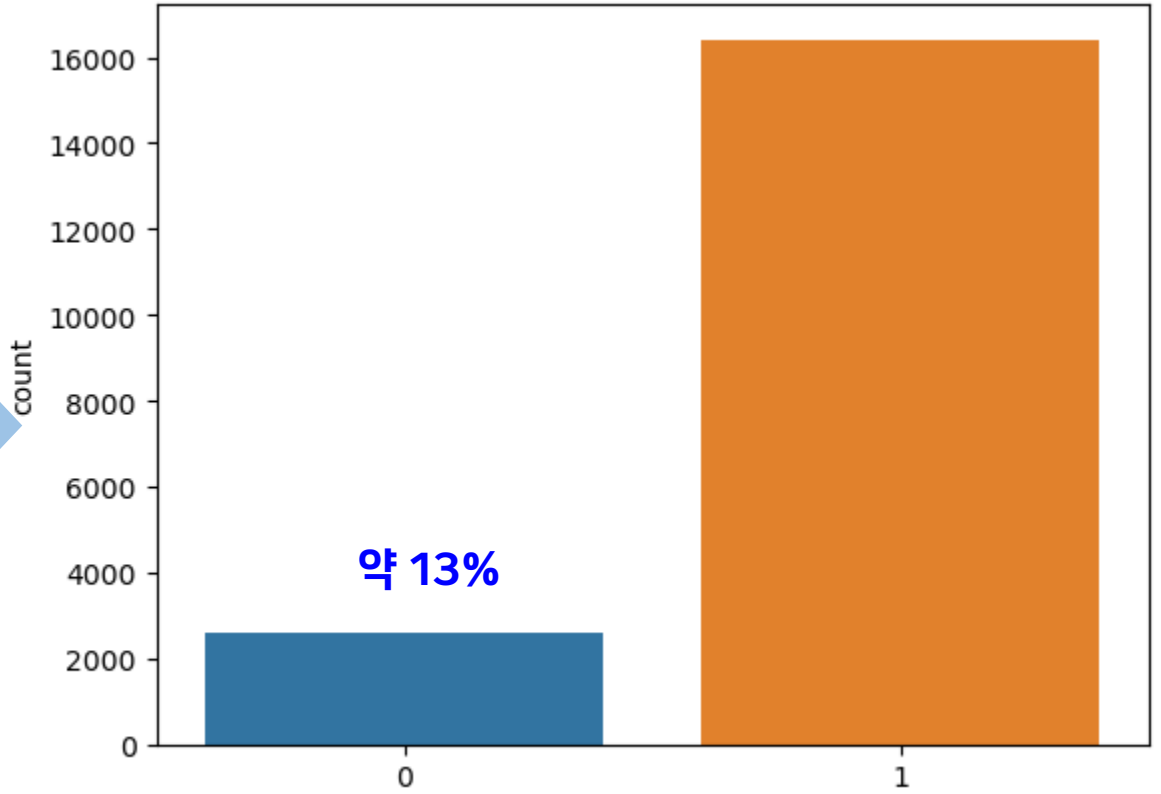
5. 모델 평가

가설 채택

MUSINSA review 데이터에서 별점으로만 봤을 때 불만족인 경우는 3%이지만, review 데이터를 분석해 보면 실제로는 그 이상이다



MUSINSA 원본 데이터의
만족(1), 불만족(0) 분포



LSTM 모델에 적용한 후
MUSINSA 데이터의
만족(1), 불만족(0) 분포

6. 한계점, 추후 해결 방안

1. 한글 텍스트의 경우 띄어쓰기나 단어가 아닌 형태소로 나누어 분석하는데 있어서 형태소에 대한 지식이 부족했음.

2. transformer나 다른 모델들도 활용해서 학습해볼 예정.

3. LSTM , CNN 모델의 경우 과적합이 있어서 생각보다 좋은 성능이 나온 것은 아니라고 판단되어 추후에 레이어 층을 삭제하거나 하이퍼 파라미터 튜닝을 통해 개선 예정.

Q & A