기계학습 기말고사 대체 과제 보고서

소프트웨어학과 201920918 박재호

목차

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

감성 분석은 텍스트 데이터를 분석하여 데이터의 감정이나 의견을 분석하고 분류하는 과정이다. 자연어 처리의 한 분야로 긍정적, 부정적 감정을 식별하는데 사용된다.

감정분석의 주요 요소로 텍스트 전처리, 모델 학습 및 예측, 평가 및 검증이 있다.

텍스트 전처리에서는 노이즈 제거와 결측값 처리, 토큰화, 정규화, 불용어 제거가 있다. 노이즈 제거와 결측값 처리는 데이터 정리 과정이며 토큰화는 텍스트를 단어, 문장 등 작은 단위로 나누는 과정이고 정규화는 대소문자의 통일, 불필요한 문장부호 제거, 철자 오류 수정 등 텍스트를 일관성 있게 만드는 과정이며, 불용어(Stopwords Removal) 제거는 분석에 필요하지 않은 일반적인 단어를 제거하는 과정을 의미한다.

모델 학습에서는 로지스틱 회귀, 서포트 벡터 머신, 랜덤 포레스트등 다양한 기계 학습 모델 중 하나를 선택 후 그리드 서치나 랜덤서치를 사용하여 최적의 하이퍼파라미터를 찾는다, 하이퍼 파라미터는 모델의 성능에 영향을 미치는 설정 값들 이다.

모델 평가에서는 전체 예측 중 올바르게 예측한 비율인 정확도(Accuracy), 양성으로 예측한 것 중 실제 양성인 비율인 정밀도(Precision), 실제 양성인 것 중 양성으로 예측한 비율인 재현율(Recall)을 측정하고 정밀도와 재현율의 조화 평균을 계산하여 모델의 성능을 평가하는 F1 점수와 ROC-AUC의 그래프를 활용하여 모델 평가에 활용한다. 모델이 학습 데이터에 너무 치우쳐져 일반화가 잘 안될 경우 오버피팅 및 언더 피팅을 논의 후 더 많은 데이터 수집, 정규화, 모델 단순화 등을 추가적으로 진행 후 학습한다.

응용분야로는 소셜미디어 분석, 제품/서비스 평가, 금융 시장 분석, 정치적 감성 분석 등의 응용 분야가 존재하며 이외에도 많은 산업 분야에서 중요한 정보를 제공하고 의사결정에 도움을 주는 도구 이다.

1. 데이터 전처리

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

데이터 전처리를 위하여 먼저 데이터를 불러온다.

데이터 확인 결과

텍스트, 스크린샷, 폰트, 메뉴이(가) 표시된 사진

자동 생성된 설명

32개의 열과 2890개의 데이터가 존재 하는 것을 확인 할 수 있고

감정 분석을 하기 위하여 필요한 데이터는 자연어 데이터가 있는 reiviews.text와 reviews.title, 감정 분석한 결과를 토대로 모델을 평가하기 위한 reviews.rating 데이터 만을 남긴다. 텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

정리된 데이터를 확인해 보면 누락된 데이터가 있는 것을 확인 할 수 있다. 따라서 결측값 처리를 해주어야 한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

결측값이 처리되고 2406개의 데이터가 남은 것을 확인 할 수 있다.

다음으로는 토큰화 및 불용어 제거를 진행 하였다.

텍스트, 스크린샷, 소프트웨어, 디스플레이이(가) 표시된 사진

자동 생성된 설명

불용어 제거가 정상적으로 진행 되었는지 확인 하기 위하여 먼저 reviews.text 열에서 가장 많은 단어수를 출력해 보았을 때 332개의 단어로 이루어진 리뷰 텍스트 데이터가 있는 것을 확인 할 수 있다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

NLTK 불용어 세트를 이용하여 리뷰 텍스트 데이터와 리뷰 타이틀 데이터에 불용어 제거를 진행하였고

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

불용어 처리 이후 다시 리뷰 텍스트 데이터에서 가장 많은 단어수를 확인 해 보았을 때 250 단어로 불용어 처리가 제대로 진행 된 것을 확인 할 수 있었다.

다음으로 스테머와 표제어 추출기를 이용하여 텍스트 정규화를 진행 하였다.

텍스트, 스크린샷, 소프트웨어, 운영 체제이(가) 표시된 사진

자동 생성된 설명

1. 모델 선택 및 학습

텍스트, 스크린샷, 소프트웨어, 디스플레이이(가) 표시된 사진

자동 생성된 설명

수업에서 활용한 vader 모델을 활용하여 리뷰 텍스트와 리뷰 타이틀 데이터에 대하여 감정 계산 함수를 적용시켜 각 sentiment 열에 입력하였다.

Vader (Valence Aware Dictionary and sEntiment Reasoner) 감정 분석기는 긍정적, 부정적, 중립적 감정 점수를 계산하여 이 점수를 합산한 점수를 통해 전체 감정의 강도를 평가한다.

VADER의 강점

1. 간편함 : 사전 학습된 모델로, 추가학습이 필요없고 복잡한 전처리 과정 없이 바로 사용 할 수 있다.
2. 구어체와 이모티콘 처리 : 이모티콘과 약어를 인식 할 수 있다.
3. 속도 : 사전 기반 접근 방식으로, 감성 분석 속도가 매우 빠르다.

VADER의 약점

1. 한계된 어휘 범위 : 사전에 정의된 어휘에 의존하여, 신조어에 대하여 잘못된 점수를 부여할 수 있다.
2. 문맥 이해 부족 : 단어의 위치나 문맥을 고려하지 않아 복잡한 문맥을 파악하지 못할 수 있다.
3. 평가

최종적으로 5점 리뷰 데이터를 포함한 데이터셋과 5점 리뷰를 포함하지 않은 데이터셋을 이용하여 모델을 평가하였다.

* 5점 리뷰 데이터를 포함한 데이터셋

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명



5점 리뷰에 대한 f1 점수는 0.9로 높고 나머지 점수 리뷰에 대한 f-1 점수는 처참하다.

정확도는 0.82 로 상당히 정확하며 ROC-AUC 점수는 0.6669로 준수하다.

* 5점 리뷰 제외한 데이터셋

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명



5점 리뷰에 대한 f1 점수는 0.76으로 줄었지만나머지 점수 리뷰에 대한 f-1 점수는 많이 증가 하였다.

정확도는 0.59 로 낮아졌지만 준수하고 ROC-AUC 점수는 0.67968 아주 조금 올랐다. 해당 모델은 0.67 정도의 점수를 받을 만한 모델이라 평가한다.

1. 시각화 및 해석

먼저 5점 리뷰를 포함한 전체 데이터 감성 분석 결과 이고

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

5점을 제외한 데이터셋 분석 결과 이다.

텍스트, 스크린샷, 디스플레이, 직사각형이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

5점 리뷰를 포함한 데이터의 분석 결과에서는 대부분의 데이터가 5점이었기 때문에 거의 모든 예측을 5점으로만 하여 정확도는 높지만 제대로 예측을 하고 있는 것인 것 판단이 불가능 한 상태라고 판단하여 5점 리뷰를 제외한 데이터셋을 이용하여 다시한번 평가를 진행 한 것이었으며 두개 모두의 ROC-AUC 점수는 비슷한 결과를 보여준 것을 보아 모델의 점수는 대략 0.67으로 판단하게 되었다.

5점을 제외한 데이터의 경우 데이터의 총량이 부족하여 결과가 아쉽게 나왔다고 생각하며 만약 5점을 제외한 데이터셋의 총량도 충분히 컸다면 해당 모델의 평가가 오히려 더 높게 나올 수 있지 않았을까 생각해 본다.

최종적으로 와인 리뷰 데이터셋은 기본적으로 감성 분석 모델이 평점을 예측은 하기 쉬우나 해당 모델의 성능을 평가하기에는 아쉬운 데이터 셋이라고 생각한다.