|  |
| --- |
| 기계학습 기말고사 |
| Wine review 데이터의 감성 분석 |

|  |
| --- |
| 201921044 전태윤  2024-6-21 |

* 감성 분석의 기본 이해

감성 분석(Sentiment Analysis)은 텍스트 데이터에서 감정, 감성, 또는 주관적인 의견을 추출하거나 이를 분류하는 자연어 처리 기술입니다. 이 기술은 특정 텍스트가 나타내는 감정이 긍정적인지, 부정적인지, 중립적인지를 판단하거나, 보다 구체적으로는 감정의 강도나 종류를 분석하는 데 사용됩니다.

감성 분석의 주요 기능과 응용

감정 분류: 주어진 텍스트가 긍정적인지, 부정적인지, 중립적인지 분류합니다. 예를 들어, "이 제품은 좋다"와 "이 제품은 나쁘다"라는 두 문장을 각각 긍정과 부정으로 분류할 수 있습니다.

감정 강도 분석: 텍스트에서 나타나는 감정의 강도를 평가합니다. 예를 들어, "이 영화는 정말 멋지다!"와 "이 영화는 그저 그렇다."에서 각각 '정말 멋지다'의 감정이 강하고, '그저 그렇다'의 감정이 약하다고 분석할 수 있습니다.

주관적인 의견 분석: 텍스트에서 나타나는 주관적인 의견을 추출하거나 분류합니다. 예를 들어, 소셜 미디어에서 사용자들의 제품 리뷰나 정치적인 발언을 분석하는 데 사용될 수 있습니다.

감성 분석의 응용 사례

제품 리뷰 분석: 제품에 대한 리뷰를 분석하여 소비자의 반응을 이해하고 제품 개선에 활용할 수 있습니다.

SNS 분석: 특정 주제에 대한 공개적인 의견을 분석하여 대중의 감정 상태를 파악하거나 정책 변화에 대한 반응을 모니터링할 수 있습니다.

* 데이터 전처리
  + 결측값 제거 및 컬럼 선택

감성 분석에서 필요한 컬럼인 'reviews.text'만을 선택하여 데이터프레임을 구성합니다. 이후, 결측값을 확인하고, 있으면 해당 행을 제거하여 데이터를 정리합니다.

* + 텍스트 토큰화 및 정규화

텍스트 토큰화는 주어진 문장을 단어 단위로 분할하는 과정을 말합니다. NLTK 라이브러리를 사용하여 텍스트를 소문자로 변환하고, 단어 단위로 토큰화한 후 불용어를 제거합니다. 불용어는 문장의 의미를 이해하는 데 기여하지 않는 단어들로, 예를 들어 'a', 'the', 'is'와 같은 단어들이 포함됩니다.

* + 텍스트 정규화 (표제어 추출)

텍스트 정규화는 텍스트에서 단어의 원형을 찾아내는 과정입니다. 이 과정은 단어들을 문법적으로나 의미적으로 관련된 단일 형태(표제어)로 변환하여 처리의 용이성을 높이는데 도움을 줍니다.

* 모델 선택 및 학습
  + 모델 구현

NLTK 패키지의 VADER 감정 분석기를 사용합니다. VADER는 감성 분석을 위한 사전 기반의 감정 분석 도구로, 텍스트의 각 단어와 문맥을 고려하여 감정 점수를 계산합니다.

감정 점수는 -1에서 1 사이의 값으로, 값이 클수록 긍정을, 값이 작을수록 부정을 의미합니다.

각각의 리뷰를 입력으로 받아 감정 점수를 계산하며, 리뷰에 부정적인 내용이나 긍정적인 내용이 없는 경우에는 감성 점수로 0을 반환합니다.

그렇지 않은 경우에는 리뷰의 감정 점수를 계산하고 데이터프레임의 “Sentiment” 열에 저장됩니다.

감성 점수가 0 이상이면 긍정(1), 그렇지 않으면 부정(0)으로 라벨링합니다.

* + 데이터 분할

데이터의 80%는 train, 20%는 test로 분할합니다.

* + random\_state=42를 지정하여 재현성을 보장합니다.
  + 파이프라인 구성

TfidfVectorizer:

텍스트 데이터를 TF-IDF(Term Frequency-Inverse Document Frequency) 벡터로 변환합니다.

단어의 빈도와 중요도를 반영하여 수치 벡터를 생성합니다.

Logistic Regression:

벡터화된 데이터를 입력으로 받아 감성 분류를 수행하는 로지스틱 회귀 모델을 사용합니다.

solver='liblinear'를 사용하여 작은 데이터셋에 적합한 solver를 지정합니다.

* + 하이퍼파라미터 그리드 설정

ngram\_range: 단어의 조합을 하나로 할지(1, 1), 두 개까지 할지(1, 2) 설정합니다.

max\_df: 너무 빈번하게 등장하는 단어를 무시할 기준을 설정합니다.

C: 로지스틱 회귀 모델의 규제 강도를 조절합니다. 값이 클수록 규제가 약해지고, 작을수록 규제가 강해집니다.

* + 그리드 서치 적용

설정한 하이퍼파라미터 그리드를 사용하여 최적의 모델을 찾습니다.

교차 검증(cv=5)을 통해 데이터를 5등분하여 각각의 부분을 테스트 셋으로 사용하면서 훈련을 반복합니다.

모델의 성능을 accuracy로 평가합니다.

* + 최적 하이퍼파라미터 출력

최적의 하이퍼파라미터 조합을 출력합니다.

이 조합은 그리드 서치를 통해 가장 높은 정확도를 기록한 파라미터입니다.

* 평가
  + 정확도 (Accuracy): 약 92.56%

정확도는 모델이 전체 샘플 중 올바르게 분류한 비율을 나타냅니다. 이 모델은 약 92.56%의 정확도를 보입니다.

* + 정밀도 (Precision): 약 92.86%

정밀도는 모델이 긍정 클래스로 예측한 것 중 실제로 긍정인 비율을 나타냅니다. 이 모델은 약 92.86%의 정밀도를 가집니다.

* + 재현율 (Recall) 또는 민감도 (Sensitivity): 약 99.43%

재현율은 실제 긍정 중 모델이 올바르게 예측한 긍정의 비율을 나타냅니다. 이 모델은 약 99.43%의 재현율을 보입니다.

* + F1 점수 (F1 Score): 약 96.03%

F1 점수는 정밀도와 재현율의 조화 평균으로, 모델의 성능을 종합적으로 평가합니다. 이 모델은 약 96.03%의 F1 점수를 가집니다.

* + ROC-AUC 점수 (ROC-AUC Score): 약 0.88

ROC-AUC 점수는 ROC 곡선 아래 영역으로, 모델의 분류 성능을 전반적으로 나타냅니다. 이 모델은 약 0.88의 ROC-AUC 점수를 가집니다.

* + 평가 결과 해석:
  + 정확도가 92.56%로 비교적 높지만, 데이터셋의 클래스 불균형이 있을 경우에는 정확도만으로 모델의 성능을 판단하기 어렵습니다.
  + 정밀도와 재현율 모두 높으므로, 긍정과 부정 모두에 대해 모델이 균형 잡힌 예측을 수행하고 있습니다.
  + F1 점수가 높고, ROC-AUC 점수도 0.88로 양호한 수치를 보이므로, 이 모델이 감성 분석을 수행하는 데 효과적이라고 평가할 수 있습니다.
* 시각화 및 해석

Confusion Matrix

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

Confusion Matrix에서 알 수 있는 결과는 다음과 같습니다.

TP: 520, TN:15, FP:40, FN:3

* + 위의 결과를 종합해 본다면, 이 감성 분석 모델은 TN이 매우 낮고 FP가 상대적으로 높습니다. TN이 15, FP가 40이라는 것은 부정적인 감정을 예측하는 데 문제가 있다는 것을 나타냅니다. 부정적인 감정을 잘 예측하지 못하고, 오히려 많은 경우를 긍정으로 잘못 예측하고 있습니다.
  + TP (520)와 FN (3)을 보면 긍정적인 예측은 매우 잘하고 있지만, 부정적인 예측에서는 상당한 오차가 있습니다.

이러한 경우 데이터가 긍정적인 레이블(1)로 편향되어 있거나, 모델이 긍정적인 경우를 더 잘 예측하도록 학습된 것을 의미할 수 있습니다.

True Negative Rate를 계산해 본 결과 약 0.273입니다. 이것은 부정적인 감정을 잘 예측하고 있지 못하는 것을 의미합니다.

ROC Curve

텍스트, 라인, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

ROC Curve를 통해 알 수 있는 결과는 다음과 같습니다.

1. ROC Curve의 대각선은 랜덤으로 분류하는 모델의 결과를 나타냅니다. 학습한 모델은 대각선 위에 Curve가 존재하므로 랜덤으로 분류하는 모델보다 효과적으로 감성 분석을 하고 있음을 알 수 있습니다.
2. ROC Curve의 아래 면적인 AUC는 모델의 전반적인 성능을 나타냅니다. AUC=0.88으로 1에 가깝다고 볼 수 있으며 이는 모델이 효과적으로 분류하고 있음을 나타냅니다.

* Overfitting과 Underfitting

Overfitting의 주요 특징은 훈련 데이터에 대한 예측력은 높지만, 일반화 성능이 낮아지는 것입니다. 이 모델의 경우, 긍정적인 감정에 대한 예측이 매우 좋지만, 부정적인 감정에 대한 예측이 매우 나쁩니다. 이는 모델이 긍정적인 예제에 너무 맞춰져 있어서 부정적인 예제에 대한 일반화가 부족하다는 것을 의미할 수 있습니다.

따라서, 이 모델은 부정적인 감정에 대해 Overfitting 문제가 발생했다고 볼 수 있습니다. Overfitting을 해결하기 위해서는 다음과 같은 접근 방법을 고려할 수 있습니다:

* + 데이터의 다양성 확보: 부정적인 감정에 대한 데이터를 추가적으로 수집하거나 다양한 데이터로 확장합니다.
  + 모델 복잡도 조정: 모델의 복잡도를 줄이거나, 규제를 강화하여 Overfitting을 방지합니다.
  + 교차 검증 사용: 모델의 일반화 성능을 더 정확히 평가하기 위해 교차 검증을 사용합니다.
  + 하이퍼파라미터 튜닝: 모델의 최적화를 위해 하이퍼파라미터를 조정하고, 테스트 데이터나 검증 데이터에 대한 성능을 평가합니다.
* 모델 분석 결과 요약
  + 클래스 불균형 문제: 데이터가 긍정적인 감정에 편향되어 있을 가능성이 높습니다.
  + 부정적인 감정 예측의 성능 부진: TN이 낮고 FP가 높은 것은 부정적인 감정을 잘 예측하지 못하고 있음을 의미합니다. 이는 부정적인 감정 예측에 대해 Overfitting 문제가 발생했다고 볼 수 있습니다.
  + 그럼에도 불구하고 AUC=0.88으로 랜덤으로 분류하는 경우보다 좋은 분류 결과를 나타냅니다.