스크린샷, 패브릭, 디자인이(가) 표시된 사진

자동 생성된 설명ㅡ

|  |
| --- |
| **기계학습 제품 review 데이터 감성 분석 보고서** |
| 스크린샷, 패브릭, 디자인이(가) 표시된 사진  자동 생성된 설명 |

텍스트, 원, 엠블럼, 로고이(가) 표시된 사진

자동 생성된 설명

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **과목** | **기계학습** |  | **담당교수** | **백우진 교수님** | |
| **이름** | **전동찬** |  | **제출일** | | **~2024-06-22** |

**목 차**

**1. 데이터 전처리**

1.1 데이터 전처리

1.2 모델 선택 및 학습

1.3 텍스트 정규화

**2. 감정분석모델구현**

**3. 모델 학습 및 튜닝**

3.1 데이터셋을 학습 및 테스트 세트로 분할

3.2  랜덤 서치

3.3 교차 검증

**4. 모델 평가**

4.1 정확도, 정밀도, 재현율, F1 점수 및 ROC-AUC와 같은 지표를 사용하여 모델 평가

4.2 오버 피팅과 언더 피팅을 논의하고 이를 해결하는 방법 학습

**5. 시각화**

5.1 Matplotlib 또는 Seaborn과 같은 라이브러리를 사용하여 결과 시각화

5.2 혼동 행렬 및 ROC 곡선 생성

1. 데이터 전처리
   1. 데이터 정리데이터 정리 (노이즈 제거, 결측값 처리 등).

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

데이터의 결측값을 해보니 helpful\_count, review\_title, review\_text 열에 결측값을 확인했습니다.

불필요한 열과 결측값을 제거하여 데이터의 결측값을 모두 처리했습니다.

* 1. 텍스트 토큰화 및 불용어 제거.

텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* 1. 텍스트 정규화 (스테밍 또는 표제어 추출).

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명

텍스트 정규화를 통해 스테밍, 표제어를 추출하고 따옴표 및 괄호를 제거했습니다.

1. **감정분석모델구현**

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

Review\_text를 감정 분석했습니다. 감정 분석 하는데 2..32초가 소모되었습니다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

감정 분석 결과를 확인했습니다. -1에 가까울수록 부정적인 감정을 나타내고, 1에 가까울수록 긍정적인 감정을 나타냅니다.

1. **모델 학습 및 튜닝:**

3-1. 학습 데이터와 테스트 세트 분할

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_tfidf, y, test\_size=0.2, random\_state=42)

3-2.랜덤서치

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

최적의 깊이: 7

최적의 트리 개수: 91

랜덤 서치 결과 점수는 0.7572

3-3. 교차검증

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

5회 교차검증 결과 평균: 0.7572점이 측정되었습니다.

1. **모델 평가**

4-1. 정확도, 정밀도, 재현율, F1 점수 및 ROC-AUC와 같은 지표를 사용하여 모델 평가.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

모델 평가 결과 🡪 정확도: 0.8158, 정밀도: 0.8160, 재현율: 0.6525, F1 점수: 0.8886,

ROC-AUC: 0.8978의 결과가 측정되었습니다.

4-2. 오버피팅과 언더피팅을 논의하고 이를 해결하는 방법 학습

텍스트, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

오버피팅: 모델이 학습 데이터에 너무 잘 맞춰져서 새로운 데이터에 대한 일반화 성능이 떨어지는 현상.

언더피팅: 모델이 학습 데이터의 패턴을 충분히 학습하지 못하는 현상.

학습데이터의 점수가 검증데이터의 점수보다 높은 오버피팅 그래프입니다.

학습데이터의 성능은 높지만 검증 데이터나 새로운 데이터에서 성능이 떨어질 수 있습니다.

1. **시각화**

5-1. Matplotlib 또는 Seaborn과 같은 라이브러리를 사용하여 결과 시각화.

텍스트, 스크린샷, 직사각형, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

Matplotlib라이브러리를 이용하여 정확도, 정밀도, 재현율, F1 점수, ROC-AUC 점수를 막대 그래프 형식으로 시각화 하였습니다.

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

Seaborn 라이브러리를 이용하여 review\_rating 분포도를 그렸습니다.

5-2. 혼동행렬 및 ROC 곡선

텍스트, 스크린샷, 도표, 다채로움이(가) 표시된 사진

자동 생성된 설명

혼동 행렬 생성 및 분석 결과

모델이 긍정적이라고 예측한 리뷰중 긍정적 리뷰 1001

모델이 부정적이라고 예측한 리뷰중 부정적 리뷰 111

모델이 긍정적이라고 예측한 리뷰중 부정적 리뷰 226

모델이 부정적이라고 예측한 리뷰중 긍정적 리뷰 25

로 분석되었습니다.

텍스트, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

ROC 곡선이 왼쪽 위에 가까울수록 모델의 분류 성능이 좋습니다. ROC 곡선이 이 대각선 위에 있으면 모델이 무작위 예측보다 성능이 좋다는 것을 의미합니다. 따라서ROC-AUC 값이 0.90으로 상당히 좋은 분류 성능을 보인다고 할 수 있습니다.

대부분의 평가가 높은 수준으로 나타났지만 재현율이 다른 평가와 비교해서 살짝 낮은 수치입니다.그렇지만 혼동 행렬 그래프가 정상적으로 출력되었습니다.

AUC 곡선 그래프 생성 결과 AUC의 값이 0.9로 성능 평가에서 긍정적인 결과를 확인했습니다.또한 모델 평가 결과 대부분의 수치에서 80% 이상의 수치가 측정되었기 때문에 분석 시스템의 성능이 긍정적이라고 말할 수 있습니다.