**기계학습 보고서**

**제품 review 데이터의 감성 분석**

달력이(가) 표시된 사진

자동 생성된 설명

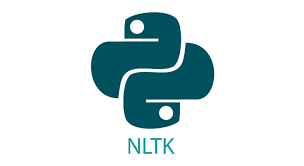
|  |  |  |  |
| --- | --- | --- | --- |
| 제출일 | 2024.06.21 | 학과 | 컴퓨터공학과 |
| 과목 | 기계학습(8585) | **학번** | 201921064 |
| 담당 교수님 | 백우진 교수님 | **이름** | 최요한 |

**목차**

1. 감정 분석의 기본 이해
2. 데이터 전처리
3. 모델 선택 및 학습
4. 평가
5. 시각화 및 해석

**1. 감성 분석의 기본 이해**

텍스트 내용에서 사람의 감정, 의견, 태도를 식별하고 추출하는 기술입니다. 이 기술은 다양한 소스의 텍스트 데이터에서 주관적인 정보를 자동으로 분석하는 데 사용됩니다.



감정분석은 **NLTK**의 **VADER**를 통해 진행됩니다.

**특성 추출**:

* + 감정 분석을 수행하기 위해 텍스트 데이터에서 유용한 특성을 추출합니다. 가장 기본적인 방법은 Bag of Words (BoW) 모델을 사용하는 것이며, TF-IDF(Term Frequency-Inverse Document Frequency) 점수를 사용하여 텍스트 내의 각 단어가 갖는 중요도를 계산할 수 있습니다.

**감정 분석 모델 사용**:

* + NLTK는 **VADER**라는 감정 분석 도구를 제공합니다.
  + VADER가 제공하는 점수를 바탕으로 텍스트가 긍정적인지, 부정적인지, 아니면 중립적인지를 판단합니다. 점수는 일반적으로 -1에서 1 사이의 값으로, 프로젝트에서는 **0.2**를 기준으로 긍정과 부정이 나뉩니다.

이번 프로젝트에서는 로지스틱 회귀를 사용하기 때문에 이진 분류를 위해 긍정적, 부정적 감정으로만 분류합니다.

텍스트, 폰트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

**2. 데이터 전처리**

프로젝트에서 사용된 데이터셋은 **‘wine\_review.csv’**입니다.

아래 데이터는wine\_review.csv의 항목 중 감정과 가장 밀접한 관련이 있어보이는 text와 rating열입니다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

**didPurchase**와 **doRecommend**의 항목도 있었지만 결측값이 너무 많아 머신러닝 모델을 학습하기에는 매우 적고 후에 데이터 증강을 하더라도 이러한 결측값을 모두 제거한다면 데이터의 원본이 매우 적게 남아 배보다 배꼽이 커지는 상황이 발생할 것으로 우려되기 때문에 데이터에 포함시키지 않았습니다.



텍스트, 영수증, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

따라서 다음과 같이 **rating**, **text**, **title** 3개의 열만 사용하였습니다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

또한 데이터를 확인해보니 아래 그림과 같이 리뷰 점수가 5점에 집중되어 있는 모습입니다.

텍스트, 스크린샷, 디스플레이, 도표이(가) 표시된 사진

자동 생성된 설명텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

이렇게 데이터가 5점에만 치중되어 있다면 학습이 제대로 이루어지지 않고, 긍정적 감정이라고만 판정을 내려도 정밀도가 높아질 수 있을 것 같아 데이터를 더 처리한 후에 **데이터 증강**을 통해 표본의 수를 늘려주었습니다.

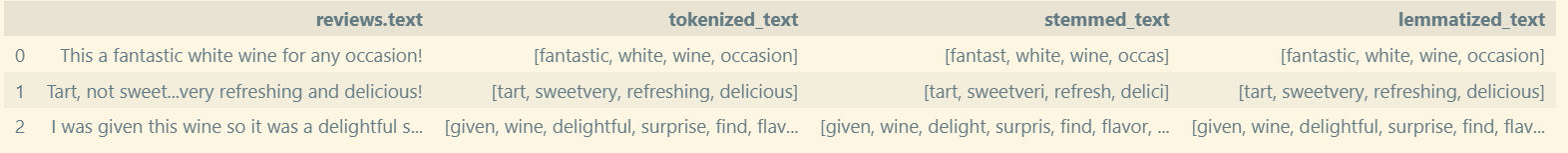
아래 코드의 **re(Regular Expression)**을 통해 추가 공백과 특수 문자를 제거하고 소문자로 변화시켜 영문자 중에서도 소문자만 남기게 하였습니다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

토큰화 -> 스테밍 -> 표제어 추출의 과정을 담았습니다.



텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

후에 제목도 표제어 추출을 해준 뒤 제목과 내용은 감정 상태가 크게 다르지 않을 것이라고 판단하여, 하나의 **combined\_text**로 묶어주었습니다.

이제 데이터들은 표제어 추출이 되어있고, 제목과 내용이 합쳐져 있습니다. 이 시점에서 데이터 증강을 진행해줍니다.

텍스트, 스크린샷, 문서, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명텍스트, 폰트, 스크린샷, 영수증이(가) 표시된 사진

자동 생성된 설명

리뷰의 수를 3000개씩으로 증강해준 모습입니다. 이제 데이터가 고르게 있어 학습하는 데 문제가 없을 것으로 보입니다.

**3. 모델 선택 및 학습**

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명감성 분석을 위해 NLTK 라이브러리의 VADER 모델을 선택하였습니다.

감성 점수(compound)를 -1부터 1 사이로 제공합니다.



0.2를 기준으로 **긍정적 감정**과 **부정적 감정**으로 감정 이진화를 진행하였고, 로지스틱 회귀 모델을 사용하여 감성 분류를 학습하였습니다.

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명



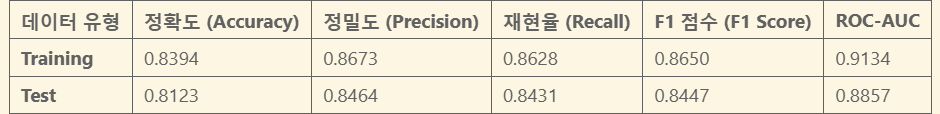
총 15000개 중 **80%**의 데이터를 훈련용으로 사용하였습니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

그리드 서치를 통해 하이퍼파라미터를 튜닝했습니다.

**4. 평가**

****

모델의 성능은 정확도, 정밀도, 재현율, F1 점수, ROC-AUC로 평가하였습니다.

* **Training 데이터**에서의 정확도는 0.8394, 정밀도는 0.8673, 재현율은 0.8628, F1 점수는 0.8650, **ROC-AUC는 0.9134**로 나타났습니다.
* **Test 데이터**에서의 정확도는 0.8123, 정밀도는 0.8464, 재현율은 0.8431, F1 점수는 0.8447, **ROC-AUC는 0.8857**로 나타났습니다.
* 전반적으로, Training 데이터와 Test 데이터 간의 성능 지표가 유사하여 모델이 비교적 잘 일반화된 것으로 보입니다.

**오버피팅과 언더피팅**

* Training 데이터와 Test 데이터의 성능 지표가 유사하므로, 모델이 훈련 데이터에 과도하게 적합되지 않은 것으로 보입니다. 따라서 오버피팅의 가능성은 낮습니다.
* 반면, Training 데이터와 Test 데이터 모두에서 높은 성능을 유지하고 있어, 모델이 충분히 복잡한 패턴을 학습하고 있는 것으로 보입니다. 이는 언더피팅의 가능성도 낮다는 것을 의미합니다.

전반적으로 모델은 좋은 성능을 보이며, 과적합이나 과소적합의 문제를 크게 보이지 않습니다.

특히 ROC-AUC 점수는 0.89로, 우수한 예측 성능을 보여주었습니다.

**5. 시각화 및 해석**

모델의 예측 결과를 혼동 행렬로 시각화하여 긍정적 및 부정적 리뷰의 분류 정확도를 분석하였습니다. 이를 통해 모델의 강점과 약점을 더 명확하게 이해할 수 있었습니다.

스크린샷, 텍스트, 다채로움, 직사각형이(가) 표시된 사진

자동 생성된 설명

**혼동 행렬의 시각적 해석**

* 이 혼동 행렬을 통해 볼 때, 모델은 긍정적 감정을 식별하는 데 있어 상당히 효과적입니다(True Positive가 높음).
* 그러나 여전히 상당수의 부정적 사례를 긍정으로 잘못 분류하고 있음(False Positive)을 볼 수 있습니다.
* 이는 모델이 특정 조건이나 특성에서 오류를 범할 수 있음을 암시합니다.
* False Negative의 수치 또한 상대적으로 낮지만, 이를 줄이는 것도 중요하므로 모델의 센서티비티(민감도)를 향상시키는 전략을 고려해야 할 수 있습니다.

종합하면, 모델은 전반적으로 긍정적 사례에 대한 예측에 강점을 보이지만, 일부 부정적 사례에 대해서는 오류를 범할 가능성이 있으므로 추가적인 조정과 분석이 필요합니다. 이러한 결과는 모델이 직면할 수 있는 편향이나 한계를 이해하는 데 도움이 됩니다.

텍스트, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

**ROC 곡선 해석**

* **곡선 형태**: ROC 곡선은 왼쪽 상단 모서리를 향해 굽어 있습니다. 이는 모델이 임의 추측보다 훨씬 나은 성능을 보이고 있음을 의미합니다.
* **AUC (Area Under the Curve) 값**: AUC 값은 0.88로, 이는 모델이 88%의 확률로 무작위로 선택된 양성 샘플을 무작위로 선택된 음성 샘플보다 높은 점수를 할당할 것임을 의미합니다.

**아쉬운 점**

**데이터 불균형**

* 리뷰 점수의 빈도가 높은 점수(예: 5점)에 집중되어 있어, 특정 점수에 대한 모델의 예측이 더 잘 이루어질 가능성이 큽니다.
* 데이터 증강을 통해 원래 데이터의 5배에 가까운 데이터를 생성하였기 때문에 이에 유의할 필요가 있습니다.

**오류 분석의 부족**

* 혼동 행렬에서 확인할 수 있듯이, 여전히 False Positive와 False Negative가 존재합니다.
* 이러한 오류가 발생하는 패턴을 분석하고, 이에 대한 개선 방안을 모색할 필요가 있습니다.

**모델의 단순성**

* Logistic Regression 모델을 사용하여 이진분류를 하였습니다.
* 리뷰의 평점이 1점부터 5점까지 있기에 범주형으로 자세히 나눠 더 복잡한 모델 (예: Random Forest, XGBoost 등)과 비교하여 성능을 평가해 볼 수도 있습니다.

**하이퍼파라미터 튜닝**

* 그리드 서치나 랜덤 서치 등을 통한 하이퍼파라미터 최적화가 어느 정도 이루어졌지만, 더 많은 파라미터를 조정하거나 더 많은 교차 검증을 수행하여 최적의 모델을 찾을 수 있습니다.

**추가적인 특성 사용의 부족**

* **reviews.getPurchase**와 **reviews.numHelpful**을 통해 결제를 하거나, 도움이 되었다는 리뷰를 통해 가중치를 더하거나 데이터 증강(Data Augmentation)을 통해 데이터를 처리하고 싶었는데 그러지 못했던 점이 아쉬웠습니다.

**결론**

모델은 전반적으로 좋은 성능을 보이며, 특히 긍정적인 리뷰 예측에서 일관성을 보여주고 있습니다. 하지만 데이터 불균형, 오류 분석의 부족, 더 복잡한 모델의 필요성 등 몇 가지 아쉬운 점을 보완함으로써 모델의 성능을 더욱 향상시킬 수 있습니다. 이를 통해 와인 리뷰 분석의 정확성과 신뢰성을 높일 수 있을 것입니다.

**출처**

**chatGPT**

**nltk -** [[Text Mining][NLP] 자연어 처리를 위한 NLTK (설치 방법 및 사용 문법) (tistory.com)](https://ysyblog.tistory.com/87)

**vader -** [Amazon.com: STAR WARS The Vintage Collection Darth Vader (The Dark Times) Toy, 3.75-Inch-Scale OBI-Wan Kenobi Figure, Toys Kids Ages 4 and Up : Toys & Games](https://www.amazon.com/STAR-WARS-Vintage-Collection-3-75-Inch-Scale/dp/B09H1B2G5Q)