**기계학습(8585)**

**기말고사 대체 과제**

**감성 분석을 통한 제품 리뷰 분석**

201920829 금동호

**1. 감성 분석의 기본 이해**

감성 분석은 텍스트 데이터에서 감정이나 사용자의 의견을 식별하고 수치화하는 과정이다. 텍스트 데이터가 긍정적인이 부정적으로 분류하는 과정을 통해 사용자 의견을 파악하고 이를 바탕으로 서비스 개선이나 다수의 반응을 파악하는 등의 결과를 기대할 수 있다. 이러한 감성 분석은 텍스트 데이터에서 다양한 정보를 얻을 수 있고, 이를 바탕으로 의사 결정에 큰 도움을 주고 있다.

**2. 데이터 전처리**

**1) 데이터셋 준비**

지정된 데이터셋은 amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv이다. 주어진 파일을 불러와서 살펴보면 구성은 다음과 같다.

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

자동 생성된 설명**

**URL 주소, 실제 구매 여부, 리뷰 작성 날짜와 같은 불필요한 데이터들이 섞여있다. 따라서 불필요한 값들을 제거해줄 필요가 있다.**

**텍스트, 스크린샷, 시계, 폰트이(가) 표시된 사진

자동 생성된 설명**

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

자동 생성된 설명**

**위의 과정을 거쳐 필요하지 않은 데이터를 걸러냈다.**

**2) 데이터 전처리**

이어서 텍스트 데이터에 대해서 전처리 과정을 진행한다. 이는 노이즈 제거, 토큰화, 불용어 제거, 표제어 추출 순으로 진행하였다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

이어서 준비가 완료된 데이터의 리뷰 점수에 따른 빈도를 시각화를 진행하였다.

**텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명**

시각화 결과, 5점의 빈도가 다른 점수에 비해 월등히 높음을 알 수 있었다. 이를 통해 데이터의 구성이 결과에서 부정적인 평점에 대한 학습 결과에 영향을 줄 수 있음을 고려하고 진행하였다.

**3. 모델 선택 및 학습**

**1) 모델 선택**

수업에서 활용했던 NLTK의 VADER를 통해 감성 분석을 진행한다. 감성 분석을 통해 리뷰에 대한 텍스트 데이터를 긍정, 부정으로 이진화한다. 감성 분석의 결과는 -1 ~ 1 사이의 값을 가지며, 0.1이 넘으면 긍정적인 것으로 여긴다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

**2) 모델 학습**

모델 학습에는 선형 회귀모델을 활용한다. 이를 위해서 앞서 이진화한 text\_token을 벡터화한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

이어서 학습 데이터와 테스트 데이터를 분할하는데, 비율은 각각 8:2이다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

이후 적절한 하이퍼 파라미터를 튜닝하기 위해 그리드 서치를 진행한다. 이를 통해 학습에 있어서 최적의 하이퍼 파라미터를 구할 수 있다. 이 과정에서 교차 검증을 위한 fold 값을 5로 지정해준다.

텍스트, 스크린샷, 소프트웨어, 폰트이(가) 표시된 사진

자동 생성된 설명

이 과정을 통해 최적 하이퍼 파라미터 값이 10임을 알 수 있다. 이 값으로 모델을 학습한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

여기에서 Box Loss는 예측된 바운딩 박스와 실제 바운딩 박스 간의 위치와 크기의 차이를 의미한다. Object Loss는 객체가 존재할 확률에 대한 손실률을, Class Loss는 객체가 0, 1, 2의 label 중 어디에 속할지 예측하는데에서 오는 손실을 의미한다.

epoch이 증가할수록 모든 손실률이 낮아지고 있음을 알 수 있으며, 이는 모델의 예측이 더욱 정확해져가는 과정임을 보여준다.

**4. 모델객체 인식 및 시각화**

훈련 모델을 평가하는 지표를 살펴보자.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

정확도, 정밀도, 재현율, F1 점수, ROC-AUC 점수 모두 원활한 점수로 나타났음을 알 수 있다. 이를 바탕으로 오버피팅이나 언더피팅의 가능성은 낮다고 볼 수 있다.

**5. 시각화 및 해석**

**1) 혼돈 행렬**

텍스트, 스크린샷, 다채로움, 도표이(가) 표시된 사진

자동 생성된 설명

혼돈 행렬을 시각화한 그래프를 살펴보면, 긍정 리뷰에 대해서는 719:57의 정확도를, 부정 리뷰에 대해서는 540:47의 정확도를 보인다. 이를 고려하면 해당 학습이 원활히 진행되었음을 알 수 있다.

**2) ROC 곡선**

마지막으로 ROC 곡선을 살펴보자.

텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

ROC 곡선은 왼쪽 모서리에 그래프가 가까울수록 학습이 잘 이루어졌다고 여긴다. 또한 0.5부터1의 값을 갖는 ROC-AUC의 값은 1에 가까울수록 학습이 잘 이루어졌다고 본다. 해당 ROC-AUC 값은 약 0.92이다. 이러한 점을 고려하면 해당 학습 과정이 원활히 이루어졌음을 알 수 있다. 따라서 학습 과정 중 언더피팅이나 오버피팅의 가능성은 낮다고 볼 수 있겠다.