**기말 대체 과제 보고서**

**제품 review 데이터의 감성 분석**

|  |  |  |
| --- | --- | --- |
| **학 과** |  | **컴퓨터공학과** |
| **학 번** |  | **202121059** |
| **이 름** |  | **정보경** |
| **담당**  **교수님** |  | **백우진 교수님** |



# **감정 분석의 기본 이해**

감정 분석(Sentiment Analysis)은 텍스트 데이터에서 감정, 의견, 또는 태도를 식별하고 추출하는 것입니다. 감정분석은 자연어 처리(NLP)의 한 분야로, 주로 사람들이 작성한 텍스트에서 긍정적, 부정적, 중립적 감정을 자동으로 감지합니다. 감정 분석은 텍스트 전처리, 특성 추출, 감정 분류, 감정 점수 계산의 순서로 진행됩니다. 감정 분석을 응용하는 분야로는 소셜, 미디어 모니터링, 고객 리뷰 분석, 여론 분석, 고객 서비스 개선, 시장 조사가 있습니다.

# **데이터 전처리**

학번이 홀수로 끝나기 때문에 amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv 데이터를 활용하여 진행합니다.

1. 먼저 데이터를 읽어와서 데이터 프레임으로 저장하고 확인합니다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

1. 데이터의 정보를 확인하기 위해 info() 함수를 사용하여 각 열의 정보를 확인합니다. 해당 데이터에는 총 11개의 열이 있습니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

1. 각 열이 어떤 정보를 담고 있는지 자세히 확인하기 위해 0번 행을 샘플 데이터로 하여 각 열의 정보를 확인해 보았습니다.

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

1. 샘플 데이터를 통하여 확인한 각 열의 정보는 다음과 같습니다.

텍스트, 스크린샷, 폰트, 영수증이(가) 표시된 사진

자동 생성된 설명

1. 11개의 열 중에서 감정 분석에 사용할 열을 선택합니다. 총 4개 review\_title, review\_text, review\_rating, uniq\_id를 선택하였습니다. 그리고 이를 df라는 새로운 데이터프레임으로 저장합니다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

선택한 4개의 열을 선택한 이유는 다음과 같습니다.

review\_title - 리뷰 제목으로, 제목은 내용을 요약한다고 생각하여 포함하였습니다.

review\_text - 리뷰 내용으로, 감정이 가장 잘 드러나는 열이라고 생각합니다.

reveiw\_rating - 감정 분석 후 맞게 분석하였는지 확인하는 지표로 사용됩니다.

uniq\_id - 데이터의 중복을 제거하기 위해 사용됩니다.

1. 전체 데이터와 중복 데이터 개수를 확인합니다. Uniq\_id를 기준으로 개수를 확인하였으며, 총 364개의 중복 데이터가 있음을 확인하였습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

1. 어떤 형태로 중복 데이터가 존재하는지 확인하였습니다. 확인 결과 스크랩 시간과 url이 다른 것을 제품의 판매자가 다르게 여러 개 등록되었다고 예상할 수 있었습니다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

1. 중복데이터를 제거하고, 제거된 데이터를 clean\_data로 저장합니다.





텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

1. Isnull()을 사용하여 결측치를 확인합니다. Review\_title는 1개, review\_text는 8개가 존재하였습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

1. Title과 text가 비어 있는 행을 확인합니다. Title과 text가 동시에 비어 있는 경우에는 감정 분석이 어렵기 때문에 제거하는 것이 좋지만, title이 비어 있는 경우는 text가 채워져 있고, text가 비어 있는 경우에는 title이 채워져 있었습니다. Title이 비어 있는 경우에는 text를 통해 감정 분석을 진행할 수 있고, text가 비어 있는 경우는 title의 내용과 rating을 확인했을 때 두 개가 연관성이 있음을 볼 수 있었습니다. 그렇기 때문에 title과 text모두 제거하지 않고 공백으로 채우기로 결정하였습니다.

텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명



텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명



텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

1. 정규식을 활용하여 특수문제를 제거하는 함수를 작성하였습니다. 특수문제를 제거하도록 구성하였고, review\_title과 review\_text에 적용하였습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

1. Word\_tokenize를 활용하여 토큰화를 하는 함수를 작성하였습니다. 특수문자 제거와 동일하게 review\_title과 review\_text에 적용하였습니다. 결과를 확인했을 때 토큰화가 되어 리스트로 저장된 것을 확인할 수 있습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

1. Stopwords.word를 사용하여 불용어를 제거하는 함수를 작성하였고, review\_title과 review\_text에 대해 불용어 처리를 진행하였습니다. 불용어 처리 결과 0번 행의 text를 보면 [Love, these, Was, looking, for, converses, ... ] 에서 [Love, Was, looking, converses, half, price, ...]로 바뀐 것을 볼 수 있습니다.

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

1. WordNetLemmatizer와 PorterStemmer을 사용하여 표제어 추출과 스테밍을 하는 함수를 작성하였습니다. review\_title과 review\_text에 대해 토큰화를 진행하였습니다. 표제어 추출은 문맥에 따른 변형을 처리하여 단어의 기본 형태를 반환하고, 스테밍은 단어의 접미사나 변형을 규칙적으로 제거하여 어근을 반환합니다. 둘 중 하나만 진행해도 되지만, 두 개 다 처리하면 일반화를 더 잘 할 수 있을 것이라고 생각하여 둘 다 진행하였습니다. 정규화 진행 후 리스트의 값들을 다시 공백으로 합쳐주었습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

# **모델 선택 및 학습**

1. 수업 시간에 사용한 vader 감정 분석기를 활용하여 감정 점수를 계산하고 회귀 분석을 위한 모델을 구현합니다. Vader 모델은 텍스트의 감정을 분석하기 위해 설계된 감정 분석 도구입니다. Vader는 비정형 텍스트를 대상으로 할 수 있다는 점과 부정적인 단어를 처리할 수 있다는 장점이 있습니다. 그러나, 영어 외의 언어 분석은 어렵고 문맥을 이해하는 것이 어려울 수 있다는 단점이 있습니다.
2. Vader 감정 분석기를 초기화하고 감정 점수를 계산하는 함수를 정의합니다. 함수는 vader를 사용하여 계산하고 compound 점수를 반환합니다. Review\_title과 review\_text에 대해 감정 계산을 하고 각 점수는 new\_title과 new\_text로 저장합니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 영수증, 폰트이(가) 표시된 사진

자동 생성된 설명

1. New\_title과 new\_text에 해당하는 점수를 회귀분석에 사용하기 위해 0과 1로 변경하여 저장합니다. 감정 점수가 0.1이상이면 1, 그렇지 않으면 0으로 리턴하고 각 값은 new\_title\_data, new\_text\_data로 저장합니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

각 값이 0과 1로 매칭된 것을 볼 수 있습니다. 0번 행을 보면 별점은 5.0점이고, 각 분석이 1로 긍정 리뷰라고 잘 평가하였습니다. 1번 행의 경우에는 벌점이 2.0이라는 낮은 점수이므로 부정적이라고 볼 수 있습니다. title에 대해서는 부정으로 잘 평가했지만, text의 감정은 긍정이라고 잘못 평가하였습니다. 6822번 행의 경우에는 별점이 5점으로 긍정 리뷰이지만 title과 text 모두 부정으로 평가했습니다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

1. 이번에는 두 개를 모두 사용할 수 있도록 계산하였습니다. New\_title\_data와 new\_text\_data의 값을 더하여 사용하였고 하나라도 부정이 있으면, 부정으로 판단하고 계산하였습니다. 2면 긍정-긍정이므로 긍정, 1이면 긍정-부정이므로 부정, 0이면 부정, 부정이므로 부정으로 계산하였습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

1. 감정 분석이 잘 진행되었는지 확인하기 위해 별점을 기준으로 확인합니다. 별점은 3점을 기준으로 이상이면 긍정, 이하면 부정으로 생각하고 계산을 진행하였습니다. Title, text, 두 개 합으로 진행하였습니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

세 개 중 가장 잘 예측한 것은 text를 사용하여 예측한 것입니다. 그래서 회귀분석은 text를 사용하여 진행합니다.

1. 데이터 셋을 분할하고 학습합니다. 벡터화를 진행하고 학습 데이터는 train과 test 8:2로 분할하였습니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

1. 로지스틱 회귀 모델을 만들고 그리드 서치를 사용하여 하이퍼 파라미터 튜닝을 진행합니다. 교차 검증은 5로 설정하였습니다. 그리드 서치가 선택한 최적의 하이퍼 파라미터는 c = 10이었습니다. C는 로지스틱 회귀 모델의 규제 강도를 제어하는 하이퍼파라미터로, 값이 클수록 규제가 약해지고, 값이 작을수록 규제가 강해지는데, 10을 통해 규제가 약간 작용하면서도 모델이 잘 학습되었다는 것을 알 수 있습니다. 모델의 정확도는 92.63%로 높은 성능을 갖고 있다는 것을 알 수 있습니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

1. 최적의 모델을 사용하여 학습을 진행합니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

# **모델 평가**

1. Accuracy (정확도): 0.93의 높은 수치로, 모델이 대부분의 경우 올바르게 예측했다는 것을 알 수 있습니다.
2. Precision (정밀도): 0.93의 높은 수치로, 모델이 양성 클래스를 예측할 때 대부분 올바르게 예측했다는 것을 알 수 있습니다.
3. Recall (재현율): 0.93의 높은 수치로, 모델이 실제 양성 중 모델이 양성으로 올바르게 예측했다는 것을 알 수 있습니다.
4. F1 Score: 0.93의 높은 수치로, 모델이 양성과 음성 예측 모두에서 균형 잡힌 성능을 보이고 있다는 것을 알 수 있습니다.
5. ROC-AUC Score: 0.926로, 1에 가까운 수치인데, 이는 모델이 양성과 음성을 잘 구분하고 있다는 것을 알 수 있습니다.
6. Classification Report: 모델은 전체적으로 높은 성능(정확도 93%, 정밀도 93%, 재현율 93%, F1 스코어 93%)을 보이며, 긍정 리뷰(94% 정밀도, 93% 재현율)와 부정 리뷰(91% 정밀도, 92% 재현율) 모두에서 균형 잡힌 예측을 하고 있습니다.
7. 오버 피팅: 현재 모델의 정확도가 93%로 매우 높고, 다른 성능 지표들도 균형 잡혀 있습니다. 모델의 성능이 테스트 데이터에서도 높게 유지되므로, 오버피팅이 없다고 볼 수 있습니다.
8. 언더 피팅: 현재 모델의 성능 지표(정확도, 정밀도, 재현율, F1 점수)가 모두 높습니다. 모델이 데이터를 잘 학습하고, 테스트 데이터에서도 좋은 성능을 보임으로써, 언더피팅 문제는 없음을 확인할 수 있습니다.

이를 통해 모델은 매우 우수한 성능을 갖고 있다고 볼 수 있습니다.

텍스트, 스크린샷, 영수증, 폰트이(가) 표시된 사진

자동 생성된 설명

# **시각화**

1. confusion\_matrix를 matplotlib을 활용하여 시각화합니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

* True Negative (0.92)
  + 모델이 부정 리뷰를 부정으로 정확히 예측한 비율입니다.
  + 모델이 부정 리뷰를 잘 분류하고 있음을 나타냅니다.
* False Positive (0.075)
  + 모델이 부정 리뷰를 긍정으로 잘못 예측한 비율입니다.
  + 7.5%의 부정 리뷰가 잘못 분류되었다는 것을 알 수 있습니다.
* False Negative (0.072)
  + 모델이 긍정 리뷰를 부정으로 잘못 예측한 비율입니다.
  + 긍정 리뷰의 7.2%가 잘못 분류되었으며, 재현율이 높다는 것을 나타냅니다.
* True Positive (0.93)
  + 모델이 긍정 리뷰를 긍정으로 정확히 예측한 비율입니다.
  + 긍정 리뷰의 93%가 정확하게 분류되었다는 것을 알 수 있습니다.
* 모델의 정확도는 약 92.5%로, 전반적인 성능이 매우 좋습니다.
* 정밀도는 92.5%로, 모델이 긍정 리뷰로 예측한 것 중 대부분이 실제 긍정 리뷰임을 나타냅니다.
* 재현율은 92.9%로, 모델이 실제 긍정 리뷰를 거의 모두 정확히 예측하고 있음을 보여줍니다.
* F1 점수는 92.7%로, 정밀도와 재현율의 균형을 나타내며, 모델의 전반적인 성능을 종합적으로 평가합니다.

모델은 부정 리뷰와 긍정 리뷰 모두에서 충분히 높은 성능을 갖고 있기 때문에 성능 개선을 필요성이 낮다고 볼 수 있습니다.

1. ROC 곡선을 matplotlib을 활용하여 시각화합니다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

라인, 텍스트, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

ROC 곡선이 대각선 기준선보다 위에 위치할수록 모델의 성능이 우수함을 나타내는데, 주황색 실선이 대각선보다 위에 있으므로, 모델이 무작위 분류기보다 성능이 좋음을 알 수 있습니다. AUC는 1에 가까울수록 성능이 좋다는 것을 의미하는데, 0.926이라는 것을 통해 모델이 분류를 잘 한다는 것을 알 수 있습니다. 즉, ROC 곡선과 AUC 값 모두 높은 값을 가지고 있으므로 모델의 성능이 매우 좋다는 것을 알 수 있습니다.