**REPORT**

제품 review 데이터의 감성 분석



|  |  |
| --- | --- |
| **과목명** | 기계학습(8585) |
| **학과** | 컴퓨터공학과 |
| **담당교수** | 백우진 교수님 |
| **학번 이름** | 202121873 황호준 |
| **제출일자** | 2024.06.21 |

1. **감정 분석의 기본 이해**

감정 분석(Sentiment Analysis)은 텍스트 데이터에서 감정이나 의견을 식별하여 메시지의 어조가 긍정적인지, 부정적인지, 또는 중립적인지 분류하는 자연어 처리(NLP) 기술입니다. 이를 통해 텍스트의 전체적인 분위기나 감정적 톤을 파악할 수 있습니다. 감정 분석은 리뷰, 설문 조사, 소셜미디어 등 다양한 분야에서 활용되며 기업은 감정 분석을 활용하여 고객 서비스를 개선하고 브랜드 평판을 높입니다. 해당 프로젝트에서는 아마존의 신발 리뷰 데이터를 사용하여 감정 분석을 수행합니다.

1. **데이터 전처리**

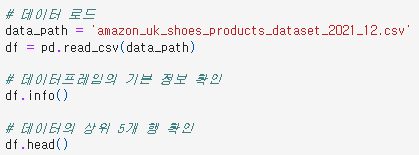
데이터 전처리는 감정 분석 모델의 성능을 좌우하는 중요한 단계입니다. 홀수 학번은 amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv의 데이터를 사용하여 감정 분석을 수행함에 따라 데이터 로드, 결측치 제거, 데이터 프레임 정리, 노이즈 및 불용어 제거, 텍스트 토큰화 등의 전처리 과정을 거쳤습니다.

1. **필요 라이브러리 설치 및 임포트**



데이터 전처리에 필요한 라이브러리를 설치 및 임포트합니다.

1. **데이터 로드**

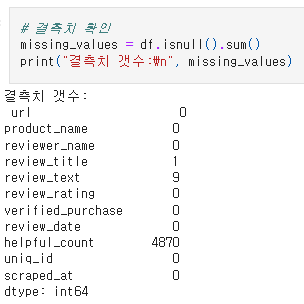


라이브러리의 설치 및 임포트가 끝나면 데이터의 구조를 확인하기 위해 데이터 로드를 진행합니다. 이 과정에서는 데이터의 구조와 각 열의 타입을 파악하며, 데이터의 상위 5개 행을 출력하여 데이터의 샘플을 확인합니다.

로드된amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv의 데이터 구조를 확인해보면 다음과 같은 11가지의 열을 확인할 수 있습니다.

1. url
2. product\_name
3. reviewer\_name
4. review\_title
5. review\_text
6. review\_rating
7. verified\_purchase
8. review\_date
9. helpful\_count
10. uniq\_id
11. scraped\_at
12. **결측치 제거**

결측치란 데이터셋에서 값이 누락된 경우를 의미합니다. 결측치는 다양한 이유로 발생할 수 있으며, 데이터 수집 과정에서 누락되거나 잘못 입력된 경우가 대표적입니다. 결측치가 존재하면 데이터 분석과 모델링의 정확도와 신뢰성을 저해할 수 있기 때문에 이를 처리하는 것은 매우 중요합니다.



* **결측치 제거 전**

먼저 데이터프레임의 각 열에 결측치가 얼마나 있는지 확인합니다. 해당 결과에서 ‘helpful\_count’ 열에 많은 결측치가 존재하며, ‘review\_title’과 ‘review\_text’에도 일부 결측치가 존재하는 것을 확인할 수 있습니다. 결측치가 너무 많은 열은 학습에 유용하지 않다고 판단하여 제거하고, 남아있는 열에서 결측치가 존재하는 행을 제거합니다.

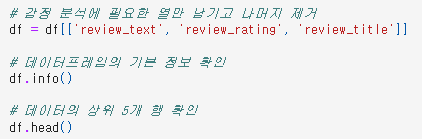


* **결측치 제거 후**

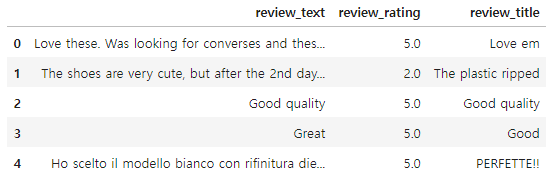
결측치 제거 후, 데이터프레임의 각 열에 남아있는 결측치를 다시 확인하여 모든 결측치가 제거되었는지 확인합니다. 이를 통해 분석과 모델링에 사용할 수 있는 완전한 데이터를 확보할 수 있습니다.

1. **데이터 프레임 정리**

감정 분석에 필요한 열만 남기고 나머지 불필요한 열을 제거하여 데이터 프레임을 정리하는 과정입니다. 이를 통해 분석의 집중도를 높입니다.



감정 분석을 위해서는 주로 텍스트 데이터와 감정 레이블이 필요합니다. 해당 프로젝트에서는 'review\_text', 'review\_rating', 'review\_title' 열이 감정 분석에 직접적으로 의미가 있다고 판단하였고, 따라서 이 세 가지 열만 남기고 나머지 열은 제거합니다.

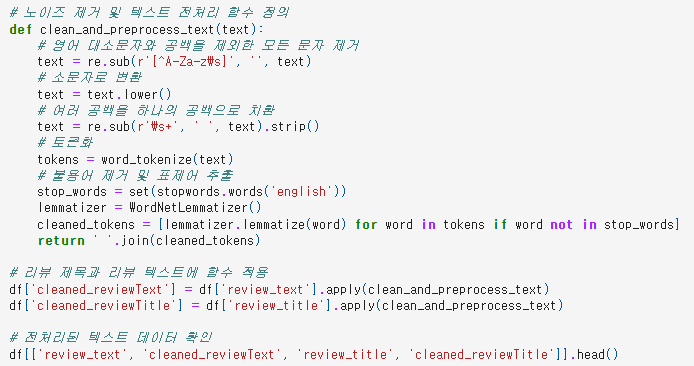


* **감정분석에 직접적으로 필요한 열만 남기고 정리된 데이터 프레임**

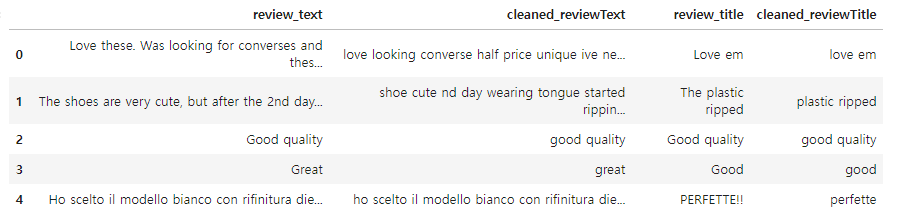
제거 후 데이터프레임의 기본 정보를 다시 확인하여 데이터 프레임의 구조와 누락된 데이터의 유무를 다시 한 번 점검합니다.

1. **노이즈, 불용어 제거 및 토큰화 진행**

이 과정에서는 텍스트 데이터를 분석에 적합한 형태로 변환하여 모델의 성능을 최적화하는 과정을 거칩니다.

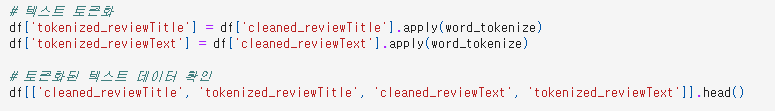


먼저 텍스트 데이터를 정제하기 위한 함수를 정의합니다. 해당 함수는 텍스트에서 불필요한 문자와 노이즈를 제거하고, 소문자로 변환한 후, 불용어를 제거하고 표제어 추출을 통해 텍스트를 정제합니다. 이후 정의된 전처리 함수를 리뷰 제목과 리뷰 텍스트에 적용하여 텍스트 데이터를 정제하고 확인합니다.

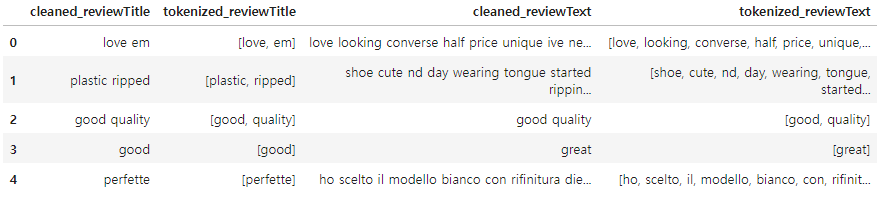


* **정제된 텍스트 데이터 확인**

전처리가 완료된 텍스트 데이터를 확인하여 정제가 올바르게 수행되었는지 검증합니다. 이를 통해 원본 텍스트와 정제된 텍스트를 비교할 수 있습니다.



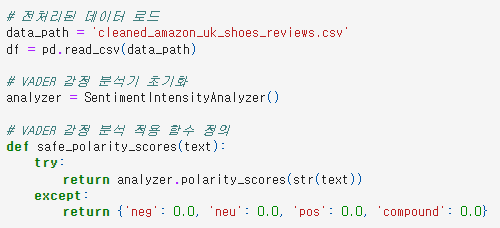
이후 정제된 텍스트 데이터를 분석에 사용할 수 있도록 토큰화를 진행합니다. 토큰화는 텍스트를 개별 단어 단위로 나누는 과정입니다.



* **토큰화된 텍스트 데이터 확인**

1. **모델 선택 및 학습**
2. **VADER 모델 활용 감정 분석 진행**

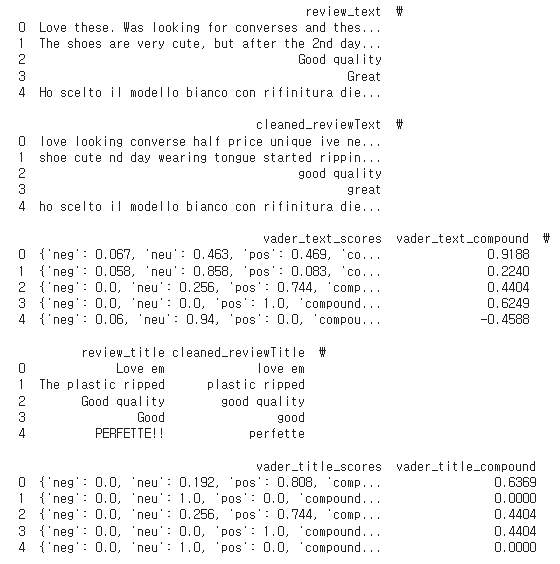
이번 단계에서는 VADER(Valence Aware Dictionary and sEntiment Reasoner) 모델을 사용하여 리뷰 데이터의 감정을 분석합니다. VADER는 소셜 미디어 텍스트와 같은 짧고 비공식적인 문서에서 감정을 분석하는 데 적합한 모델입니다. 해당 프로젝트에서는 VADER를 사용하여 전처리된 리뷰 텍스트와 제목에서 감정 점수를 계산하고, 이를 이진 값으로 변환하여 긍정과 부정을 분류합니다.



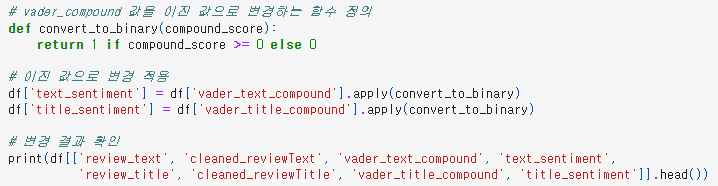
먼저 전처리된 데이터를 로드하고 VADER 감정 분석기를 초기화합니다. 이후 VADER 감정 분석을 적용할 함수를 정의합니다. 이 함수는 텍스트를 문자열로 변환한 후 감정 점수를 계산하고 예외가 발생하면 모든 감정 점수를 0.0으로 설정합니다.



정의된 함수를 사용하여 전처리된 리뷰 텍스트와 제목에 대해 VADER 감정 분석을 적용합니다. 각 리뷰에 대해 긍정적, 중립적, 부정적 감정 점수와 종합 점수를 계산합니다.



* **각 열에 대한 VADER 감정 분석 점수 결과**

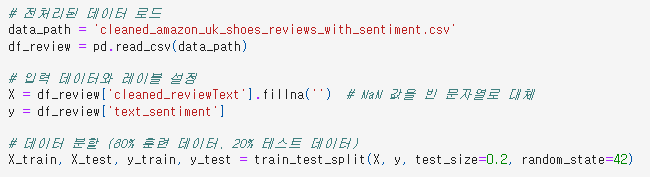
****

VADER 감정 분석의 종합 점수를 이진 값으로 변환하여 긍정(1) 과 부정(0)으로 구분합니다. 종합 점수가 0 이상이면 긍정으로 간주하여 1을 반환하고, 0 미만이면 부정으로 간주하여 0을 반환하도록 합니다.

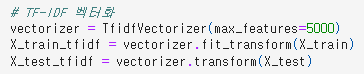


* **종합 점수를 이진 값으로 변환**

1. **데이터 레이블 설정 및 훈련 테스트 데이터 분할**



로지스틱 회귀모델을 사용하여 감정 분석 모델을 학습시키기 위해, 먼저 훈련 데이터와 테스트 데이터로 분할하고, 80%를 훈련 데이터로 20%를 테스트 데이터로 사용합니다.



로지스틱 회귀모델은 숫자 데이터를 입력으로 받기 때문에, TF-IDF 벡터화를 사용하여 텍스트 데이터를 수치 데이터로 변환하는 과정을 거칩니다. TF-IDF 벡터화는 텍스트 데이터에서 단어의 빈도를 계산하고, 단어의 중요도를 고려하여 가중치를 부여합니다. max\_features=5000은 벡터화 과정에서 최대 5000개의 특징을 사용할 것을 지정합니다. 너무 많은 특징을 사용하면 모델이 복잡해지거나 커널이 다운될 수 있고 너무 적으면 중요한 정보를 놓칠 수 있어 5000으로 설정하였습니다.

1. **Logistic Regression 모델 초기화 및 그리드 서치 활용**



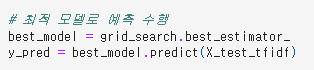
로지스틱 회귀모델을 초기화하고 max\_iter=1000은 모델이 수렴할 때 까지 최대 1000번의 반복을 허용하는 것을 의미합니다. 하이퍼파라미터 그리드 설정에서는 그리드 서치에서 탐색할 하이퍼파라미터의 범위를 설정합니다. 여기서는 ‘C’, ‘solver’, ‘penalty’ 세 가지 하이퍼파라미터의 다양한 값을 탐색합니다. 이후 그리드 서치를 통해 최적의 하이퍼파라미터를 찾고, 교차 검증을 통해 모델의 성능을 평가합니다. cv=5는 5-폴드 교차 검증을 의미하며, scoring=’accuracy’는 모델의 정확도를 기준으로 최적의 하이퍼 파라미터를 찾음을 의미합니다. 이후 그리드 서치를 통해 찾은 최적의 하이퍼파라미터와 최고 교차 검증 점수를 출력합니다.



* **그리드 서치로 찾아낸 최적의 하이퍼파라미터 및 교차 검증 최고점수**

여기서 ‘C=10’, ‘penalty=’l1’’, ‘solver=’liblinear’’의 조합이 최적의 하이퍼파라미터로 선택되었고, 교차 검증의 최고 점수는 94.53%로 나타났습니다. 이는 모델이 훈련 데이터에서 매우 높은 성능을 보였음을 의미합니다.

1. **최적 모델 활용 학습 및 예측**

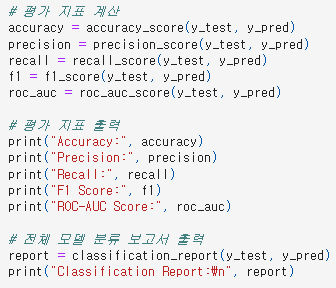


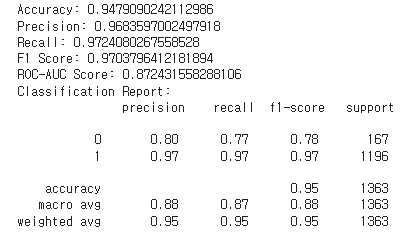


그리드 서치를 통해 찾은 최적의 모델을 ‘best\_estimator\_’ 속성에서 가져와 테스트 데이터를 예측하고, 예측 결과를 확인합니다. 예측 결과는 이진 값으로 1은 긍정적인 감정을, 0은 부정적인 감정을 나타냅니다.

1. **평가**

모델의 성능을 평가하기 위해 정확도(accuracy), 정밀도(precision), 재현율(recall), F1 점수, ROC-AUC 점수를 사용하여 모델의 성능을 평가합니다. 이러한 지표들은 모델이 예측한 결과와 실제 결과를 비교하여 얼마나 잘 맞추었는지를 나타냅니다.





**정확도(Accuracy)**: 94.79%로, 모델이 전체 데이터 중 약 95%를 정확하게 예측했음을 의미합니다.

**정밀도(Precision)**: 96.84%로, 모델이 긍정으로 예측한 데이터 중 약 97%가 실제로 긍정임을 의미합니다.

**재현율(Recall)**: 97.24%로, 실제 긍정인 데이터 중 약 97%를 모델이 정확하게 긍정으로 예측했음을 의미합니다.

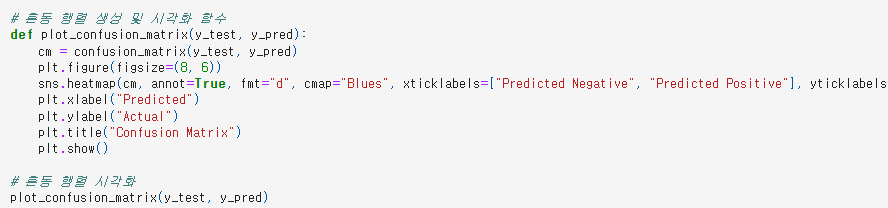
**F1 점수(F1 Score)**: 97.04%로, 정밀도와 재현율의 균형을 잘 맞춘 점수를 나타냅니다.

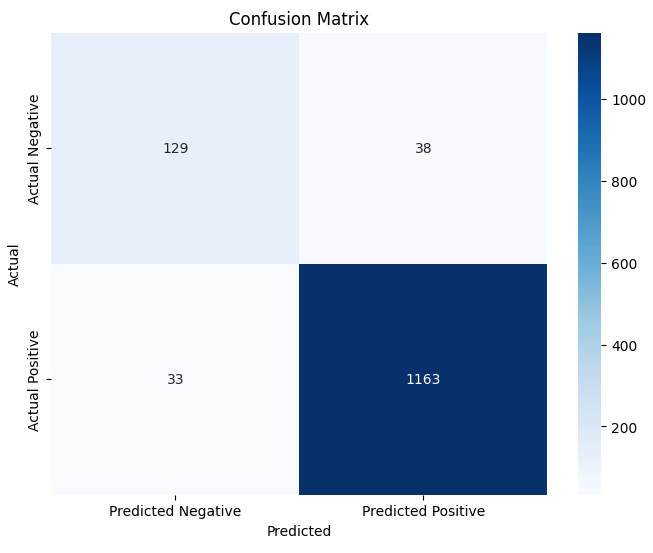
**ROC-AUC 점수(ROC-AUC Score)**: 0.872로, 모델이 긍정과 부정을 잘 구분하고 있음을 의미합니다.

모델은 전체적으로 높은 정확도, 정밀도, 재현율, F1 점수에서 우수한 성능을 보여주고 있습니다. 이 결과를 통해 보았을 때, 언더피팅의 문제는 없다고 판단할 수 있습니다. 오버피팅의 경우 최고 교차 검증 점수 0.9453와 테스트 데이터 점수 정확도 0.9479를 비교해보았을 때, 두 값이 유사하다고 볼 수 있기 때문에 학습 데이터에 대해서만 잘 맞추어진 것이 아니라, 테스트 데이터에 대해서도 잘 일반화되었음을 나타냅니다.

추가적으로, 긍정 감정(1)에서 매우 높은 성능을 기록하였으며, 부정 감정(0)에 대해서는 다소 낮은 성능을 보이고있습니다. 이는 데이터 불균형 문제로 성능이 낮게 나왔을 것으로 생각되며, 부정 리뷰에 대한 샘플 수를 늘리거나 샘플링 기법을 사용하여 불균형 문제를 해결할 수 있습니다.

1. **시각화 및 해석**
2. **혼동행렬 생성 및 시각화**

****

****

**True Negative (TN): 129** - 모델이 실제 부정을 부정으로 올바르게 예측한 경우입니다.

**False Positive (FP): 38** - 모델이 실제 부정을 긍정으로 잘못 예측한 경우입니다.

**False Negative (FN): 33** - 모델이 실제 긍정을 부정으로 잘못 예측한 경우입니다.

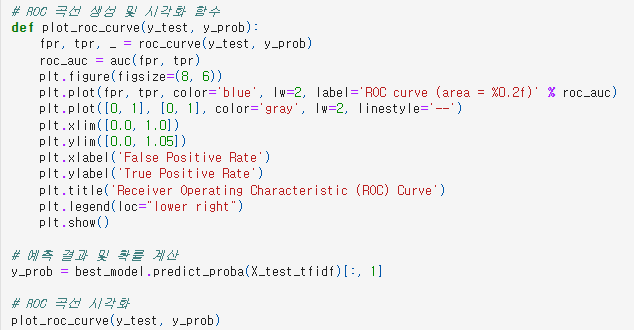
**True Positive (TP): 1163** - 모델이 실제 긍정을 긍정으로 올바르게 예측한 경우입니다.

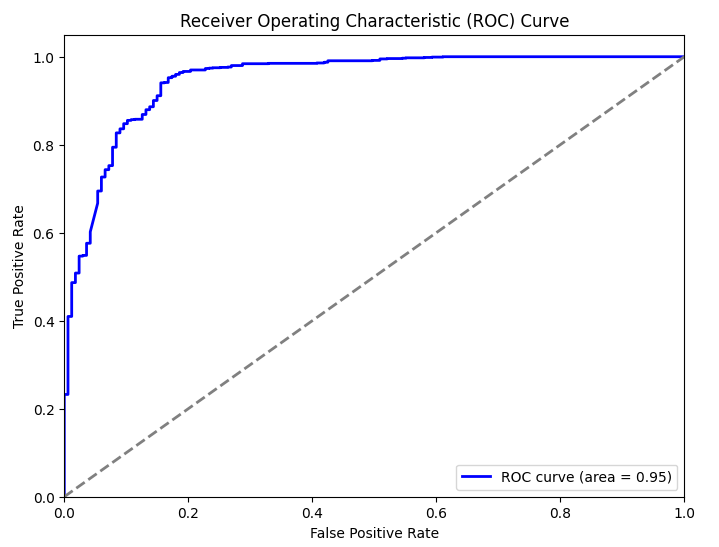
모델은 1163개의 실제 긍정을 정확하게 예측하였으며, 이는 모델의 높은 재현율과 정밀도로 이어집니다.

부정 샘플에 대해서는 129개의 실제 부정을 정확하게 예측하였으나, 38개의 부정을 긍정으로 잘못 예측했습니다.

전반적으로 혼동 행렬을 통해 모델이 긍정 샘플을 잘 예측하고 있음을 알 수 있습니다. 부정 샘플에 대한 예측 성능을 향상시키기 위해 추가적인 조치가 필요할 수 있습니다.

1. **ROC 곡선 생성 및 시각화**

****

****

ROC 곡선의 아래 면적 (AUC, Area Under Curve)은 0.95로, 이는 모델의 분류 성능이 매우 뛰어남을 나타냅니다. 대각선의 회색 점선은 랜덤 분류기를 나타내며, ROC 곡선이 이 선에서 멀어질수록 모델의 성능이 좋음을 의미합니다.

ROC 곡선의 AUC 값이 0.95라는 것은 모델이 실제 긍정과 부정을 잘 구분하고 있음을 나타내며, FPR이 낮고 TPR이 높은 이상적인 ROC 곡선을 보여주고 있다고 볼 수 있습니다.

1. **결론**

이번 프로젝트에서는 아마존 영국 신발 리뷰 데이터를 이용하여 감정 분석을 수행하였습니다. 평가 지표들을 통해 모델의 예측 성능을 평가한 결과, 전반적으로 우수한 성능을 보였음을 알 수 있었습니다.

그럼에도 불구하고 아쉬운 점은 부정 리뷰 예측 성능입니다. 혼동 행렬을 통해 확인한 결과, 부정 리뷰에 대한 예측 성능이 긍정 리뷰에 비해 다소 낮았습니다. FP와 FN이 존재하여 부정 리뷰 예측 성능을 향상시킬 필요가 있었습니다. 또한, 긍정 리뷰의 수가 부정 리뷰에 비해 많아 데이터 불균형이 존재했습니다. 이는 모델이 긍정 리뷰에 더 최적화될 가능성이 있어 부정 리뷰의 예측 성능이 낮아질 수 있습니다.

예측 성능을 향상시킬 수 있는 주요 방법은 다음과 같습니다. 데이터 불균형을 해결하기 위해 SMOTE와 같은 오버샘플링 기법을 사용하거나, 여러 모델을 앙상블하여 성능을 향상시킬 수 있습니다. 이 외에도 더 정교한 하이퍼파라미터 튜닝 기법을 사용하면 모델의 성능을 최적화할 수 있을 것 입니다.