기계학습(8585) 기말고사 대체 과제(보고서)

201920948 신주현

**[데이터 전처리]**

**<데이터 로드 및 구조 확인>**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

데이터를 로드한 후, df.info()를 통해 데이터프레임의 기본적인 정보를 확인함. 이를 통해 각 열(column)에 대한 정보와 결측치 여부를 파악할 수 있었음.

**<결측치 처리 및 중복 제거>**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

리뷰 제목의 결측값을 빈 문자열로 대체하고, 리뷰 텍스트의 결측값을 제거함.

또한 중복된 리뷰를 제거하여 데이터의 중복성을 제거함.

**<텍스트 데이터 전처리>**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

텍스트 데이터를 분석하기 위해 정규 표현식을 사용하여 전처리 진행. (숫자, 공백, 특수 문자를 제거, 소문자 변환, 토큰화, 불용어 제거, 어간 추출)

**<감정 분석>**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

VADER를 사용하여 리뷰 텍스트와 제목의 감성을 분석하고, 리뷰 텍스트와 제목에 대해 각각 감성 점수를 계산한 후, 이를 종합하여 레이블을 생성

**<입력 데이터 생성>**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

리뷰 제목과 텍스트를 합쳐서 입력 데이터 생성.

**[모델구현]**

**<데이터 분할>**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

학습 데이터(70%)와 테스트 데이터(30%)로 분할하여 모델 학습 및 평가에 사용.

**<파이프라인 및 모델 학습>**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

**텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명**

TF-IDF 벡터화와 로지스틱 회귀를 사용하여 파이프라인을 구성하고, 그리드 서치를 통해 최적의 하이퍼파라미터를 찾았음. 이를 기반으로 모델 학습 진행.

**[결과 및 분석]**

**<모델 평가>**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

**[학습 데이터 평가 결과]**

정확도는 0.90980로 90.98%의 정확도로 예측할 수 있음.

class 0에서는 비교적 낮은 재현율과 낮은 f1 Score을 가지고 있다.

class 1에서는 높은 정밀도와 재현율, f1 Score을 가지고 있다.

이를 통해 긍정적인 반응(class 1)에 대해서는 잘 예측하고 있으나, 부정적인 반응(class 2)에 대해서는 모델 개선이 필요하다는 점을 알 수 있음.

**[ROC-AUC 평가 결과]**

ROC-AUC는 10에 가까울 수록 분류 성능이 높다는 것을 의미한다. 따라서 약 0.8859의 결과는 높은 분류 성능을 가진다는 것을 알 수 있다.

**<혼합 행렬>**

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

**[혼합 행렬 분석]**

[0][0] / True Negative (TN): 19 / 모델이 Actual Negative를 Negative로 정확하게 예측한 경우.

[0][1] / False Positive (FP): 65 / 모델이 Actual Negative를 Positive로 잘못 예측한 경우.

[1][0] / False Negative (FN): 4 / 모델이 Actual Positive을 Negative로 잘못 예측한 경우.

[1][1] / True Positive (TP): 677 / 모델이 Actual Positive을 Positive로 정확하게 예측한 경우.

혼합 행렬을 통해 모델이 Negative를 Positive로 잘못 예측하는 비율이 비교적 높다는 것을 확인할 수 있었음.

**<ROC curve 및 AUC>**

텍스트, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

**[ROC 곡선 분석]**

ROC 곡선은 모델의 분류 능력을 시각적으로 평가하는 도구이며 곡선 아래의 면적(AUC)이 1에 가까울수록, 좌상단에 가까울수록 모델의 성능이 우수함을 나타냄. 본 모델의 경우, AUC 값은 0.89로 나왔으며, 이는 모델이 양호한 성능을 보인다는 것을 의미함.

**<결과 요약 및 분석>**

정확도(Accuracy): 0.91

정밀도(Precision), 재현율(Recall), F1 점수(F1 Score)는 모두 양호한 수준을 보였음.

혼합 행렬에서 Negative를 Positive로 잘못 예측한 비율(FP)이 비교적 높았으나, 전반적인 성능은 양호함.

ROC-AUC 점수: 0.89로, 모델의 분류 능력이 우수함.

**<결론>**

본 과제에서는 와인 리뷰 데이터를 기반으로 텍스트 감성 분석을 수행하고, 로지스틱 회귀 모델을 통해 리뷰의 감성 분류를 진행함. 데이터 전처리, 모델 학습 및 평가 과정을 거쳐 모델의 성능을 종합적으로 분석하였음.

최종적으로 본 모델의 경우, AUC 값이 0.89로 좋은 성능의 모델임을 알 수 있었음.