**목차**

감성 분석의 기본 이해 ……………………………………………………………………….. 2

데이터 전처리 ……………………………………………………………………………………… 3

모델 선택 및 학습 …………………………………………………………………………....... 8

모델 평가 …………………………………………………………………………………………….12

시각화 및 해석 …………………………………………………………………………………...14

결론 ……………………………………………………………………………………………………..16

**1.감성 분석의 기본 이해**

감성 분석(Sentiment Analysis) 또는 의견 마이닝(Opinion Mining)은 텍스트 데이터에서 감정이나 의견을 추출하고 분류하는 과정. 감성은 -1에서 1의 극성 (polarity)으로 측정되며, -1은 가장 부정적인 감성을, 1은 가장 긍정적인 감성을 의미 감성은 객관성 (objectivity, 0)과 주관성(subjectivity, 1)에 대해 0~1 점수로 측정한다. 감성 분석은 자연어 처리(NLP)의 한 분야로, 컴퓨터가 인간의 언어를 이해하고 해석하는 것을 목표로 한다. 이번 프로젝트에서는 이를 수행하기 위해 감성 분석기 중 하나인 VADER를 사용하도록 한다.

응용 분야

* 마케팅 및 브랜드 모니터링: 고객의 리뷰나 소셜 미디어 게시글을 분석하여 제품이나 서비스에 대한 감성을 파악하고, 브랜드의 인식을 모니터링하여 마케팅 전략 조정
* 고객 서비스: 고객의 피드백을 분석하여 불만사항을 신속하게 식별하고 해결
* 제품 개발: 사용자 리뷰를 분석하여 제품의 장단점을 파악하여, 제품을 빠르게 개선
* 정치 및 사회 연구: 정치적 연설이나 소셜 미디어 게시글을 분석하여 공공의 의견과 감정을 파악하고, 사회적 동향을 연구할 수 있다.
* 금융 분석: 뉴스 기사나 소셜 미디어 데이터를 분석하여 주식 시장의 동향을 예측하고 투자 결정을 내릴 수 있다.

VADER(Valence Aware Dictionary and sEntiment Reasoner)

소셜 미디어 텍스트를 중심으로 한 감성 분석에 특화된 도구.

작동방식

텍스트 토큰화: 입력 텍스트를 단어 또는 구로 나눈다.

감성 어휘 사전 매칭: 각 토큰을 VADER 감성 어휘 사전과 매칭하여 긍정 또는 부정 감성 점수를 할당

규칙 적용: 강조 표현, 해시태그, 인터넷 속어 등과 같은 특수 규칙을 적용하여 감성 점수를 조정

compound score 계산: 긍정 및 부정 점수를 종합하여 최종 감성 점수 (compound score) 계산

한계점: 단어 토큰을 어휘 사전과 매칭하여 작동하는 구조상 언어 표현의 다양성, 중의성, 다중성, 은유 및 비유적인 상황을 분석하는데 어려움이 있을 수 있다.

**2. 데이터 전처리**

2.1 데이터 분석

분석 데이터: wine\_review.csv

데이터 사이즈: 2891행 32열

각 열 설명

|  |  |
| --- | --- |
| 속성 명 | 설명 |
| id | 고유 식별자. |
| asins | Amazon Standard Identification Numbers, 아마존 제품 식별 번호. |
| Brand | 브랜드 이름. |
| categories | 제품이 속한 카테고리 |
| dateAdded | 데이터가 추가된 날짜 |
| dateUpdated | 데이터가 업데이트된 날짜 |
| descriptions | 제품에 대한 정보를 JSON 방식으로 가짐  필드: "dateSeen", "sourceURLs", "value" |
| dimension | 제품의 크기. |
| ean | European Article Number, 유럽 표준 제품 번호 |
| flavors | 와인의 맛 설명 |
| keys | 키워드 |
| manufacturer | 제조사 |
| manufacturerNumber | 제조사 번호 |
| name | 제품명 |
| reviews.date | 리뷰 작성 날짜 |
| reviews.dateAdded | 리뷰가 추가된 날짜 |
| reviews.dateSeen | 리뷰가 확인된 날짜 |
| reviews.didPurchase | 구매 여부 |
| reviews.doRecommend | 추천 여부 |
| reviews.id | 리뷰 고유 ID |
| reviews.numHelpful | 리뷰가 도움이 된 횟수 |
| reviews.rating | 리뷰 평점 |
| reviews.sourceURLs | 리뷰 출처 URL |
| reviews.text | 리뷰 텍스트 |
| reviews.title | 리뷰 제목 |
| reviews.userCity | 리뷰 작성자의 도시 |
| reviews.userProvince | 리뷰 작성자의 주 |
| reviews.username | 리뷰 작성자 이름 |
| sizes | 제품 크기(?) 잘 모르겠음 |
| sourceURLs | 출처 URL. |
| upc | Universal Product Code, 범용 제품 코드 |
| weight | 제품 무게 |

이 중에서 실질적으로 감성 분석할 열은 reviews.text 열로 리뷰 내용을 담고 있다.

이러한 리뷰들의 긍정 및 부정 평가와 정량적으로 연결 지을 수 있는 열로는 reviews.rating 열이 있다. 해당 열을 분석하기 위하여 빈도수를 파악한다.

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

자동 생성된 설명

reviews.rating의 평점 빈도수를 분석하였을 때 5점이 절대 다수를 차지 하고 있다.

빈도수만을 감안했을 때는 최소한 4점 미만은 모두 부정평가로 볼 여지가 있으나. 실제 리뷰를 함께 볼 경우 명확하게 부정평가를 하는 것은 대부분 1점에서 발생하고 3점만 돼도 긍정적인 평가를 내리는 리뷰어들이 존재함으로 빈도수를 기준으로 섣불리 판단하기는 어렵다.

그저 단순히 리뷰어들이 대체적으로 긍정적인 리뷰를 하는 경향이 있음을 확인하였다.

2.2 데이터 정리

텍스트, 전자제품, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

먼저 데이터를 로드하고 결측 값을 처리하며 필요 없는 열을 제거한다.

결측 치 처리 이전과 이후를 비교하여 데이터셋에 1개의 결측 치가 존재하였음을 확인할 수 있다.

2.3 텍스트 토큰화, 노이즈 제거, 정규화

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

NLTK 라이브러리를 사용하여 텍스트 데이터를 토큰화

분석에 앞서 데이터를 최적화하기 위하여 결측 값과 노이즈를 제거한다.

불용어(stop words)는 문맥상 큰 의미가 없는 단어들로, 감성에 영향을 주지 않기 때문에 제거하는 것이 좋다. 특수 문자와 숫자들 역시 감성에 직접적인 영향을 주지 않음으로 함께 제거한다.

각 토큰을 스테밍하여 단어의 어간을 추출한다. 이를 통해 텍스트를 정규화 하면 단어의 일관성을 높여 모델의 복잡성을 줄이고 학습 효율성을 높일 수 있다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명



평균적으로 절반 이상의 단어를 제거하는 효과가 있었다. 이런 식으로 노이즈를 제거하면 감성 분석의 정확도에 영향을 주지 않으면서 최적화를 달성할 수 있다.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

위와 같은 전처리 과정을 통해 텍스트가 상당히 간결 해졌음을 확인할 수 있다.

**3. 모델 선택 및 학습**

3.1 모델 선택

전처리 된 데이터로 감성분석을 수행한다. 수행 방법으로는 로지스틱 회귀를 사용한다.

로지스틱 회귀는 데이터를 0과 1로 분류하는데 있어 우수한 성능을 보임으로 긍정과 부정으로 분류하는 감성분석 상황에 적합하다. 그러나 비선형 회귀의 특성상 과적합에 유의하며 학습할 필요가 있다. 로지스틱 회귀분석한 결과를 평가하기 위해서는 평가 기준이 필요한데 이러한 평가 기준으로는 rating(평점)열을 사용하는 방법과 VADER 감성분석기의 결과를 기준으로 사용하는 방법 2가지를 사용한다.

교차 검증(cross-validation)

데이터셋을 여러 번 분할하여 모델의 성능을 평가하는 방법. 그중 가장 일반적으로 사용되는 것은 K-폴드 교차 검증(K개의 동일한 크기의 fold로 분할)

교차 검증을 하는 이유

알고리즘을 학습시키기 위해서는 학습 데이터와 이에 대한 예측 성능을 평가하기 위한 별도의 테스트 데이터가 필요하다. 그러나 이 방법은 과적합에 취약하다. 교차검증을 이용해 다양한 학습과 평가를 수행함으로써 과적합을 개선할 수 있다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명평점 기준 감성 분석

앞서 평점열을 직관적으로 분석하였을 때 2점 이하에서 부정적 리뷰가 주로 관측되고, 3점부터 긍정적 성향이 보이기 시작함으로 분류 기준을 3점으로 잡는다.

3점 이상의 평점은 긍정, 그 미만은 부정으로 간주한다.

전처리 된 데이터셋을 학습 및 테스트 세트로 분할한다.

K-폴드 교차 검증을 수행하고 로지스틱 회귀로 분류작업을 수행한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 소프트웨어, 폰트이(가) 표시된 사진

자동 생성된 설명

그리드 서치로 하이퍼파라미터 튜닝을 수행한다. 튜닝 효과를 확인하기 위하여 그리드 서치를 적용한 결과와 적용하지 않은 결과 모두를 출력한다.

VADER 기준 감성 분석

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

Compound score가 0에 가까우면 분리하기 어렵고 노이즈가 될 수 있음으로 0.05이상이면 긍정(1), -0.05 이하면 부정(0)으로 처리한 뒤 0과 1의 감성 레이블만 사용한다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명  
이 외에는 평점 기준 분석 방식과 동일하다.

**4. 평가**

교차 검증 정확도

 평점 기준

 VADER 기준

교차 검증 정확도는 평점 기준 보다 VADER 기준이 더 높은 정확도를 보인다.

이는 VADER 기분 레이블링이 리뷰 텍스트의 실제 감성을 더 잘 반영하고 있음을 나타낸다.

평점은 사람마다 기준이 다를 수 있음으로 평점과 리뷰를 1 대 1 매칭하는 것은 부정확할 수 있으며 긍정 부정 기준도 임의로 지정한 것이기 때문에 아주 높은 정확도를 보이기 어려워 보인다.

텍스트, 스크린샷, 폰트, 디자인이(가) 표시된 사진

자동 생성된 설명하이퍼파라미터 미 최적화텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명 하이퍼파라미터 최적화 후

최적화 성능차이

앞선 코드에서 하이퍼파라미터 최적화의 효과를 보기 위하여 양쪽 상황 모두를 볼 수 있도록 하였다. 평점을 기준으로 한 감성 분석의 경우 성능상 양쪽의 차이가 극히 미미하여 ROC-AUC를 제외하면 수치상 드러나는 차이가 전혀 없었다. (때문에 최적화 이후는 ROC-AUC만 보고서에 포함)

성능 평가

부정 리뷰에 대해 precision이 0.71이고 recall이 0.15 인 동안 긍정 리뷰에 대해 precision이 0.82이고 recall이 0.98 이라는 것은 모델이 부정보다 긍정으로 분류하려는 경향이 매우 강하다는 것을 의미한다. 이는 회기 분석 자체의 분제보다 3점 이상 긍정, 미만 부정으로 분류한 현재 분류 기준이 한쪽으로 치우쳐졌을 가능성이 높다.

F1-score의 경우 부정 0.25, 긍정 0.90으로 부정 분류에 있어서 성능이 매우 낮음을 알 수 있다.

Accuracy는 0.82 로 좋다고 하기는 부족한 수준.

Confusion Matrix와 ROC-AUC는 시각화에서 단원에서 마저 설명한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

최적화 성능차이

VADER를 기준으로 한 경우에는 최적화 이후에 상당한 개선이 있었다. 최적화 이전에는 특히 부정 분류가 정상적으로 이루어지지 못하였는데, 최적화 이후에는 수치가 효과적으로 개선됐다.

성능 평가 (최적화 이후를 기준)

평점 기준 분류일 때만큼은 아니지만 마찬가지로 모델이 부정보다 긍정 평가하려는 경향이 강하다. 이는 VADER가 사용하는 감성 어휘 사전에 편향이 있거나, 긍정 리뷰에 비해 부정 리뷰가 대체적으로 모호하기 때문일 수도 있다.

**5. 시각화 및 해석**

평점 기준 감성 분석 VADER 기준 감성 분석

텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명텍스트, 라인, 스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명

ROC 그래프를 보았을 때 ROC 평점 기준 방식에서는 파라미터 최적화로 인한 성능 변화가 거의 미미한 반면 VADER 기준 방식에서는 최적화로 인한 성능 향상이 꽤 있었음을 확인할 수 있다.

ROC 수치는 VADER 기준 방식에서 살짝 더 높아 그만큼 성능이 더 뛰어남을 알 수 있다.

평점 기준

텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

자동 생성된 설명’ 텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

자동 생성된 설명

VADER 기준 감성 분석

텍스트, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명

혼동행렬표는 TRUE, FALSE, POSITIVE, NEGATIVE 관계를 나타내는 도표로, 요소 간의 관계를 이용하여 성능지표를 만들어낸다. 성능 지표에 관해서는 앞서 언급하였음으로 여기서는 간단히 정리한다.

1. 평점 기준 모델 (기본 모델)

17개의 실제 부정 리뷰를 정확히 부정으로 예측하였음

97개의 실제 부정 리뷰를 긍정으로 잘못 예측하였음

7개의 실제 긍정 리뷰를 부정으로 잘못 예측하였음.

457개의 실제 긍정 리뷰를 정확히 긍정으로 예측하였음

2. 평점 기준 모델 (튜닝 된 모델)

기본 모델과 동일

3. VADER 기준 모델 (기본 모델)

모든 부정 리뷰를 긍정으로 잘못 예측

모든 긍정 리뷰를 정확히 긍정으로 예측

4. VADER 기준 모델 (튜닝 된 모델)

20개의 실제 부정 리뷰를 정확히 부정으로 예측

43개의 실제 부정 리뷰를 긍정으로 잘못 예측

1개의 실제 긍정 리뷰를 부정으로 잘못 예측

462개의 실제 긍정 리뷰를 정확히 긍정으로 예측

**결론**

정리하자면 하이퍼파라미터 튜닝에 있어서 평점 기준 모델은 튜닝 후에도 성능 향상이 없으며, 특히 부정 리뷰에 대한 예측 성능이 매우 낮다. VADER 기준 모델은 튜닝 후 성능이 향상되었으며, 특히 부정 리뷰에 대한 예측 성능이 크게 개선되었다.

리뷰어의 평가점수와 실제 리뷰 사이에는 괴리가 있는 경우가 많아 평가점수보다 VAPER의 감성 분석 기능을 사용하는 것이 실제로 높은 정확도를 가진다는 것을 알 수 있다.