

**기말고사 대체과제**

기계학습

201920915 박영빈 | 2024.06.20

**목차**

1. 감정 분석의 기본 이해
2. 데이터 전처리
3. 모델 선택 및 학습
4. 평가
5. 시각화 및 결론
6. **감정 분석의 기본 이해**

감정 분석(Sentiment Analysis)은 텍스트 데이터를 분석하여 그 감정 또는 의견을 긍정적, 부정적, 중립적 등으로 분류하는 자연어 처리(NLP) 기술입니다. 이는 주로 소셜 미디어, 리뷰, 고객 피드백 등에서 사람들의 감정을 파악하는 데 사용됩니다. 감성 분석은 텍스트 마이닝(Text Mining)과 머신 러닝(Machine Learning)을 결합하여 감정의 뉘앙스를 파악합니다.

**-감정 분석의 활용**

1. 고객 피드백 분석: 기업은 고객의 리뷰나 피드백을 분석하여 제품 또는 서비스에 대한 고객의 감정을 이해하고, 이를 통해 제품 개선이나 마케팅 전략을 조정할 수 있습니다.
2. 소셜 미디어 모니터링: 소셜 미디어 플랫폼에서 브랜드나 제품에 대한 감정을 모니터링하여, 고객의 의견을 실시간으로 파악하고 대응할 수 있습니다.
3. 여론 분석: 정치나 사회적 이슈에 대한 여론을 파악하여, 특정 주제에 대한 대중의 감정을 이해하고, 이를 통해 정책 결정을 지원할 수 있습니다.
4. 영화 및 제품 리뷰 분석: 영화나 제품에 대한 리뷰를 분석하여, 일반적인 평판을 파악하고, 향후 제품 개발에 반영할 수 있습니다.

**-감성 분석의 주요 단계**

1. **텍스트 전처리**: 데이터 전처리 과정에서는 노이즈 제거, 결측값 처리, 텍스트 정규화(토큰화, 불용어 제거, 스테밍 또는 표제어 추출) 등을 수행하여 텍스트를 분석에 적합한 형태로 변환합니다.
2. **특징 추출**: TF-IDF(Term Frequency-Inverse Document Frequency) 같은 방법을 사용하여 텍스트 데이터를 수치화된 특징 벡터로 변환합니다.
3. **모델 학습**: 다양한 머신 러닝 알고리즘(SVM, Naive Bayes, 딥러닝 등)을 사용하여 모델을 학습시킵니다.
4. **모델 평가**: 학습된 모델을 평가 지표(정확도, 정밀도, 재현율, F1 점수, ROC-AUC 등)를 사용하여 성능을 평가합니다. 이 과정에서 모델의 강점과 약점을 분석합니다.
5. **데이터 전처리**
6. amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv 1

1개의 열(변수)과 6823개의 행

1. 감정 분석에 사용하지 않을 변수를 제거

제거된 변수: url, reviewer\_name, review\_date, uniq\_id, scraped\_at, verified\_purchase

남은 변수: product\_name, review\_title, review\_text, review\_rating, helpful\_count

1. 결측치를 확인 후 처리

review\_title, review\_text에 결측치가 있는 경우 행 자체를 제거

helpful\_count에 결측치는 0일 때이므로 변경해서 처리

1. 중복된 데이터가 존재하는지 확인 후 처리

중복된 데이터 없음

1. review\_title, review\_text 변수의 텍스트 데이터를 토큰화하고 기호와 불용어 제거한 후 표제어를 추출하여 review 변수로 통합

nltk 데이터를 다운로드 해 위 과정에 사용

1. review\_rating 변수의 값을 확인 후 이것을 학습 라벨로 사용할 수 있게 변경

중앙값 기준으로 중앙값 이상 긍정과 이하 부정 라벨로 sentiment 변수에 저장

1. 데이터셋을 학습용, 테스트용으로 분할

X\_train, X\_test, y\_train, y\_test = train\_test\_split(df["review"], df["sentiment"], test\_size=0.2, random\_state=42)

1. 텍스트를 TF-IDF(Term Frequency-Inverse Document Frequency) 백터화

학습에 사용할 수 있게 백터화를 진행

vectorizer = TfidfVectorizer(max\_features=2000)

X\_train\_tfidf = vectorizer.fit\_transform(X\_train)

X\_test\_tfidf = vectorizer.transform(X\_test)

1. **모델 선택 및 학습**

**선택한 모델: SVC (Support Vector Classifier)**

서포트 벡터 머신(SVM, Support Vector Machine)은 지도 학습 기반의 분류 및 회귀 분석 방법입니다. SVM은 데이터 포인트를 분류하는 초평면(hyperplane)을 찾는 데 중점을 둡니다. SVC (Support Vector Classifier)는 SVM을 사용하여 분류 작업을 수행하는 모델입니다.

**-강점**

1. SVC는 고차원 데이터셋에서도 매우 효율적으로 작동합니다.

2. SVC는 클래스 간의 마진을 최대화하여 데이터 포인트를 분류합니다. 이는 일반화 성능을 향상시켜 과적합을 방지하는 데 도움이 됩니다.

3. SVC는 다양한 분류 문제에서 견고하고 높은 성능을 보입니다.

4. 노이즈에 민감하지 않고, 데이터의 이상치(outliers)에 대해 상대적으로 영향을 덜 받습니다.

**-단점**

1. SVC는 큰 데이터셋에서 계산 비용이 많이 들기 때문에 비효율적일 수 있습니다. 특히, 훈련 시간이 길어질 수 있습니다.

2. SVC 모델은 결정 트리나 회귀 모델처럼 직관적으로 해석하기 어려워 모델이 어떻게 결정을 내리는지 설명하기 어려운 경우가 많습니다.

3. 클래스 불균형 문제에서 SVC의 성능이 저하될 수 있습니다. 이는 모델이 주요 클래스에 더 많은 가중치를 두기 때문입니다.

**하이퍼파라미터 최적화**

주로 조절하는 C와 gamma값을 조절 해보았습니다.

* gamma: 하나의 훈련 샘플이 미치는 영향의 범위를 결정합니다
* C: 각 포인트의 중요도를 제한하는 매개변수로, 해당 값이 커질수록 결정경계가 데이터에 정확하게 맞춰집니다.

sklearn모듈의 GridSearchCV를 이용하여 교차 검증과 아래의 하이퍼파라미터 최적화를 한번에 수행했습니다.

'C': [0.001, 0.01, 0.1, 1, 10, 100]

'gamma': [0.001, 0.01, 0.1, 1, 10, 100]

학습 결과 C: 1, gamma: 1에서 베스트 모델이 나왔습니다.

1. **평가**

* **정확도 Accuracy: 0.8165810711665444**

정확도는 전체 예측 중에서 올바르게 예측한 비율을 나타냅니다.

81.7%의 정확도로 모델이 전체 데이터를 잘 분류하고 있음을 의미합니다.

* **정밀도 Precision: 0.8130939809926082**

정밀도는 모델이 긍정으로 예측한 것들 중 실제로 긍정인 비율을 나타냅니다.

81.3%의 정밀도로 모델이 긍정 예측을 할 때 비교적 신뢰할 수 있음을 의미합니다.

* **재현율 Recall: 0.9134045077105575**

재현율은 실제 긍정 중에서 모델이 올바르게 예측한 비율을 나타냅니다.

91.3%의 재현율로 모델이 대부분의 긍정 사례를 잘 예측하고 있음을 의미합니다.

* **F1 Score: 0.8603351955307262**

F1 점수는 정밀도와 재현율의 조화 평균입니다. 이 값은 두 지표 간의 균형을 제공합니다.

86.0%의 F1 점수로 모델이 정밀도와 재현율 사이에서 균형을 잘 맞추고 있음을 나타냅니다.

* **ROC-AUC Score: 0.786509946162971**

ROC-AUC 점수는 ROC 곡선 아래의 면적을 나타냅니다. 1에 가까울수록 좋습니다.

0.787의 ROC-AUC 점수는 모델이 긍정과 부정을 잘 구분할 수 있음을 의미하지만, 완벽하지는 않습니다.

1. **시각화 및 결론**

**혼동 행렬**

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

TP (770): 모델이 긍정으로 올바르게 예측한 수

FP (177): 모델이 긍정으로 잘못 예측한 수

FN (73): 모델이 부정으로 잘못 예측한 수

TN (343): 모델이 부정으로 올바르게 예측한 수

이 혼동 행렬을 통해 모델이 대부분의 긍정 사례와 부정 사례를 올바르게 예측하고 있음을 알 수 있습니다.

**ROC 곡선**

텍스트, 도표, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

**결론 및 평가**

전처리 과정에서 review\_rating 데이터 거의 대부분이 5점에 몰려 있어 긍정과 부정 라벨로 변경할 때 기준점을 어디에 두어야 하는지가 힘들었습니다. 텍스트를 불용어를 제외하고 표제어 추출해 백터화 하는 등의 작업은 모든 라이브러리가 있어 처리하는데 크게 어려움은 없었던 것 같습니다.

정확도 (Accuracy): 81.7%로 모델이 전체적으로 좋은 성능을 보이고 있습니다.

정밀도 (Precision): 81.3%로 긍정 예측의 신뢰성이 높습니다.

재현율 (Recall): 91.3%로 모델이 대부분의 긍정 사례를 잘 예측하고 있습니다.

F1 점수: 86.0%로 정밀도와 재현율 사이에서 균형을 잘 맞추고 있습니다.

ROC-AUC 점수: 0.787로 모델이 긍정과 부정을 잘 구분하고 있습니다.