제품 review 데이터의 감성 분석

|  |  |
| --- | --- |
| **과목** | **기계학습** |
| **담당 교수** | **백우진 교수님** |
| **학과** | **문헌정보학과** |
| **학번 이름** | **202120634 이은비** |
| **제출 일자** | **2024.06.19** |



**목 차**

**I. 서론**  3p

1. 프로젝트 개요 3p

2. 감성 분석의 기본 이해 3p

3. 데이터셋 소개 3p

**II. 프로젝트 수행** 4p

1. 데이터 전처리 4p

2. 모델 선택 및 학습 7p

3. 평가 13p

4. 시각화 및 해석 17p

**III. 결론**  20p

**I. 서론**

1. 프로젝트 개요

이 프로젝트의 목표는 텍스트 데이터를 감성으로 분류하는 감성 분석 시스템을 개발하는 것이다. 이를 통해 데이터 전처리, 모델 선택, 학습, 평가 및 결과 시각화를 경험한다.

2. 감성 분석의 기본 이해

감성 분석(Sentiment Analysis)은 텍스트 데이터를 분석하여 그 감정을 분류하는 기술이다. 이는 제품 리뷰, 소셜 미디어 포스트 등에서 사용자 감정을 파악하는 데 사용된다. 감성 분석의 응용은 마케팅 전략 수립, 고객 만족도 분석 등 다양한 부분에서 이루어진다.

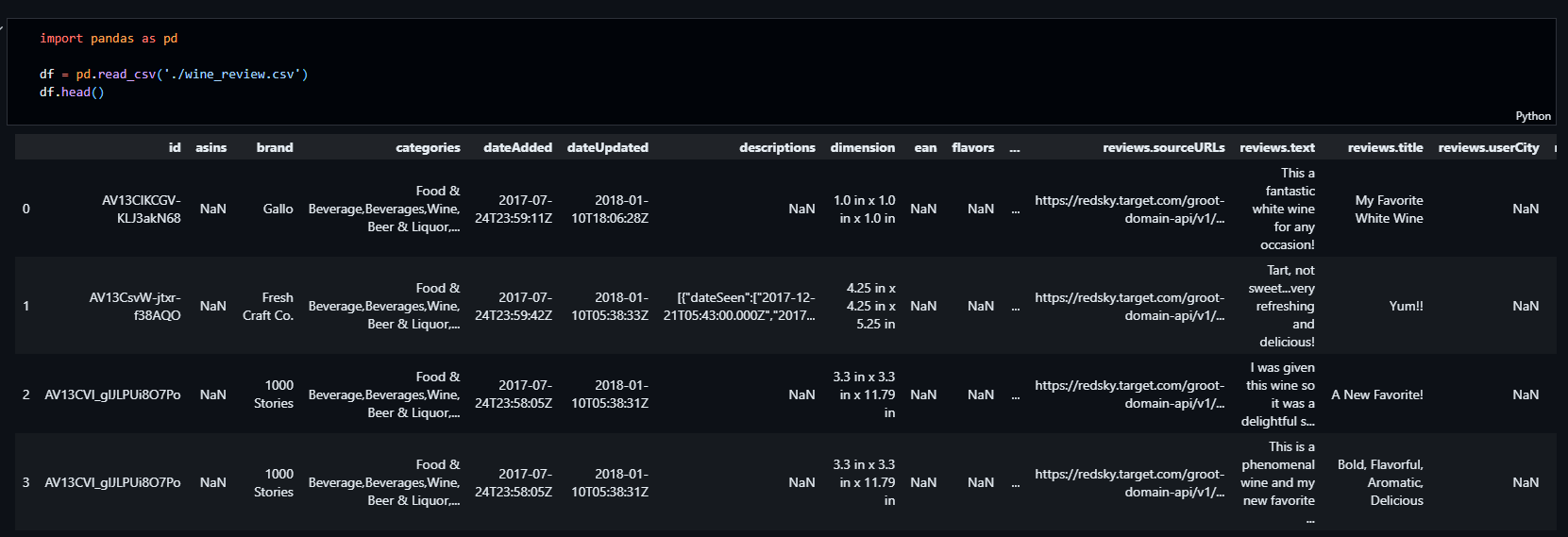
3. 데이터셋 소개

학번이 홀수로 끝나는 학생은 amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv를 짝수로 끝나는 학생은 wine\_review.csv를 분석한다. 나는 202120634로 짝수로 끝나기 때문에 wine\_review.csv를 사용하였다. 이 데이터셋은 다음과 같은 항목으로 구성되어 있다.

id: 아이디 (상품의 고유 식별자) ,asins: ASIN 번호 (아마존 상품 식별 번호), brand: 브랜드, categories: 카테고리, dateAdded: 추가된 날짜, dateUpdated: 업데이트된 날짜, descriptions: 상품 설명, dimension: 상품 크기 (차원), ean: EAN 번호 (유럽에서 사용되는 상품 식별 번호), flavors: 맛, keys: 키워드, manufacturer: 제조사, manufacturerNumber: 제조사 번호, name: 상품 이름, reviews.date: 리뷰 작성 날짜, reviews.dateAdded: 리뷰 추가된 날짜, reviews.dateSeen: 리뷰 확인된 날짜, reviews.didPurchase: 구매했는지 여부, reviews.doRecommend: 추천 여부, reviews.id: 리뷰 ID, reviews.numHelpful: 유용한 리뷰 수, reviews.rating: 평점, reviews.sourceURLs: 리뷰 출처 URL, reviews.text: 리뷰 내용, reviews.title: 리뷰 제목, reviews.userCity: 사용자 도시, reviews.userProvince: 사용자 주, reviews.username: 리뷰 작성자 이름, sizes: 사이즈, sourceURLs: 출처 URL, upc: UPC 번호 (북미에서 사용되는 상품 식별 번호), weight: 무게

**II. 프로젝트 수행**

1. 데이터 전처리



▲ 데이터를 처리하기 전 데이터의 구조를 확인하기 위해 데이터프레임을 불러온다.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

▲ 데이터 전처리에 필요한 라이브러리를 설치하고 임포트한다.

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

▲ 'reviews.text'와 'reviews.rating' 열의 결측치를 제거하고 빈 문자열로 채운다. 결측치가 포함된 데이터를 사용하여 모델을 학습할 경우, 결측치 때문에 모델이 잘못된 패턴을 학습할 수 있다. 이는 모델의 성능을 저하로 이어진다. 따라서 결측치를 제거하여 모델이 정확한 패턴을 학습할 수 있도록 한다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

▲ 결측값을 제거하고 빈 문자열로 채운 결과값이다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

▲ 와인 리뷰 데이터를 전처리하여 텍스트를 토큰화하고 불용어를 제거하는 과정을 수행하였다. 이 과정은 텍스트 데이터를 보다 분석하기 쉽게 만들기 위해 중요 단어만을 남기고, 의미 없는 단어들을 제거한다. 불용어란 분석에 불필요한 단어들로, 예를 들어 "the", "is", "in" 등이 있다. 강조어는 분석에 영향을 미칠 수 있다고 생각하여 제거하지 않았다.

텍스트, 스크린샷, 폰트, 메뉴이(가) 표시된 사진

자동 생성된 설명

▲ 텍스트 토큰화와 불용어 제거의 결과는 다음과 같다.

2. 모델 선택 및 학습

텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명  
▲ 텍스트 데이터의 표제어 추출(lemmas)을 수행하여 processed\_text 열을 생성한 후 VADER 감정 분석기를 사용하여 텍스트의 감정 점수를 계산한다.

VADER는 주어진 텍스트의 감정을 양수, 음수, 중립으로 점수화 하며, compound 점수는 이 세 점수를 합친 최종 감정 점수이다. 이 점수를 vader\_compound 열에 저장한다. 원본 리뷰 텍스트(reviews.text), 전처리된 텍스트(processed\_text), 리뷰 평점(reviews.rating)을 포함한 cleaned\_wine\_review.csv 파일을 생성하여 저장한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

▲ 그 결과는 위와 같다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

▲ 다음으로 주어진 텍스트에 대해 VADER 감정 분석기를 사용하여 감정 점수를 계산하는 함수 analyze\_sentiment를 정의한다.

이 함수는 텍스트를 입력 받아 VADER의 polarity\_scores 메서드를 사용해 감정 점수를 계산하고, compound 점수를 반환한다. VADER 감정 분석기를 사용하여 각 리뷰 텍스트의 감정 점수가 sentiment\_score 열에 저장하였다. 이 과정은 와인 리뷰 텍스트 데이터를 정규화하고 감정 점수를 수치적으로 평가하여, 이후의 분석 작업에 사용할 수 있는 기초 데이터를 제공한다. sentiment\_score는 텍스트의 전체적인 감정을 나타내며, -1(매우 부정적)에서 1(매우 긍정적) 사이의 값을 가진다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

▲ 첫번째 리뷰에서 추출한 단어는 fantastic, white, wine, occasion이고, 감정 점수는 0.5574이다. 감정 분석기는 첫번째 리뷰를 긍정에 가깝게 판단했다.

fantastic이라는 단어가 포함되어 있어 그런 것으로 추측한다.

세번째 리뷰에서 추출한 단어는 given, wine, delightful, surprise, find, flavorful이고, 감정 점수는 0.9118이다. 감정 분석기는 세번째 리뷰를 긍정으로 판단했다. delightful, surprise, flavorful가 긍정의 의미를 보이기 때문이다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

▲ VADER 감정 분석기를 사용하여 전처리된 텍스트(processed\_text)의 감정 점수를 계산하고, sentiment\_score 열에 저장하였다.

리뷰 평점(reviews.rating)과 VADER 감정 점수(sentiment\_score) 간의 일치 여부를 확인하기 위해 각 리뷰에 대해 두 값을 비교하였다.

비교 기준은 다음과 같다

리뷰 평점

2.5 초과: 긍정적

2.5 미만: 부정적

2.5: 중립적

감정 점수

0.2 이상: 긍정적

-0.2 이하: 부정적

-0.2 초과, 0.2 미만: 중립적

텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

▲ 결과는 위와 같았다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

▲ 위의 결과를 바탕으로 Match가 Ture인 비율을 계산해보았다.



▲ 결과는 68.75%로 많은 부분 일치한다는 것을 알 수 있었다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

▲ 불일치하는 리뷰를 알아보기 위해 필터링한 후 reviews.rating과 sentiment\_score, processed\_text를 출력해보았다.

텍스트, 스크린샷, 폰트, 메뉴이(가) 표시된 사진

자동 생성된 설명

▲ 출력은 위와 같았다. 2871번째 리뷰의 실제 reviews.rating은 2.0이었지만, sentiment\_score는 0.9456으로 긍정이었다. reviews.text는 아래와 같았다.

"I wish I had returned this wine, but only gradually used and each bottle was less than satisfying. Have ordered Vintage Wine Estates wine a few times before and very much enjoyed. O'Lerary's reserve wines were not of the same quality in my opinion and in the opinion of my book club members. Not sure if they were left in the heat somewhere along the way or just what, but just no happy with these and would not order again. Kevin presents himself as quite the wine producer and he is likely quite knowledgeable, but there is better wine available at lower cost."

("이 와인을 반품하고 싶었지만, 점차 사용되었고 각 병이 만족스럽지 못했습니다. 이전에 빈티지 와인 에스테이트 와인을 몇 번 주문했고 매우 즐겼습니다. 오리어리의 예비 와인들은 제 생각과 독서 클럽 회원들의 생각에 같은 품질이 아니었습니다. 도중에 어딘가에 더위 속에 방치되었는지 아니면 그냥 무엇인지는 확실하지 않지만, 이것들에 만족하지 않고 다시 주문하지 않을 것입니다. 케빈은 자신을 꽤 와인 생산자라고 소개하고 꽤 박식할 것 같지만, 더 저렴한 가격에 구할 수 있는 더 나은 와인이 있습니다.")

반면, processed\_text는 아래와 같았다.

wish returned wine gradually used bottle le satisfying ordered vintage wine estate wine time very much enjoyed olerarys reserve wine quality opinion opinion book club member sure left heat somewhere along way happy would order kevin present quite wine producer likely quite knowledgeable better wine available lower cost

(희망하다 반품 와인 점차적으로 사용된 병 만족 주문 빈티지 와인 에스테이트 와인 시간을 매우 즐겼습니다 olerary 예약 와인 퀄리티 의견 북클럽 회원은 분명히 어딘가에 열기를 남겨두었습니다 행복한 시간을 보내려면 kevin 생산자 선물 꽤 지식이 풍부한 더 나은 와인을 제공할 것입니다 저렴한 비용)

Wish, satisfying, enjoyed 등 긍정적 의미의 단어가 많지만 실제 리뷰를 확인해보면 그렇지 않은 경우이다. 감정 분석기가 평가의 문맥을 제대로 파악하지 못한 경우이다.

3. 평가

모델의 성능을 평가하기 위해 다양한 지표를 사용하여 분석하려 한다. 로지스틱 회귀 모델의 성능을 평가하는데 사용할 지표는 정확도(accuracy), 정밀도(precision), 재현율(recall), F1 점수(F1 score), ROC AUC 점수(ROC AUC score), 혼동 행렬(confusion matrix) 등이다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

▲ 먼저, 그리드 서치(Grid Search)를 통해 최적의 하이퍼파라미터를 찾은 로지스틱 회귀 모델을 사용하여 테스트 데이터 세트에 대한 예측을 수행하고 성능을 평가했다.

C: 규제 강도를 조절하는 하이퍼파라미터로, 값이 작을수록 강한 규제를 의미한다.

penalty: 사용되는 규제 유형 (l1 또는 l2)을 나타낸다.

solver: 최적화 알고리즘 (liblinear)을 나타낸다.

각 조합에 대해 5번의 교차 검증을 수행하여 모델을 평가했다

텍스트, 스크린샷, 메뉴, 모노크롬이(가) 표시된 사진

자동 생성된 설명

▲ 결과는 위와 같다. 대부분의 C=0.001 조합은 매우 빠르게 수행되었으며, 시간이 거의 소요되지 않았다. C=10 이상의 일부 조합은 회귀가 수렴하지 않아 Convergence Warning이 발생했다. C=100 이상의 조합도 빠르게 수행되었지만, 경고가 발생했다.

이제 그리드 서치 결과 중 최적의 하이퍼파라미터를 선택하고, 최적화된 모델로 테스트 세트에 대한 성능을 평가할 수 있다.

텍스트, 스크린샷, 폰트, 디스플레이이(가) 표시된 사진

자동 생성된 설명

▲ 위의 코드를 사용하여 최적의 하이퍼파라미터와 모델을 출력하였다.

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

▲ 교차검증을 진행한 코드와 결과이다. 모델이 5-폴드 교차 검증에서 평균적으로 72.71%의 정확도를 보였다. 이 결과를 통해 모델의 신뢰성을 평가할 수 있으며, 72.71%의 평균 정확도는 모델이 학습한 데이터에 대해 비교적 좋은 성능을 보인다고 해석할 수 있다.

텍스트, 스크린샷, 소프트웨어, 운영 체제이(가) 표시된 사진

자동 생성된 설명

▲ Accuracy(정확도)는 전체 샘플 중 올바르게 예측된 비율이다. 71.45%는 나쁘지 않은 성능이지만, 별점 예측의 도메인에 따라 다르게 평가될 수 있다.

Precision(정밀도)는 모델이 긍정 클래스로 예측한 샘플 중 실제 긍정 클래스인 샘플의 비율이다. 정밀도는 67.25%는 중간 정도의 성능을 나타낸다.

Recall(재현율)은 실제 긍정 클래스 샘플 중 모델이 올바르게 예측한 샘플의 비율로 71.45%이다. 이는 비교적 높은 값으로, 모델이 실제 긍정 클래스를 잘 찾아내는 능력이 있다는 것을 나타낸다.

F1 Score(F1 점수)는 정밀도와 재현율의 조화 평균, 모델의 전체적인 성능을 평가한다. 63.14%는 중간 정도의 성능을 나타낸다. 이는 모델이 정밀도와 재현율 사이에서 균형을 잘 맞추고 있음을 의미한다.

마지막으로 Area Under the ROC Curve(ROC-AUC)는 다중 클래스 분류에서 ROC 곡선 아래의 면적으로 1에 가까울수록 완벽한 분류를 의미한다.

일반적으로 0.7 이상이면 좋은 성능으로 간주하는데 77.43%로 좋은 성능이라고 볼 수 있다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

▲ 교차검증 정확도는 교차 검증을 통해 얻은 모델의 평균 정확도로 여러 폴드에 걸쳐 모델을 평가한 결과이다.

테스트 세트 정확도는 학습하지 않은 테스트 세트에서의 모델 정확도로 모델의 일반화 성능을 평가한다. 교차 검증 정확도와 테스트 세트 정확도가 비슷한 수준을 유지하고 있다. 이는 모델이 과적합(overfitting)되지 않았고, 새로운 데이터에서도 일관된 성능을 보임을 의미한다.

4. 시각화 및 해석

결과를 시각화 하기 위해 혼동행렬과 ROC 곡선을 사용하였다.

텍스트, 스크린샷, 폰트, 디스플레이이(가) 표시된 사진

자동 생성된 설명

▲ grid\_search.best\_estimator\_를 통해 최적의 모델을 불러온 후 X\_test 데이터를 이용하여 예측(y\_pred) 및 확률(y\_pred\_proba)을 계산하여

혼동 행렬 계산 및 시각화하였고, confusion\_matrix(y\_test, y\_pred)를 사용하여 실제 값(y\_test)과 예측 값(y\_pred)의 혼동 행렬을 생성한 후 seaborn의 heatmap 함수를 사용하여 혼동 행렬을 시각화하였다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

▲ 결과는 위와 같으며, True Positive (TP): 정확하게 예측된 5점의 개수는 388개이다.

False Positive (FP): 5점으로 예측되었지만 실제로는 다른 등급인 경우의 개수는 혼동 행렬의 첫 번째 열의 합 (24+1+0+2+0+8 = 35개)이다.

True Negative (TN): 정확하게 예측된 나머지 등급 (0~4)의 개수는 혼동 행렬의 주 대각선 값들의 합이다.

False Negative (FN): 다른 등급으로 예측되었지만 실제로는 5점인 경우의 개수는 첫 번째 행의 마지막 열의 값인 74개이다.

모델은 5점을 예측하는데 강한 성능을 보이고 있지만, 다른 등급에 대해서는 예측 성능이 부족함을 확인할 수 있다.

텍스트, 전자제품, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

▲ ROC 곡선을 시각화 하는 코드를 작성하였다.

텍스트, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

▲ 결과는 위와 같다. 그래프에서 점선은 무작위 분류기의 성능을 나타낸다. 이 곡선이 왼쪽 상단 모서리에 가까울수록 분류기의 성능이 좋음을 의미한다.

AUC 값은 ROC 곡선 아래의 면적을 나타내며, 이 값이 1에 가까울수록 분류기의 성능이 좋음을 의미한다. 현재 AUC 값은 0.9163으로 매우 높다. 이는 모델이 양성과 음성을 구분하는 능력이 뛰어남을 의미한다.

**III. 결론**

텍스트 데이터를 감성으로 분류하는 감성 분석 시스템을 개발하기 위해 와인 리뷰를 분석해 보았다.

모델은 ROC-AUC 점수 0.9163으로 높은 예측 성능을 보였지만, 혼동 행렬에서 확인한 바와 같이 데이터 불균형 문제로 인해 대부분의 리뷰를 5점으로 예측하는 경향이 있었다. 언더샘플링, 오버샘플링, SMOTE 등의 기법을 사용하거나, 다른 머신러닝 알고리즘을 시도해 이러한 데이터 불균형 문제를 해결할 수 있을 것이라 생각한다. 또한 텍스트 외에 다른 특징들을 추가하여 특징 공학을 수행하면 예측 성능을 더욱 향상시킬 수 있을 것이다.

전반적으로 모델의 예측 성능은 우수하였지만, 데이터 불균형 문제를 해결하여 모델을 개선한다면 보다 정확한 감성 분석 시스템을 구축할 수 있을 것이다.