**기계학습(8585) 기말대체 과제**

**보고서**

**제품 review 데이터의 감성 분석**

**201921018 이준배**



**목차**

**1. 감정 분석의 기본 이해-----------------------------------------------------------3P**

**2. 데이터 전처리---------------------------------------------------------------------3P**

**2-1. 결측 값이 많은 열 제거-------------------------------------------------------3P**

**2-2. 결측 값이 있는 행 제거-------------------------------------------------------3P**

**2-3. 필요한 열 선택----------------------------------------------------------------4P**

**2-4. 노이즈 제거--------------------------------------------------------------------4P**

**2-5. 불용어 제거--------------------------------------------------------------------4P**

**2-6. 토큰화--------------------------------------------------------------------------5P**

**2-7. 표제어 추출하기---------------------------------------------------------------6P**

**2-8. 스테밍--------------------------------------------------------------------------6P**

**3. 모델 선택 및 학습------------------------------------------------------------------------6P**

**3-1. 모델 선택 및 전처리----------------------------------------------------------6P**

**3-2. 학습-----------------------------------------------------------------------------8P**

**4. 평가----------------------------------------------------------------------------------------9P**

**5. 시각화 및 해석--------------------------------------------------------------------------10P**

**5-1. 혼동 행렬 시각화-------------------------------------------------------------10P**

**5-2. ROC 곡선 시각화-------------------------------------------------------------12P**

**6. 결론--------------------------------------------------------------------------------------13P**

**1. 감정 분석의 기본 이해.**

감정 분석이란 영어로 Sentiment Analysis로 텍스트에 있는 단어들 속에 어떠한 감정이 들어있는지 분석하는 자연어 처리 기법입니다. 오피니언 마이닝이라고도 불리며 텍스트에 담긴 의견, 태도, 감정과 같은 정적인 요소를 뽑아내는 작업입니다. 이번 과제에서는 와인들에 달린 리뷰들에서 어떠한 감정이 있는지 전처리 과정 후 감정 점수를 뽑아 확인한 후 이를 긍정인지 부정인지 라벨링하여 사용하였습니다. VADER감정 분석기를 사용하였고 -1~1 사이에 값이 배정되게 됩니다. 작을수록 부정적인 내용이라는 뜻이고 클수록 긍정적인 내용입니다. 즉 와인 리뷰 텍스트에 들어있는 단어들을 활용하여 리뷰 자체가 긍정적인 내용인지 부정적인 내용인지 알아내는 작업입니다. 이를 토대로 리뷰가 달린 와인의 브랜드를 모니터링 할 수도 있고 마케팅 전략에 사용될 수도 있습니다. 리뷰 분석을 통해 제품을 개선하는 것도 좋은 방법입니다.

**2. 데이터 전처리**

학번이 짝수로 끝나므로 wine\_review.csv 를 사용하였습니다. 2890행과 32개의 열의 와인 과 관련된 데이터 파일입니다.

**2-1. 결측 값이 많은 열 제거**

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

결측 값이 너무 많게 된다면 의미가 없을 수 있습니다. 그 행들만 지운다면 너무 많은 데이터들이 사라지게 됩니다. 그러하여 1000개 이상의 결측 치가 있는 열은 30%이상이 결측 값이므로 과감하게 삭제하였습니다. 또한 삭제되는 열들이 중요하지 않은 열이었습니다.

**2-2. 결측 값이 있는 행 제거**

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

결측 값이 많은 열을 제거하였으니 이번엔 결측 값이 있는 행들을 제거하였습니다. 그러하여 데이터 프레임에는 결측 값이 없어진 것을 알 수 있습니다.

**2-3. 필요한 열 선택**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

작업에 모든 열이 필요한 것이 아니라 필요한 다음과 같은 열들만 뽑아 데이터프레임에 저장하였습니다.

**2-4. 노이즈 제거**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

감정분석이 사용될 열인 reviews.text 열에 노이즈를 제거하였습니다. 모든 문자들을 소문자로 교체하고 특수문자를 제거하는 작업입니다.

**2-5. 불용어 제거**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

**불용어 제거는 nltk 라이브러리의 stopwords 불용어들을 활용하여 제거하였습니다. 불용어 제거란 큰 의미가 없는 단어들을 제거 시켜주는 작업입니다. 해당 작업으로 reviews.text에 있는 this 나 I 와 같은 단어들이 사라졌습니다.**

**2-6. 토큰화**

**텍스트, 스크린샷, 디스플레이, 폰트이(가) 표시된 사진

자동 생성된 설명**

**토큰화란 주어진 텍스트를 작은 단위로 나누는 작업입니다.**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

**위 사진과 같이 reviews.text 열이 토큰으로 나눠진 것을 볼 수 있습니다.**

**2-7. 표제어 추출하기**

텍스트, 스크린샷, 폰트, 디스플레이이(가) 표시된 사진

자동 생성된 설명

표제어 추출하기란 영어 단어의 원형을 변환하는 작업입니다. 예를들어 am, are is를 원형인 be로 바꿔주는 작업입니다. 마찬가지로 감정 분석에 사용될 reviews.text열에 적용하였습니다.

**2-8. 스테밍**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

스테밍이란 단어의 어간을 추출하는 작업입니다. 예를들어 running, runs, runner와 같은 단어를 run으로 바꾸는 작업입니다. 또한 reviews.text열에 적용하였습니다.

위 8과정을 거쳐 데이터 전처리를 완료 하였습니다.

**3. 모델 선택 및 학습**

**3-1. 모델 선택 및 전처리**

감정 분석 모델은 실습에 사용된 nltk에 있는 vader\_lexicon를 사용하겠습니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

다음과 같은 코드를 사용하여 있는 vader\_lexicon사용 하였습니다. 전처리가 완료된 reviews.text 열의 감정 점수를 추출하여 Sentiment\_Score로 데이터프레임에 추가하였습니다. 감정 점수는 -1~1 사이의 값이며 클수록 긍정적인 내용이라는 뜻입니다. 여기서 reviews.text는 토큰화가 진행되어 리스트 형태입니다. 그로 인해 리스트를 문자열로 변환해주는 작업을 해야 합니다.

긍정적인 점수를 받은 리뷰를 보면 love와 같은 긍정적인 단어들이 많은 반면 부정적인 점수를 받은 리뷰들을 보면 stress, small과 같은 부정적인 단어들이 많이 있습니다.

스크린샷, 텍스트이(가) 표시된 사진

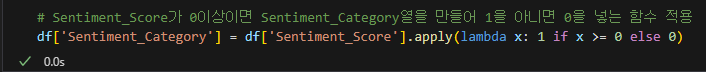
자동 생성된 설명

위 코드를 활용하여 Sentiment\_Score와 reviews.rating 점수가 서로 맞는지 확인하는 작업을 해보았습니다. Sentiment\_Score의 기준은 0으로 reviews.rating의 기준은 3으로하여 두 값이 같은 분류에 있으면 True를 아니면 False를 match라는 열을 만들어 추가하였습니다.

텍스트, 스크린샷, 폰트, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

잘 매칭된 값은 1229개 아닌 경우는 150개로 대부분 잘 매칭된 것을 볼 수 있습니다. 감정 분석이 잘되었거나 reviews.rating 열이 의미 있는 열임을 뜻합니다.



다음으로는 학습을 위해 위 코드를 활용하여 감정 분석 점수를 이진법으로 새로운 열을 만들었습니다. 감정 점수가 양수이면 1 음수이면 0 값을 가지게 됩니다. 즉 긍정적인 리뷰이면 1 부정적인 리뷰이면 0을 가지게 되는 것입니다.

**3-2. 학습**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

학습에 필요한 패키지들을 미리 설치한 후 진행합니다.

텍스트, 스크린샷, 폰트, 디스플레이이(가) 표시된 사진

자동 생성된 설명

학습을 하기 위해 x값과 y값을 지정해 줍니다. x값은 reviews.text 열을 공백으로 구분된 문차열로 변환 한 후 TF-IDF벡터화를 진행하여 사용합니다. 벡터화를 해줘야 학습이 가능해집니다. y값은 감정 점수를 이진분류한 Sentiment\_Category열을 사용합니다.



학습 데이터와 테스트 데이터를 8:2로 나누는 작업입니다. Random\_state을 지정하여 다시 코드를 돌려도 같은 결과값이 나오게끔 설정해줍니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

다음 그리드서치를 활용하여 최적의 파라미터를 찾는 작업입니다. 그리드서치란 정의된 하이퍼 파라미터들로 모든 조합을 생성 후 각 조합에 대해 학습 후 최고의 하이퍼 파라미터를 찾는 방식입니다. 위 코드에서는 c값을 [0.001, 0.01, 0.1, 1, 10, 100]로 주었으며 그중 10이 최적의 파라미터로 출력 되었습니다. 코드에서 cv=5는 교차 검정을 활용하는 코드 입니다. Cv=5는 k-fold 교차 검증에서 5개의 폴드를 사용한다는 뜻입니다. 즉 데이터를 5부분으로 나눠 사용하고 각 폴드를 학습하고 나머지 폴드를 활용하여 검증하는 것을 5번 반복하는 것입니다. 최적의 하이퍼 파라미터를 찾는 것에 도움을 줍니다.



그 후 그리드 서치로 찾은 최적의 하이퍼 파라미터로 튜닝된 모델을 활용하여 학습을 진행하게 됩니다.

**4. 평가**

텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

위 코드를 활용하여 정확도, 정밀도, 재현율, F1점수 ROC-AUC를 계산하였습니다.

텍스트, 스크린샷, 폰트, 블랙이(가) 표시된 사진

자동 생성된 설명

출력 결과는 위 사진과 같습니다.

정확도는 0.913으로 전체 예측 중 올바르게 예측한 비율을 나타냅니다. (맞춘 예측 수 / 전체 예측 수)으로 계산됩니다.

정밀도는 0.9148로 모델이 True로 예측한 것이 실제로 True인 비율을 나타내는 지표입니다. TP / (TP + FP)으로 계산됩니다.

재현율은 0.996으로 실제 True인 데이터를 모델이 True로 예측한 비율입니다. TP / (TP + FN)으로 계산됩니다.

F1점수는 0.9537으로 정밀도와 재현율의 조화 평균으로 계산되는 지표입니다. 2 \* (정밀도 \* 재현율) / (정밀도 + 재현율) 으로 계산됩니다.

ROC-AUC는 0.8090으로 Receiver Operating Characteristic - Area Under the Curve의 약자로 모델의 성능을 평가하는데 사용되는 지표입니다. ROC 곡선의 아래 면적을 나타냅니다. 보통 0.5~1사이의 값을 가지며 1에 가까울 수록 성능이 좋은 모델입니다.

성능이 나쁘지는 않지만 또 매우 좋다고도 할 수 없습니다. 정확도와 정밀도도 0.95이상이 나오면 좋을거 같고 ROC-AUC도 0.9이상이 나오면 좋을거 같습니다. 재현율은 현재도 좋은 상태입니다. 좋은 데이터를 추가하여 수행하거나 여러 학습 모델을 학습한 후 제일 좋은 모델을 선택하는 앙상블 기법을 사용하거나 데이터 전처리에 좀더 신경을 써 모델의 성능을 올려보는것도 좋아보입니다.

정확도와 정밀도, 재현율은 잘 나오지만 ROC-AUC는 다소 낮게 나오는것으로 보아 오버피팅이 일어나진 않은거 같습니다. 하지만 정확도가 낮은게 아니라 언더피팅인것도 아닐거 같습니다. 만약 언더피팅이라면 더 복잡한 모델인 SVM을 사용하거나 하이퍼 파라미터를 다시 조정하는 방법들이 있습니다. 반면 오버피팅이라면 더 많은 데이터를 사용하거나 앙상블 기법을 사용하는 방법이 있습니다.

**5. 시각화 및 해석**

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

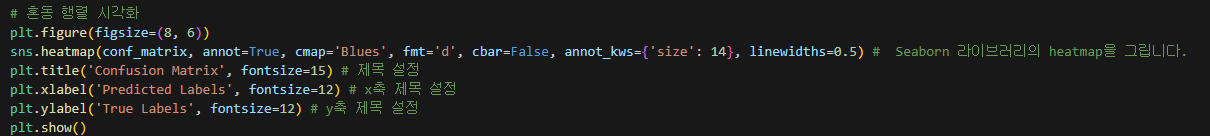
시각화에 필요한 라이브러리들을 설치 한 후 진행합니다.

**5-1. 혼동 행렬 시각화**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

위 와 같은 코드를 활용하여 혼동 행렬을 계산하고 변수에 정합니다.



다음과 같은 코드를 활용하여 혼동 행렬을 시각화 합니다.

텍스트, 스크린샷, 폰트, 도표이(가) 표시된 사진

자동 생성된 설명

혼동 행렬이란 TN, FP, FN, TP값을 행렬로 표현한 것으로

[[TN, FP]

[FN,TP]] 형식으로 나타내집니다.

TN은 실제 Negative인 데이터를 Negative로 정확히 예측한 경우의 수입니다. 저희는 5가 나왔습니다.

FP는 실제 Negative인 데이터를 Positive로 잘못 예측한 경우의 수입니다. 23이 나왔습니다.

FN은 실제 Positive인 데이터를 Negative로 잘못 예측한 경우의 수입니다. 1개가 나왔습니다.

TP는 실제 Positive인 데이터를 Positive로 정확히 예측한 경우의 수입니다. 247개가 나왔습니다.

잘못 분류한 데이터는 23+1으로 24개 데이터입니다. 잘 분류한 데이터는 247+5으로 252개입니다. 대부분 잘 맞춘 것을 볼 수 있습니다. 또한 혼동 행렬이 한쪽으로 치우쳐진 이유는 학습에 사용된 감정 분석 이진분류 값이 0보다 1의 개수가 많기 때문이라고 생각됩니다.

텍스트, 스크린샷, 디스플레이, 폰트이(가) 표시된 사진

자동 생성된 설명

**5-2. ROC 곡선 시각화**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

다음과 같은 코드를 활용하여 ROC 곡선을 시각화 하였습니다.

텍스트, 라인, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

시각화 한 그래프는 위와 같습니다.

파란색 점선은 x=y 그래프이며 기준선을 표시한 것입니다. 노란색 그래프가 roc그래프 입니다. 오른쪽 아래 범례를 보면 roc그래프의 아래 면적이 0.81이라고 나온 것을 볼 수 있습니다. 보통 0.8 이상이면 좋은 성능을 나타내며 0.9 이상이 우수한 성능을 나타냅니다. 0.8초반이므로 좋기는 하지만 많이 좋지는 않고 나쁘지도 않은 성능을 보여주는 모델이라는 것입니다.

**6. 결론**

모델의 전반전인 성능은 괜찮게 나왔습니다. 하지만 위에서 말한 것처럼 roc 곡선의 면적이 0.8후반이나 0.9까지 올라오면 좋고 정확도와 정밀도 재현율도 조금 더 올리면 좋을 거 같습니다. svm과 같은 좀 더 복잡한 모델을 사용하거나 데이터 추가하기, 혹은 앙상블 기법을 사용하면 지금보다 조금 더 좋은 결과가 나올 것으로 보입니다. 현재 기준에서는 위에서 말했다시피 오버 피팅의 가능성은 적어 보입니다.