AI 01 김세진

Sentiment Analysis Project

프로젝트 주제 및 배경

네이버 쇼핑 리뷰 데이터를 이용하여 감성 분석하기

제품이나 서비스의 사용 리뷰를 남기는 사용자들은, 보통 자신의 만족도에 따라 별점으로 서비스의 질을 평가하고 리뷰를 남긴다. 하지만 종종 그렇지 않은 리뷰들도 있다. 별점을 낮게 주면 판매자가 신고 혹은 삭제를 하는 경우도 있기 때문에, 별점은 높게 주고 리뷰 내용에만 자신이 느꼈던 불편한 점을 기재하는 경우도 존재한다. 그래서 별점으로만 리뷰를 분류하게 되면 그 리뷰가 실제로 긍정적인 리뷰인지, 부정적인 리뷰인지 파악하기 어렵다. 그래서 NLP를 활용하여 리뷰 내용 자체를 분석하여 리뷰에 담긴 긍정/부정적 감정들을 알아보고 싶었다.

사용한 데이터

	ratings	reviews
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

네이버 쇼핑 홈페이지에서 수집한 제품별 후기 데이터(별점 1,2,4,5점에 해당하는 리뷰만 수집)

수집 기간 : 2020.06~2020.07

데이터 건수 : 20만 건

데이터 분석 목표

1. 한국어 형태소 분석기를 통해 리뷰 데이터를 토큰화 한다. 이 과정에서 두 가지의 분석기를 사용해 보고, 성능이 더 좋은 분석기를 통해 이후 작업을 진행한다.
2. 토큰화한 리뷰 데이터를 시각화하여 리뷰에 가장 많이 쓰이는 단어들을 찾아본다.
3. GRU와 LSTM 모델을 사용하여 리뷰 데이터 속에서 드러나는 긍정적/부정적 단어들을 통해 데이터를 분류하고, 모델 학습에 사용하지 않은 리뷰를 사용하여 리뷰가 긍정적인지, 부정적인지 예측해 본다.

전처리 과정

중복 데이터 제거

별점으로 라벨링 : 별점이 4점 이상이면 긍정, 2점 이하이면 부정 리뷰로 분류

정규표현식을 사용하여 한글 이외의 영어, 특수문자 모두 제거

한국어 형태소 분석기 중 Okt와 Mecab 을 통해 각각 토큰화

시각화 - Mecab

'네요': 33946, '는데': 21646, '안': 21063, '어요': 15822, '너무': 14183, '있': 14107, '했': 12389, '좋': 10472, '배송': 10306, '같': 9620, '거': 9467, '어': 9425, '구매': 9383, '아요': 9281, '없': 9273

'좋': 42133, '아요': 22433, '네요': 21289, '어요': 19862, '잘': 19824, '구매': 17280, '습니다': 14421, '있': 13208, '배송': 12925, '는데': 12389, '합니다': 10439, '했': 10437, '먹': 10389, '재': 9918, '너무': 8986

시각화 - Okt

'너무': 14161, '안': 12449, '배송': 9771, '그냥': 9085, '잘': 8416, '로': 6177, '했는데': 5988, '별로': 5967, '못': 5485, '제품': 5419, '으로': 5306, '생각': 5276, '좀': 5217, 'ㅠㅠ': 5191, '사용': 5173

'잘': 16320, '좋아요': 15337, '배송': 12884, '너무': 9712, '재구매': 9184, '구매': 7629, '사용': 5850, '가격': 5346, '같아요': 4747, '으로': 4682, '로': 4661, '좋네요': 4471, '제품': 4228, '보다': 4209, '빠르고': 4190

한국어 형태소 분석기 Mecab과 Okt 비교

Mecab : 속도가 빠르지만, 토큰화 과정에서 어간과 어미를 분리하기 때문에 의미적 특징을 찾아내기 어렵다.

Okt : 속도는 Mecab보다 다소 느리지만, 사람이 이해하기 쉬운 정도로 문장을 분리하고 경우에 따라서는 오타까지 수정해서 정규화 하기 때문에 의미적 특징을 확실히 찾아낼 수 있다.

Mecab

'좋': 42133, '아요': 22433, '네요': 21289, '어요': 19862, '잘': 19824, '구매': 17280, '습니다': 14421, '있': 13208, '배송': 12925, '는데': 12389, '합니다': 10439, '했': 10437, '먹': 10389, '재': 9918, '너무': 8986

Okt

'잘': 16320, '좋아요': 15337, '배송': 12884, '너무': 9712, '재구매': 9184, '구매': 7629, '사용': 5850, '가격': 5346, '같아요': 4747, '으로': 4682, '로': 4661, '좋네요': 4471, '제품': 4228, '보다': 4209, '빠르고': 4190

모델 학습 – 딥러닝 모델

GRU

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, None, 100)         4361200
_________________________________________________________________
gru_1 (GRU)                  (None, 128)               88320
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 129
=================================================================
Total params: 4,449,649
Trainable params: 4,449,649
Non-trainable params: 0
_________________________________________________________________
```

Accuracy : 0.91481

LSTM

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, None, 100)         4361200
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 101
=================================================================
Total params: 4,441,701
Trainable params: 4,441,701
Non-trainable params: 0
_________________________________________________________________
```

Accuracy : 0.91547

모델 학습 – 머신러닝 모델

Random Forest

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=500,
                       n_jobs=None, oob_score=False, random_state=11, verbose=0,
                       warm_start=False)
```

Accuracy : 0.69786

Logistic Regression

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

Accuracy : 0.52675

긍정/부정 리뷰 예측

GRU

좋았습니다 주문을 미루고 미루다보니 시간이 급해서 당일제작 옵션 선택해서 진행했습니다.	100.00% 확률로 긍정 리뷰입니다.
재질도 좋고 색감도 예뻐요 만족스러워서 서랍장 하나 더 구매했습니다	100.00% 확률로 긍정 리뷰입니다.
라면 냄비로 구입 했었는데 제가 생각 했던것보다 많이 작았네요	91.62% 확률로 부정 리뷰입니다.
최악이다 기포 안없어져서 다시 뜯었는데 그대로 굳었어 ㅋ	100% 확률로 부정 리뷰입니다.

LSTM

좋았습니다 주문을 미루고 미루다보니 시간이 급해서 당일제작 옵션 선택해서 진행했습니다.	100.00% 확률로 부정 리뷰입니다.
재질도 좋고 색감도 예뻐요 만족스러워서 서랍장 하나 더 구매했습니다	100.00% 확률로 부정 리뷰입니다.
라면 냄비로 구입 했었는데 제가 생각 했던것보다 많이 작았네요	99.62% 확률로 부정 리뷰입니다.
최악이다 기포 안없어져서 다시 뜯었는데 그대로 굳었어 ㅋ	98.34% 확률로 부정 리뷰입니다.

긍정/부정 리뷰 예측

Random Forest

좋았습니다 주문을 미루고 미루다보니 시간이 급해서 당일제작 옵션 선택해서 진행했습니다.	100.00% 확률로 부정 리뷰입니다.
재질도 좋고 색감도 예뻐요 만족스러워서 서랍장 하나 더 구매했습니다	100.00% 확률로 긍정 리뷰입니다.
라면 냄비로 구입 했었는데 제가 생각 했던것보다 많이 작았네요	100.00% 확률로 긍정 리뷰입니다.
최악이다 기포 안없어져서 다시 뜯었는데 그대로 굳었어 ㅋ	100.00% 확률로 부정 리뷰입니다.

Logistic Regression

좋았습니다 주문을 미루고 미루다보니 시간이 급해서 당일제작 옵션 선택해서 진행했습니다.	100.00% 확률로 부정 리뷰입니다.
재질도 좋고 색감도 예뻐요 만족스러워서 서랍장 하나 더 구매했습니다	100.00% 확률로 부정 리뷰입니다.
라면 냄비로 구입 했었는데 제가 생각 했던것보다 많이 작았네요	100.00% 확률로 부정 리뷰입니다.
최악이다 기포 안없어져서 다시 뜯었는데 그대로 굳었어 ㅋ	100.00% 확률로 부정 리뷰입니다.

결론

1. 한국어 형태소 분석기는 Okt가 더 나은 성능을 보였기 때문에 감성분석에는 Okt를 사용한 토큰화 결과를 학습시켰다.
2. 시각화 결과, Mecab과 Okt를 사용하여 토큰화 한 데이터를 시각화했을 때 긍정적인 리뷰에서는 '좋아요', '너무', '잘'과 같은 단어가 많이 나타났고, 부정적인 리뷰에서는 '안', '너무' 와 같은 단어가 많이 나타났다.
3. GRU 모델에서는 감성 분석이 제대로 되었던 반면, LSTM 모델은 학습한 모델의 정확도는 높았지만 실제 리뷰를 분석해 보았을 때 제대로 된 결과가 나오지 않았다. 머신러닝을 활용한 모델들과 비교했을 때는, 딥러닝 모델의 정확도가 더 높게 나타나고 감성분석도 비교적 더 잘 되었다는 것을 볼 수 있었다.

Thank You