**제품 review 데이터의 감성 분석**



|  |  |
| --- | --- |
| 과목명 | 기계학습 |
| 담당 교수님 | 백우진 교수님 |
| 학과 | 소프트웨어전공 |
| 학번 | 202020963 |
| 이름 | 유제희 |
| 제출일 | 2024/6/22 |

**목차**

1. 감성 분석의 기본 이해

2. 데이터 전처리

3. 모델 선택 및 학습

4. 평가

5. 시각화 및 해석

**감성 분석의 기본 이해**

텍스트를 분석하여 메시지의 감성이나 의견, 평가 등을 주관적으로 컴퓨터를 통해 분석하는 과정을 말한다. 이러한 감성 분석은 객관적인 인사이트 제공, 더 나은 제품 및 서비스 빌드, 대규모 분석, 실시간 결과 제공이라는 장점을 가진다.

이런 감성 분석은 여러 분야에서 활용이 된다. 예를 들어, 쇼핑몰, 영화와 같은 리뷰를 작성할 수 있는 분야에서 활용 중이다. 감성 분석을 통해 소비자의 리뷰를 분석하여 더 나은 서비스 개선 및 대응이 가능하기 때문이다.

과제에서는 쇼핑몰 리뷰에 대한 감성 분석이며, nltk-vader 모델과 logistic regression을 통해 분석을 해보았다

데이터 전처리

- url : 상품 URL 주소

- product\_name : 상품명

- reviewer\_name : 리뷰 작성자 이름

- review\_title : 리뷰 제목

- review\_text : 리뷰 내용

- review\_rating : 리뷰 별점

- verified\_purchase : 구매 확인 여부

- review\_date : 리뷰 작성 날짜

- helpful\_count : 리뷰가 도움이 된 사람의 수

- uniq\_id : 리뷰 ID

- scraped\_at : 리뷰를 가져온 시점

|  |  |
| --- | --- |
| 데이터 불러오기 |  |
| 결측치 확인 후 제거 | helpful\_count에 결측치가 많아 우선적으로 이 열만 제거 |
| 구매 여부 확인 후 구매하지 않은 사람 리뷰 삭제 |  |
| 영어 로 된 데이터만 남기고 삭제(리뷰 제목, 리뷰 내용) |  |
| 노이즈 제거 (소문자, 특수문자) |  |
| 데이터 토튼화(리뷰제목, 리뷰내용) |  |
| 불용어 함수 정의하여 불용어 제거 |  |
| 스테밍과 표제어 작업 수행 | stemming와 표제어 모두 추출 후 분석, stemming추출은 어간을 추출하는 규칙을 세워, 규칙대로 어간을 추출, 이상한 결과가 발생하는 현상 발견하였다.  표제어추출은 의미가 동일한 경우에만 단어를 얻는 결과를 보여주었다. |

**감성분석 값에 대한 토큰화(리뷰 제목, 리뷰 내용)**

|  |  |
| --- | --- |
| 모델 vader모델의 감성 초기화, 점수 컬럼 | 데이터 감성 점수 추출, 추출한 결과로 평균 내서 프레임에 추가한다.  이때 추출한 값들은 -1 ~ 1 값을 가진다. |
| 분석 결과 값이 0.01이상이면 1값 반환 그 외 0 |  |

**모델 선택 및 학습**

|  |  |
| --- | --- |
| 학습과 행렬, 그리드를 위한 라이브러리 다운로드 |  |
| 초기화 및 train, test 데이터 나누기 | X, Y 특성에 맞게 정의를 해준 후 정의를 한다. X 값은 리뷰 제목과 리뷰 내용에 대한 값으로 정의를 해줬으며, y는 평균 위에서 감성 분석을 하여 평균을 낸 값으로 정의하였다.  모델을 학습 시킬 때 최대 반복 횟수는 1000으로 정의하였다. random\_state 값은 마음대로 정의해도 상관 없다.  모델 생성 후, 하이퍼파라미터 값은 0.001 ~ 100 사이의 값으로 지정하여 그리드를 설정한다. 그리드 서치를 통한 최적의 파라미터를 탐색하는 과정 진행한다.  이후 최적의 파라미터와 최고의 cross-validation(교차 검증)의 값을 출력해 결과를 확인하였다. |
| 분류 보고서 분석 | 최적의 하이퍼파라미터 값 C:10을 기반으로 최적 모델을 bast\_model로 정의하여 최종 모델 예측을 진행한다  이렇게 학습 시킨 모델을 가지고 평가 지표와 분류 보고서를 출력한다.  정확도, 정밀도, 재현율, F1 점수, ROC-AUC 지표를 통해 모델 평가를 진행한다. 이후에는, 예측 결과를 통해 오버피팅, 언더피팅을 확인 |
|  | **1. 정확도 (Accuracy: 0.9248251748251748)**  : 전체 예측 중에서 정확하게 맞춘 비율을 의미한다.  정확도가 0.9248이라는 것은 전체 테스트 데이터 중 약 92.48%의 예측이 정확하다는 의미한다.  **2. 정밀도 (Precision: 0.9470588235294117)**  : 모델이 긍정으로 예측한 사례 중에서 실제로 긍정인 사례의 비율을 의미한다.  정밀도가 0.9471이라는 것은 모델이 긍정으로 예측한 사례의 약 94.71%가 실제로 긍정이라는 의미한다. 모델의 긍정 예측이 높음을 확인 가능하다.  **3. 재현율 (Recall: 0.9679358717434869)**  : 실제 긍정인 사례 중에서 모델이 긍정으로 정확하게 예측한 비율을 의미  재현율이 0.9679라는 것은 실제 긍정 사례의 약 96.79%를 모델이 정확히 예측했다는 의미한다.. 이는 모델이 실제 긍정을 잘 놓치지 않고 예측하는 능력이 뛰어나다는 것을 나타낸다.  **4. F1 점수 (F1 Score: 0.9573835480673935)**  : 정밀도와 재현율의 조화 평균으로, 두 지표 간의 균형을 평가한다.  F1 스코어가 0.9574라는 것은 정밀도와 재현율 간의 균형이 매우 좋음을 나타낸다.  모델이 긍정을 예측하는 데 있어 전체적으로 높은 성능을 보이고 있음을 의미한다.  **5. ROC-AUC 점수 (ROC-AUC Score: 0.7990364290224286)**  : ROC 곡선 아래의 면적을 나타내며, 0.5는 무작위 예측, 1은 완벽한 예측을 의미한다.  이 값은 모델의 분류 능력을 측정한다.  ROC-AUC가 0.7990이라는 것은 모델이 이진 분류 문제에서 약 79.90%의 경우에 대해 무작위 추측보다 더 나은 예측을 한다는 의미한다.  이는 비교적 좋은 성능을 나타내지만, 다른 지표들에 비해서는 다소 낮은 편이다.  **전체 점수 보고(Classification Report)**  이 모델은 전체적으로 높은 성능을 보이며, 특히 정밀도와 재현율에서 뛰어난 성능을 나타낸다.  다만, ROC-AUC가 약간 낮은 것은 모델이 일부 경우에서 예측에 대한 확신이 부족할 수 있음을 시사합니다.  이는 모델의 분류 임계값을 조정하거나, 추가적인 특성 엔지니어링을 통해 개선될 수 있다. |

**시각화 및 해석**

|  |  |
| --- | --- |
| 혼동행렬 시각화 |  |
| ROC에 대한 그래프 시각화 |  |

혼동 행렬은 분류 문제에서 예측 결과를 요약한 표이다. 올바른 예측과 잘못된 예측의 수를 클래스별로 나누어 요약한다.

참 긍정 (TP): 실제 클래스가 긍정(1)이고, 모델이 긍정으로 예측한 경우

참 부정 (TN): 실제 클래스가 부정(0)이고, 모델이 부정으로 예측한 경우

거짓 긍정 (FP): 실제 클래스가 부정(0)인데, 모델이 긍정으로 예측한 경우

거짓 부정 (FN): 실제 클래스가 긍정(1)인데, 모델이 부정으로 예측한 경우

참 긍정 (TP): 483 , 참 부정 (TN): 46, 거짓 긍정 (FP): 27, 거짓 부정 (FN): 16로 결과를 분석하면 예측 실패한 데이터가 있다는 것을 알 수 있다.

AUC (Area Under Curve): 값이 1에 가까울수록 모델의 성능이 우수함을 나타낸다. 분석한 모델의 AUC는 0.80로 실제로 유의미한 분류 성능을 가지고 있음을 나타낸다.