**기계학습 (8585)**

**기말 대체 과제 보고서**

**과제: 제품 review 데이터의 감성 분석**

**▶ 소프트웨어 전공**

**201920879 김태현**



**1. 감성 분석의 기본 이해**

**1) 감성 분석이란?**

감성 분석(Sentiment Analysis)은 텍스트 데이터를 분석하여 그 안에 포함된 감정이나 의견을 파악하는 자연어 처리(Natural Language Processing, NLP) 기술입니다. 주로 긍정적, 부정적 등의 감정 범주로 텍스트를 분류합니다.

**2) 감성 분석의 응용 분야**

감성 분석은 다양한 분야에서 활용될 수 있습니다. 예를 들어, 소셜 미디어 모니터링에서는 트위터나 페이스북 같은 플랫폼에서 사용자들이 작성한 게시물의 감정을 분석하여 고객 만족도를 모니터링할 수 있습니다. 제품 리뷰나 고객 설문조사, 콜센터 대화에서 고객의 감정을 파악하여 제품 개선이나 서비스 품질 향상에 기여할 수 있습니다. 이와 같이 감성 분석은 텍스트 데이터를 기반으로 사람들의 감정과 의견을 파악하고 이를 통해 더 나은 의사 결정을 지원하는 데 중요한 역할을 하고 있습니다.

**2. 데이터 전처리**

**0) 환경 구성**

* 감성 분석에 필요한 모듈을 설치하고 적용

**텍스트, 전자제품, 스크린샷, 디스플레이이(가) 표시된 사진

자동 생성된 설명**

**1) 데이터를 읽고 데이터 구조를 확인**

* amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv를 pandas로 읽어와 데이터 프레임으로 저장
* 총 리뷰 데이터의 개수는 6823개
* 총 11개의 열이 존재

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

자동 생성된 설명**

* 실제 데이터 정보 확인
* 열에 대한 정보

1. url: 상품의 링크
2. product\_name: 상품의 이름
3. reviewer\_name: 리뷰 작성자의 이름
4. review\_title: 리뷰의 제목
5. review\_text: 리뷰의 내용
6. review\_rating: 리뷰의 별점
7. verified\_purchase: 실제 구매 여부
8. review\_date: 리뷰 작성일
9. helpful\_count: 리뷰에 대한 평가 정보
10. uniq\_id: 리뷰의 고유 아이디
11. scraped\_at: 리뷰 등록일

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

자동 생성된 설명**

**2) 데이터 전처리**

* 모델 성능 향상을 위해 데이터 전처리를 통해 테이터를 정제화시키는 과정이 필요

1. 중복 데이터 제거: 데이터셋 내 중복된 항목들을 식별하고 제거하여 데이터의 정확성 향상
2. 주요 특성 설정: 분석에 중요한 특성(컬럼)을 선택하여 데이터셋을 간소화
3. 결측치 확인 및 제거: 누락된 값이 있는 행 처리
4. 노이즈 제거: 불필요한 단어(숫자, 다중 공백, 특수 문자 등)을 제거, 텍스트 소문자로 변경
5. 토큰화: 텍스트 데이터를 단어 단위로 토큰화
6. 불용어 제거: 분석에 불필요한 단어 제거(Ex: 'is', "the', 'and' 등)
7. 표제어 추출: 단어의 기본 형태(표제어)로 변환(Ex: 'running', 'ran', 'runs' → 'run')
8. 중복 데이터 제거

* 총 리뷰 데이터의 개수는 6823개
* uniq\_id열 기준으로 중복된 행의 개수가 364개
* 예시로 1002번 인덱스의 리뷰 데이터와 1828번 인덱스의 리뷰 데이터가 같은 것을 확인

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

자동 생성된 설명**

* uniq\_id 열을 기준으로 첫 번째 행만 유지, 나머지 행 삭제

1. 원본 데이터: 6823
2. 중복 데이터: 364
3. 중복 데이터를 제외한 데이터: 6823 - (364 / 2) = 6641

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

자동 생성된 설명

1. 주요 특성 설정

* 11개의 열 중에 감성 분석에 사용할 열을 선택
* review\_title, review\_text, review\_rating 3가지 칼럼을 주요 특성으로 사용

1. review\_title: 리뷰의 핵심 내용을 요약하여 감성 분석에 중요한 단서
2. review\_text: 리뷰의 상세 내용을 포함하고 있어 감성 분석을 위한 정보
3. review\_rating: 감성 분석 결과에 대해 맞게 분석이 되었는지 확인할 수 있는 지표

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

자동 생성된 설명

1. 결측치 확인 및 제거

* 결측치를 확인하기 위해 isnull()과 sum()을 사용하여 결측 값의 개수를 확인

1. 제목 결측치: 1개의 결측치 발견
2. 리뷰 결측치: 8개의 결측치 발견
3. 레이팅 결측치: 0개의 결측치 발견

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* 제목, 리뷰에 결측치가 존재하는 행은 제거

1. 중복 데이터를 제외한 데이터: 6641
2. 결측치가 존재하는 데이터 : 1 + 8 = 9
3. 결측치를 제외한 데이터: 6641 - 9 = 6632

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

자동 생성된 설명

1. 노이즈 제거: 불필요한 단어(숫자, 다중 공백, 특수 문자 등)을 제거, 텍스트 소문자로 변경
2. 토큰화: 텍스트 데이터를 단어 단위로 토큰화
3. 불용어 제거: 분석에 불필요한 단어 제거(Ex: 'is', "the', 'and' 등)
4. 표제어 추출: 단어의 기본 형태(표제어)로 변환(Ex: 'running', 'ran', 'runs' → 'run')

* 노이즈 제거, 토큰화, 불용어 제거, 표제어 추출를 위한 전처리 함수 선언
* review\_text열과 review\_title열에 전처리 함수 적용

텍스트, 전자제품, 스크린샷, 디스플레이이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

**3. 모델 선택 및 학습**

**1) 모델 구현**

* 수업 시간에 배운 VADER 감성 분석기를 활용하여 감성 점수를 계산하고 해당 리뷰가 긍정적인지 부정적인지 확인

1. NLTK의 VADER 감성 분석기를 초기화하고 감성 점수를 계산하여 반환하는 함수 선언
2. review\_title과 review\_text에 대해 감성 계산을 진행
3. 각 점수는 데이터 프레임에 title\_compound와 text\_compound로 저장
4. title\_compound와 text\_compound의 평균을 리뷰의 최종 vader score인  Sentiment \_Score로 저장
5. Sentiment\_Score의 값이 0보다 크거나 같으면 1(긍정), 아니라면 0(부정)으로 라벨 분류
6. 감성 분석 결과 확인

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

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

자동 생성된 설명

* Sentiment\_Score는 -1에서 1까지 인것을 확인 가능
* Sentiment\_Score의 최저점은 -0.8726, 최고점은 0.9602

텍스트, 멀티미디어 소프트웨어, 소프트웨어, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* Sentiment\_Score 가 0이상일 경우 1(긍정), 아닐 경우 0(부정)으로 라벨 분류

텍스트, 소프트웨어, 멀티미디어 소프트웨어, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 메뉴이(가) 표시된 사진

자동 생성된 설명

* review\_rating이 3점이상인 경우에는 대부분 올바르게 처리
* review\_rating이 2점이하인 경우에는 대부분 올바르지 못하게 처리
* 더 자세히 감성 분석 결과에 대해 확인 필요

텍스트, 멀티미디어 소프트웨어, 소프트웨어, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진

자동 생성된 설명

* great kind super와 같이 긍정적인 단어가 보인 리뷰에서는 높은 점수 획득

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

자동 생성된 설명

* die와 같이 부정적인 단어가 보인 리뷰에서는 낮은 점수 획득
* *다만, 리뷰가 영어로 적혀있지 않은 것으로 보이나 영어로 판독하여 부정적인 결과가 나온 리뷰들도 있는 것으로 보입니다.*

텍스트, 소프트웨어, 멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* 정확한 분석은 5450개
* 비정확한 분석은 1182개
* 약 80퍼센트의 정확성
* *잘못 예측될만한 상황이 있음에도 불구하고 생각보다 높은 정확성을 보입니다.*

텍스트, 스크린샷, 소프트웨어, 디스플레이이(가) 표시된 사진

자동 생성된 설명

**2) 모델 학습 및 튜닝**

* 모델은 로지스틱 회귀 모델을 사용

1. 학습에 사용될 데이터(review\_title, review\_text)를 TF-IDF 벡터화
2. 로지스틱 회귀 모델 학습 및 평가에 사용될 데이터셋을 8:2로 분할
3. 하이퍼파라미터 튜닝 기법으로는 그리드 서치를 사용
4. C와 max\_iter에 대한 하이퍼파라미터 튜닝
5. 교차 검증을 위한 fold 횟수는 5회로 지정

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

자동 생성된 설명

* review\_title과 review\_text 칼럼을 하나의 칼럼으로 합친 후 TF\_IDF 백터화 진행
* TF-IDF 벡터화는 문자열 리스트를 학습을 위해 벡터로 바꿔주는 작업
* 로지스틱 회귀 모델 학습 및 평가를 위해 데이터셋 분할
* 학습 데이터와 평가 데이터는 8:2 비율로 분할

텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* 로지스틱 회귀 모델 사용을 위한 초기화
* 하이퍼마라미터 그리드 설정
* cv = 5는 k-fold 교차 검증을 사용하여 데이터를 5개 부분으로 나누어 검증
* 최적 파라미터 확인

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

자동 생성된 설명

* 그리드 서치를 통해 알아낸 최적의 하이퍼파라미터 값은 C가 10, max\_iter가 100
* 최적의 하이퍼파라미터 값으로 모델 학습

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

자동 생성된 설명

**4. 평가**

**1) 모델 평가**

* 각 지표들을 활용하여 모델 평가 진행
* 평가 지표로는 정확도(accuracy), 정밀도(precision), 재현률(recall), F1 점수 및 ROC-AUC를 사용

1. 정확도(Accuracy): 전체 예측 중 올바르게 예측한 비율
2. 정밀도(Precision): 양성으로 예측한 것 중 실제 양성인 비율
3. 재현율(Recall): 실제 양성 중 양성으로 예측한 비율
4. F1-score: 정밀도와 재현율의 조화 평균
5. ROC-AUC: ROC 곡선 아래 면적, 클래스 구분 성능 평가, 1에 가까울수록 완벽한 구분 의미

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

자동 생성된 설명

**모델 평가 결과:**

* 정확도: 93.97%로 매우 높다.
* 정밀도와 재현율: 매우 높은 편으로 특히 클래스 1에 대해 매우 높은 재현율(0.99)을 보이고 있다.
* F1-점수는 93.45%로, 정밀도와 재현율 간의 균형이 잘 잡혀 있다.
* ROC-AUC 점수는 80%로, 클래스 구분 성능이 준수하다.
* 언더피팅(Underfitting): 현재 모델의 정확도와 F1-점수가 높은 것을 보아 언더피팅의 가능성은 낮다.
* 오버피팅(Overfitting): 현재 성능 평가 지표가 재현율뿐만 아니라 정밀도와 F1-점수도 높아서 전반적으로 모델이 균형 잡힌 성능이라고 볼 수 있다.
* 클래스 불균형 존재: 클래스 0과 클래스 1의 데이터 수가 차이가 크다(177 vs 1150). 이로 인해 모델이 클래스 1에 대해 더 높은 성능을 보인다.

**5. 시각화 및 해석**

**1) 시각화**

* confusion\_matrix(혼동 행렬)와 ROC 곡선을 활용하여 시각화

1. confusion\_matrix(혼동 행렬): 분류 모델의 성능을 평가하기 위해 사용되는 표, 실제 값과 예측 값을 비교하여 모델의 예측 능력을 시각화
2. ROC 곡선: 분류기의 성능을 평가하는 그래프, 다양한 임계값에서의 True Positive Rate(TPR)와 False Positive Rate(FPR)의 변화를 시각화

**confusion\_matrix(혼동 행렬) 분석:**

1. True Positive (0.99)
   * 긍정적인 리뷰를 보고 긍정적이라고 예측한 경우
   * 모델이 실제 양성인 경우를 99% 정확하게 예측
2. False Positive (0.39)
   * 부정적인 리뷰를 보고 긍정적이라고 예측한 경우
   * 모델이 실제 음성인 경우를 39%의 비율로 양성으로 잘못 예측, 이는 양성으로 잘못 분류되는 경우가 상대적으로 많음
3. True Negative (0.61)
   * 부정적인 리뷰를 보고 부정적이라고 예측한 경우
   * 모델이 실제 음성인 경우를 61% 정확하게 예측
4. False Negative (0.096)
   * 긍정적인 리뷰를 보고 부정적이라고 예측한 경우
   * 모델이 실제 양성인 경우를 9.6%의 비율로 음성으로 잘못 예측, 이는 양성 클래스를 놓치는 경우가 상대적으로 적음을 의미

**confusion\_matrix(혼동 행렬) 종합 평가:**

* 긍정적 리뷰 예측: 모델이 긍정적인 리뷰를 거의 정확하게 예측하는 능력이 뛰어나다.
* 부정적 리뷰 예측: 모델이 부정적인 리뷰를 긍정적으로 잘못 예측하는 비율이 상대적으로 높다.

**텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명**

**텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명**

**ROC 곡선 분석:**

* AUC(Area Under the Curve):
  + ROC 곡선 아래의 면적을 나타내며, 모델의 분류 성능을 한눈에 보여주는 지표
  + AUC 값이 1에 가까울수록 모델이 좋은 성능
  + 0.5 이하일 경우 모델이 오히려 잘못된 예측을 하는 모델

**ROC 곡선 종합 평가:**

* AUC 값이 0.80030이므로 모델델의 분류 성능이 좋음을 나타낸다. 모델이 긍정 클래스와 부정 클래스를 비교적 정확하게 구분하고 있음을 의미한다.

**텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명**

**텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명**

**2) 마무리**

* amazon\_uk\_shoes\_products\_dataset\_2021\_12 리뷰 데이터를 활용하여 감성 분석을 진행하였습니다. 그 과정에서 데이터 전처리(중복 데이터 제거, 주요 특성 설정, 결측치 확인 및 제거, 노이즈 제거, 토큰화, 불용어 제거, 표제어 추출) 작업이 진행되었고 NLTK의 VADER 감성 분석 모델을 활용하여 감성 분석을 진행하였습니다.
* 그리드 서치 기법을 활용하여 최적의 하이퍼파라미터를 찾아 로지스틱 회귀 모델 학습을 진행하였습니다. 이 모델은 긍정적인 리뷰를 탐지하는 데 매우 정확하지만, 부정적인 리뷰를 식별하는 데는 상대적으로 약한 성능을 보입니다. 만약 이 모델이 긍정적인 리뷰를 정확하게 식별하는 것이 중요한 모델이라면 매우 유용할 것이지만 부정적인 리뷰를 정확히 식별하는 것이 중요한 경우라면 부적합한 모델이 될 것입니다. 이는 모델 학습에 사용한 데이터가 부정적인 리뷰 데이터가 부족했던 원인이 크다고 생각되며 데이터 전처리 과정에서 영어가 아닌 단어를 영어를 판독하여 compound 점수가 낮게 나온 점도 작용하였다고 생각합니다.