**‘wine\_review.csv’ 데이터 감성분석 보고서**



|  |  |
| --- | --- |
| **과목명** | 기계학습 |
| **학과** | 컴퓨터공학전공 |
| **학번** | 201921092 |
| **이름** | 황준석 |
| **제출일자** | 2024.06.21 |

**감성 분석의 기본 이해**

감성 분석(Sentiment Analysis)은 텍스트 데이터를 분석하여 그 감정을 긍정적, 부정적, 중립적으로 분류하는 자연어 처리(NLP) 기술이다. 감성 분석은 소셜 미디어 모니터링, 고객 피드백 분석, 영화 리뷰 분석 등 다양한 응용 분야에서 사용됩니다. 이 프로젝트에서는 와인 리뷰 데이터를 사용하여 리뷰를 긍정적 또는 부정적으로 분류하는 감성 분석 모델을 구축한다.

**데이터 전처리**

1. 데이터 로드 및 구조 파악

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

먼저 wine\_review.csv 파일을 로드하고 데이터의 구조를 파악합니다.

데이터셋에는 총 2890개의 리뷰가 포함되어 있으며, 주요 속성으로는 'reviews.text'(리뷰 텍스트)와 'reviews.rating'(리뷰 평점)이 있습니다.

데이터셋의 처음 몇 개의 행을 출력하여 데이터셋의 기본 정보가 다음과 같음을 알 수 있습니다.

1. id: 각 리뷰의 고유 식별자

2. asins: 제품의 ASIN (Amazon Standard Identification Number)

3. brand: 와인의 브랜드 이름

4. categories: 와인이 속한 카테고리 정보

5. dateAdded: 리뷰가 데이터베이스에 추가된 날짜

6. dateUpdated: 리뷰가 데이터베이스에 업데이트된 날짜

7. descriptions: 와인의 설명

8. dimension: 와인의 치수 정보

…

1. 필요한 컬럼 선택 및 결측값 처리

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

리뷰 텍스트(reviews.text)와 리뷰 평점(reviews.rating) 컬럼을 선택하고, 결측값을 제거합니다.

1. 텍스트 전처리

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명텍스트 데이터를 분석하기 위해 NLTK를 사용하여 전처리 작업을 수행합니다. 텍스트 전처리에는 다음 단계가 포함됩니다:

소문자로 변환

구두점 제거

토큰화

불용어 제거

표제어 추출

1. 감성 레이블 생성

스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명리뷰 평점을 기준으로 감성 레이블을 생성합니다. 평점이 3 이상이면 긍정(1), 그렇지 않으면 부정(0)으로 분류합니다.

1. 텍스트 데이터를 TF-IDF 벡터화

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

TF-IDF(Vectorizer)를 사용하여 텍스트 데이터를 벡터화합니다. 이는 각 단어의 중요도를 계산하여 텍스트 데이터를 숫자로 변환하는 과정입니다.

1. 데이터셋 분할

스크린샷, 텍스트, 폰트이(가) 표시된 사진

자동 생성된 설명

데이터셋을 학습 세트와 테스트 세트로 분할합니다.학습 세트는 모델을 학습하는 데 사용하고, 테스트 세트는 모델의 성능을 평가하는 데 사용합니다.

**모델 선택 및 학습**

1. 로지스틱 회귀 모델 학습

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

로지스틱 회귀(Logistic Regression) 모델을 사용하여 학습 세트를 기반으로 모델을 학습합니다.

**평가**

1. 예측 및 평가 지표 계산

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

자동 생성된 설명테스트 세트를 사용하여 모델이 감성을 예측하도록 하고, 모델의 성능을 다양한 지표로 평가합니다.

주요 결과

정확도 (Accuracy), 정밀도 (Precision): 96.1%로, 모델이 대부분의 리뷰를 올바르게 분류, 긍정 리뷰를 정확하게 예측했습니다.

재현율 (Recall): 100%로, 모든 긍정 리뷰를 정확하게 예측했습니다.

F1 스코어 (F1 Score): 98%로, 정밀도와 재현율 모두에서 높은 성능을 보였습니다.

ROC AUC (Positive Reviews): 91.6%로, 긍정 리뷰에 대한 예측 성능이 매우 우수했습니다.

**시각화 및 해석**

혼동 행렬 및 ROC 곡선 시각화

텍스트, 스크린샷, 소프트웨어, 디스플레이이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 도표, 소프트웨어이(가) 표시된 사진

자동 생성된 설명혼동 행렬과 ROC 곡선을 시각화하여 모델의 성능을 직관적으로 이해합니다.

혼동 행렬 해석:

긍정 리뷰를 잘 예측하지만, 부정 리뷰는 대부분 잘못 예측했습니다.

True Positive (470개), False Positive (19개), True Negative (0개), False Negative (0개)로 모델이 긍정 리뷰에 편향됨을 알 수 있습니다.

ROC 곡선 해석:

긍정 리뷰 (AUC = 0.91): 모델이 긍정 리뷰를 예측하는 데 있어 매우 좋은 성능을 보였습니다.

부정 리뷰 (AUC = 0.91): 모델이 부정 리뷰를 예측하는 데 있어 성능이 좋음을 보여줍니다. 하지만 혼동 행렬을 보면 모델이 부정 리뷰를 잘못 예측하는 경향이 있음을 알 수 있습니다.

결론

부정 리뷰에 대한 ROC AUC 값이 0.91인 것은 모델이 부정 리뷰를 예측하는 데 있어 좋은 성능을 보였음을 의미합니다. 그러나 혼동 행렬에서 모델이 부정 리뷰를 잘 예측하지 못한 점을 고려했을 때, 예측 성능을 향상시키기 위해 클래스 불균형 문제를 해결하고, 모델을 개선해야 할 필요성이 보입니다.