**제품 review 데이터의 감성 분석 보고서**

**wine\_review.csv 분석**

201921042 전승호

**목 차**

1. **감성 분석의 기본 이해**
2. **데이터 전처리**
3. **모델 선택 및 학습**
4. **평가**
5. **시각화 및 해석**

**1. 감성 분석의 기본 이해**

감성 분석(Sentiment Analysis)은 ‘오피니언 마이닝(Opinion Mining)’으로도 불립니다. 텍스트에 나타난 사람들의 태도, 의견, 성향과 같은 주관적인 데이터를 분석하는 자연어 처리 기술입니다. 분석한 텍스트를 긍정, 부정, 중립의 감정으로 분류하고 이는 고객 리뷰, 소셜 미디어, 시장 조사 등 다양한 분야에 응용될 수 있습니다. 예시로 기업에서 고객 서비스를 개선하기 위해 사용하는 경우 고객의 피드백을 받은 후 감성 분석을 통해 긍정인지 부정인지 판단합니다. 또한 고객이 무엇에 해당 감정을 표출하는지 연관지어 식별할 수 있어 기업이 피드백을 개선하기 용이합니다.

**2. 데이터 전처리** (wine\_review.csv 파일 분석)

1) 필요한 라이브러리 import, NLTK 불용어 리스트 다운로드

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

2) 데이터 파일 로드 후 확인

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

(info() 실행 결과) 32개의 열이 있습니다.

텍스트, 스크린샷, 메뉴이(가) 표시된 사진

자동 생성된 설명

3) reviews.text 열만 남기고 나머지 삭제, 해당 열에서 결측값이 있는 행 제거

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

4) 텍스트 토큰화 및 불용어 제거

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

전처리 함수는 숫자와 특수 문자, 구두점을 제거하고, 텍스트를 단어 단위로 분리합니다. 또한 불용어 리스트(stop\_words)에 포함된 단어를 제거합니다.

5) 텍스트 정규화 (표제어 추출)  
텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

**3. 모델 선택 및 학습**

1) 모델 구현

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

VADER 감성 분석기로 텍스트의 감성 점수를 계산하여 compound를 반환합니다.

2) 모델 학습 및 튜닝

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

감성 점수가 0 이상이면 긍정적(1)으로, 0미만이면 부정적(0)으로 변환해 새로운 열 sentiment에 저장합니다.

데이터셋을 분리해 20%를 테스트 세트, 80%를 학습 세트로 사용합니다.

**4. 평가**

1) 코드

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

학습 세트와 테스트 세트에 대해 예측 값을 생성하고, 실제 값과 비교해 학습 세트와 테스트 세트의 정확도를 계산합니다. 또한 실제 값과 예측 값의 상관 관계를 보여주는 confusion matrix를 출력합니다.

2) 보고서

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

학습 데이터와 테스트 데이터 모두 높은 정확도를 보이므로 둘 다 모델이 잘 작동한다는 것을 알 수 있습니다.

클래스 0보다 클래스 1의 정밀도와 재현율이 높습니다. 이것은 모델이 긍정적인 리뷰를 잘 예측하는 것을 의미합니다.

**5. 시각화 및 해석**

1) 혼동 행렬 시각화

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

자동 생성된 설명

실제로 Positive이고 모델이 Positive로 예측(True Positive) 🡺 450명

실제로 Positive이지만 모델이 Negative로 잘못 예측(False Negative) 🡺 17명

실제로 Negative이고 모델이 Negative로 예측(True Negative) 🡺 31명

실제로 Negative이지만 모델이 Positive로 예측(False Positive) 🡺 12명

TP = 450 / FN = 17 / TN = 31 / FP = 12

**정확도 (Accuracy):** 전체 샘플 중에서 맞게 예측한 샘플의 비율

(TP+TN) / (TP+TN+FP+FN) = 481 / 510 = **약 0.94**

**정밀도 (Precision):** 긍정으로 예측한 샘플 중에서 실제로 긍정인 샘플의 비율

TP / (TP+FP) = 450 / 462 = **약 0.97**

**재현율 (Recall):** 실제 긍정 샘플 중에서 모델이 긍정으로 정확히 예측한 비율

TP / (TP+FN) = 450 / 467 = **약 0.96**

**F1 점수 (F1-Score):** 정밀도와 재현율의 조화 평균

2 \* (정밀도 \* 재현율) / (정밀도 + 재현율) = 2 \* 0.9312 / 1.93 = **약 0.96**

**특이도 (Specificity):** 실제 부정 샘플 중에서 모델이 부정으로 정확히 예측한 비율

TN / (TN+FP) = 31 / 43 = **약 0.72**

이 모델은 재현율이 0.96으로 매우 높은 편이지만, 특이도가 0.72로 다소 낮습니다. 이를 통해 이 모델이 긍정 클래스를 매우 잘 예측하는 반면, 부정 클래스에서는 성능이 떨어질 수 있음을 알 수 있습니다.

2) ROC 곡선 및 AUC

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 라인, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

AUC (Area Under the Curve): ROC 곡선 아래 면적이며, 0과 1사이의 값. AUC가 클수록 모델의 성능이 좋음을 의미

이 모델은 AUC가 0.89이고, ROC 곡선이 좌상단에 가깝게 위치하기 때문에 좋은 모델이라고 할 수 있습니다.