기말대체과제 보고서

201920998 이성재

**프로젝트 목표 (보고서 목차)**

1. **감성 분석의 기본 이해**: 감성 분석이 무엇이며 그 응용에 대해 학습
2. **데이터 전처리**: 머신 러닝 작업을 위해 데이터 전 처리하는 경험을 쌓음
3. **모델 선택 및 학습**: 수업에서 활용한 모델을 활용해서 구현하고 그 강점과 약점을 이해
4. **평가**: 적절한 지표를 사용하여 모델의 성능을 평가
5. **시각화 및 해석**: 결과를 시각화하고 해석
6. 감성분석의 기본 이해

감성분석이란 텍스트에 들어있는 감성, 긍정과 부정, 평가, 의견 등 인간이 느끼는 주관적인 정보를 컴퓨터를 통해 분석하는 과정입니다.

감성분석을 통해 본 프로젝트에서 사용한 데이터인 wine.review와 같은 데이터에서

평점 5점, 10점과 같은 수치적인 분석보다 더 정확한 정보와 평가를 분석할 수 있습니다.

1. 데이터 전처리

텍스트, 전자제품, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

전처리전 필요한 nltk라이브러리와 nltk모듈을 임포트 해줍니다.

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

자동 생성된 설명

데이터프레임으로 데이터를 로드하고 df의 행을 출력하여 내용을 확인해줍니다.

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

자동 생성된 설명

컬럼중에서 reviews.rating,reviews.text,reviews.title,name,brand를 분석대상으로 정합니다. 다음으론 텍스트를 전처리 해줍니다.

데이터를 최대한 살리기 위해 결측값은 제거가 아닌 빈문자열로 채워줍니다.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

불용어리스트, 토큰화 도구, 표제어 추출도구, vader\_lexion을 설치합니다.

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

자동 생성된 설명

불용어 및 스테머, 표제어 추출기를 초기화시켜줍니다.

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

자동 생성된 설명

감성분석전에 사용할 텍스트 컬럼인 reviews.text, reviews.title, name, brand를 텍스트 전처리를 위해 하나로 합처줍니다. 또한 사용하지 않는 컬럼은 모두 드랍합니다.

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

자동 생성된 설명

지금까지 전처리된 내용을 cleaned\_review.csv로 저장해주고

다시 데이터를 로드합니다.

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

자동 생성된 설명

감정점수를 진행한 결과입니다. compound 점수가 0.05 이상이면 긍정,

compound 점수가 -0.05 이하이면 부정이고 나머지는 중립입니다.

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

자동 생성된 설명

처음 roc curve 시각화를 진행하였을때 auc스코어가 0.3으로 너무 낮게 나와 분류성능이 제대로 나오지 못했다고 판단했습니다.

따라서 피처 엔지니어링을 통해 리뷰텍스트길이, 키워드빈도 피처를 만들어서 분류정밀도를 높여줍니다.

키워드는 전처리된 combined\_text에서 빈도를 확인합니다.

단어선정은 직접 데이터 파일을 확인하고 자주 나오는 단어를 데이터에서 적당히 골라서

5개 넣어줬습니다.(5개가 넘어가면 모델학습에서 시간이 너무 오래걸려서 5개만 선정했습니다.)

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

reviews.rating(평점)을 라벨링 해줍니다. 1~2는 부정, 3~5 긍정으로 합니다.

긍정이면 1을 리턴 부정이면 0을 리턴하도록 코드를 짰습니다.

텍스트, 스크린샷, 소프트웨어, 폰트이(가) 표시된 사진

자동 생성된 설명

라벨의 결측값을 제거하여주고 감정점수와 일치하는지 확인해줍니다.

라벨을 분리하여 x에는 피처, y에는 rating\_label을 입력해줍니다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

매치결과는 순조운것으로 보입니다.

1. 모델 선택 및 학습

수업에서 구현한 모델중 하나인 SVR모델을 선택했습니다.

SVR모델은 SVR(Support Vector Regression)로 머신 러닝 알고리즘 중 하나입니다. SVM(Support Vector Machine)에서 회귀 문제에 적용한 것이며 SVM이 분류 문제를 해결하기 위해 개발된 알고리즘이라면, SVR은 연속적인 값을 예측하는 데 사용됩니다.

SVR의 장점과 단점으로는

장점)

고차원 데이터에서도 효과적으로 작동합니다.

과적합을 방지하는 메커니즘이 내장되어 있습니다.

커널 트릭을 사용하여 비선형 데이터도 처리할 수 있습니다.

단점)

큰 데이터셋에서는 훈련 시간이 오래 걸릴 수 있습니다.

매개변수(ε, C, 커널 매개변수)를 튜닝하는 데 많은 시간이 소요될 수 있습니다.

1. 평가

텍스트, 스크린샷, 폰트, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

데이터셋을 학습 및 데스트 세트로 분할하였으며 8:2 비율로 분할하고 random\_state를 42로 설정하여 일관된 결과를 얻도록 하였습니다.

하이퍼파라미터 튜닝으로 그리드서치을 사용했습니다.

교차검증은 5회로 하였습니다.

Accuracy(정확도), Precision(정밀도), Recall(재현율), F1점수 모두 1에 가까운 수치가 나왔습니다.

AUC 스코어 또한 0.59로 높지는 않지만 0.6 정도의 분류성능을 가지고 있다고 나타납니다.

1. 시각화 및 해석

텍스트, 스크린샷, 소프트웨어, 직사각형이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

보시는 바와 같이 분류성능은 나쁘지 않게 나왔으나 혼동행렬 시각화자료를 보면

긍정 예측으로 쏠린 것을 볼 수 있습니다.

이것을 머신 러닝 모델이 학습 데이터의 패턴을 충분히 학습하지 못하여, 학습 데이터뿐만 아니라 새로운 데이터에 대해서도 성능이 떨어지는 현상인 언터피팅으로 판단했습니다.

이는 모델이 너무 단순해서 데이터의 구조적 복잡성을 제대로 반영하지 못해 발생한 것이라 보았습니다.

현재의 분류성능을 객관적으로 보았을 때 높은 수치가 아니며 그로 인해 예측성능예 떨어진 것으로 판단됩니다. 따라서 분류성능을 늘려보도록 하겠습니다.

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

자동 생성된 설명

이번에는 키워드를 긍정 키워드, 부정 키워드로 나누어주고 텍스트 길이 피처는 분류에 적합하지 않고 학습시간을 과하게 늘리는 요인이라서 없앴습니다.

또한 키워드 빈도 피처를 combine\_text가 아닌 review.text로 분석대상을 바꿉니다.

이것은 데이터 확인결과



이처럼 combine\_text칼럼에서 텍스트 깨짐 현상이 발견되어 정확한 분류를 위해 reviews.text를 간단한 텍스트 전처리후 분석에 사용합니다.

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

자동 생성된 설명

또한 reviews.rating은 3~5를 긍정으로 하였으나 3이라는 평점은 긍정이라기엔 모호한 부분이라 생각하여 3을 긍정범위에서 제외 시킵니다.

이후 모델 구현결과

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 소프트웨어, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

혼동 행렬을 보면 여전히 긍정예측으로 쏠려있습니다.

아무래도 긍정리뷰의 수가 압도적으로 많다보니 생기는 현상으로 보입니다.

반면에 auc점수는 눈에 띄게 올랐습니다.

0.59에서 0.72까지 상승했습니다.

빈도 키워드 피처에서 reviews.text의 키워드를 사용한것과 텍스트길이 피처 삭제, 키워드 추가가 분류성능에 유의미한 결과를 준것이라 볼 수 있겠습니다.

결과요약)

데이터 전처리, 감정점수 분석, 모델구현과 평가, 시각화까지 진행되었고

긍정으로 예측이 쏠린것을 언더피팅이라 판단하고 해결하려 하였으나 자료자체가

긍정리뷰수가 압도적으로 많아 데이터 불균형이 원인인것으로 결론지었습니다.

비록 긍정으로 예측이 쏠리는 것을 막지는 못했으나 키워드 피처를 늘려

분류성능을 올리고 필요없는 데이터를 삭제하는 등의 작업을 통해 더 나은 결과를 만들어 내보았습니다.

감사합니다.