**기계학습 보고서**

담당 교수 : 백우진 교수님

이름 - 김지원

학번 - 202121855

전공 - 컴퓨터 공학과

1. **감성 분석의 기본 이해**  
   신발 제품 리뷰 중 리뷰텍스트에서 감성 단어를 추출해 긍정적 또는 부정적 감정을 분류한다. 이 작업을 통해 특정 제품 리뷰에서 소비자의 감정을 효과적으로 파악할 수 있다.
2. **데이터 전처리**

스크린샷, 텍스트, 폰트이(가) 표시된 사진

자동 생성된 설명  
데이터 전처리를 위해 csv 파일에서 변수명을 확인하고 분석에 필요하지 않은 열을 삭제한다. 감성에 관한 단어가 주로 있는 ‘review\_title’, ‘review\_text, 와 감성을 분석한 후 평점으로 분류하기 위한 ‘review\_rate’로 분석한다. ‘review\_text’ 열에 결측값이 있는 행을 삭제한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

대소문자가 구별되지 않으면 동일한 단어도 서로 다른 언어로 인식될 수 있고, 단어 집합의 크기를 줄이기 위해 소문자로 변환해 단순화한다.

1. **모델 선택 및 학습**  
    텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

   자동 생성된 설명

불용어 리스트를 불러와 porterstemmer 객체를 초기화 하고 텍스트 전처리 함수를 정의한다. ‘review\_text’에 있는 구두점을 제거하고 단어 토큰화, 불용어 제거, 어간 추출로 전처리하고 ‘cleaned\_review\_text’ 열에 저장한다.

스크린샷, 텍스트, 폰트이(가) 표시된 사진

자동 생성된 설명  
전처리된 텍스트 데이터는 X로 평점이 4점 이상이면 긍정, 그렇지 않으면 부정으로 설정한다. 이때 긍정은 1, 부정은 0으로 변환해 Y로 설정한다. 데이터를 학습, 테스트셋으로 분할하고 데이터의 40%를 테스트셋으로 사용한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명  
텍스트 데이터를 벡터화 하고 모델을 초기화 하고 학습셋으로 모델을 학습시킨다.

테스트셋에 대한 감성을 예측한다.

-평점을 4 이상으로 설정한 이유

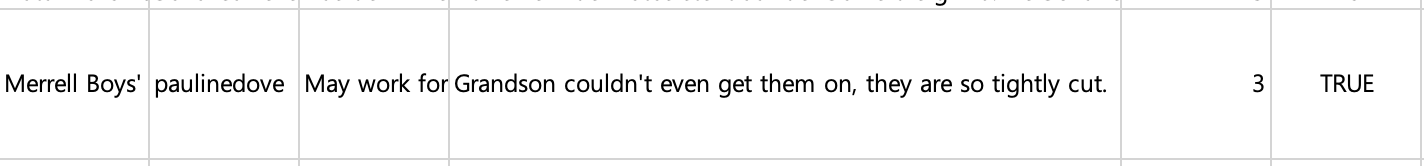
텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명  
텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

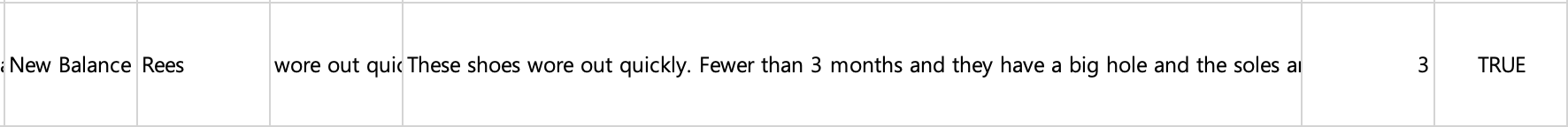
자동 생성된 설명

부정 리뷰에 대한 정밀도와 재현율이 상대적으로 낮은 결과가 나타났고 부정 리뷰를 많이 놓치고 있는 결과를 보인다.

-3점인 리뷰텍스트를 확인



Grandson couldn't even get them on, they are so tightly cut.



These shoes wore out quickly. Fewer than 3 months and they have a big hole and the soles are separating from the shoe. We bought New Balance from the local NB store and they lasted 7 months. These were disappointing.

실제 3점인 데이터를 확인했을 때 부정 리뷰인 것으로 나타났다.

더 정확한 분류를 위해 평점을 4점 이상으로 설정한 후 학습을 진행한다.

1. **평가**

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명  
\*precision(정밀도) : 모델이 양성으로 예측한 것 중 실제 양성인 비율로 0(부정)은 0.74로 모델이 부정으로 예측한 것 중 74%가 실제 부정으로 나타났고 1(긍정)은 0.86으로 모델이 긍정으로 예측한 것중 86%가 실제 긍정인 것으로 나타났다.

\*Recall(재현율) : 실제 양성 중 모델이 양성으로 정확하게 예측한 비율로 0(부정)은 0.54로 실제 부정인 것 중 54%가 모델이 부정으로 예측, 1(긍정)은 0.94로 실제 긍정인 것 중 94%를 모델이 긍정으로 예측한 것으로 나타났다.

\*F1-score : 정밀도와 재현율의 조화 평균으로 0(부정)은 0.63, 1(긍정)은 0.90으로 나타났다.

\*Accuracy(정확도) : 전체 샘플 중 모델이 정확하게 예측한 비율로 0.84

\*Macro avg(marco 평균) : 모든 클래스의 정밀도, 재현율, F1-score의 산술 평균으로 정밀도는 0.80, 재현율은 0.74, F1-score는 0.76으로 나타났다.

\*Weighted avg(가중 평균) : 각 클래스의 샘플 수를 가중치로 사용해 계산한 평균으로 정밀도는 0.83, 재현율은 0.84로 나타났다.

긍정클래스의 재현율은 0.94로 높으나 부정클래스의 재현율이 0.54로 모델이 부정 리뷰를 잘 잡아내지 못하고 있어 언더피팅이 되었다. 전체적인 정확도는 좋지만 부정 클래스에서의 성능이 떨어져 전체 데이터의 패턴을 충분히 학습하지 못한 것으로 보인다. 이를 해결하기 위해 더 복잡한 모델(Decision Tree, Random Forest 등)을 사용하거나 어간 추출이 아닌 표제어 추출로 텍스트를 전처리 하는 것으로 실험해볼 수 있다.

1. **시각화 및 해석**

텍스트, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명

* 왼쪽 위(TN) : 실제 부정클래스, 부정으로 예측된 샘플 수, 365
* 오른쪽 위(FP) : 실제 부정클래스, 긍정으로 예측된 샘플 수, 311
* 왼쪽 아래(FN) : 실제 긍정클래스, 부정으로 예측된 샘플 수, 126
* 오른쪽 아래(TP) : 실제 긍정클래스, 긍정으로 예측된 샘플 수, 1924

텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

X축 거짓양성비율(FPR) : 실제 음성인 샘플 중 양성으로 잘못 분류된 비율

Y축 진실양성비율, 재현율(TPR) : 실제 양성인 샘플 중 양성으로 잘 분류된 비율

AUC값은 0.87로 0.5보다 크므로 좋은 성능을 나타낸다. 무작위 분류 대각선에서 곡선이 멀리 떨어져 있어 잘 분류하는 성능을 보이고 있다.