**데이터 수집 및 전처리**

본 분석에 사용된 데이터는 amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv파일로부터 로드되었습니다. 데이터 전처리 과정은 다음과 같습니다:

**결측값 처리**: 리뷰 텍스트나 기타 중요한 열에 결측값이 있는 데이터 행을 제거하여 데이터의 완전성을 확보했습니다.

**텍스트 정제**: 각 리뷰에서 HTML 태그, 특수 문자, 숫자를 제거하고 모든 텍스트를 소문자로 **변환하여 표준화했습니다.**

**불용어 제거 및 표제어 추출**: NLTK 라이브러리를 활용하여 영어 불용어를 제거하고, 단어의 표제어를 추출하여 텍스트 데이터의 일관성을 높였습니다.

**감성 분석**

감성 분석을 위해 NLTK의 VADER 도구를 사용했습니다. VADER는 각 리뷰 텍스트의 긍정성, 부정성, 중립성을 수치로 평가하며, 이를 통해 다음과 같은 레이블을 할당했습니다:

**긍정적**: 감성 점수가 0.05 이상

**부정적**: 감성 점수가 -0.05 이하

**중립적**: 그 외의 경우

**모델링 및 예측**

분석을 위해 로지스틱 회귀 모델을 사용했으며, TF-IDF 벡터라이저를 통해 텍스트를 수치 데이터로 변환했습니다. 모델의 성능을 최적화하기 위해 그리드 검색을 통한 하이퍼파라미터 튜닝을 수행했습니다. 최적의 하이퍼파라미터는 C=1, solver='liblinear'로 결정되었습니다.

**결과 및 분석**

**성능 평가**: 모델의 정확도는 88%로 측정되었습니다. 또한, 5-폴드 교차 검증을 통해 모델의 일관성을 확인했으며, 평균 교차 검증 정확도는 87%였습니다.

**혼동 행렬**: 혼동 행렬 분석을 통해 모델이 긍정적 리뷰를 부정적 리뷰보다 잘 예측하는 경향을 확인했습니다.

**ROC 곡선 및 AUC**: 각 클래스에 대한 ROC 곡선을 시각화하여 모델의 성능을 평가했습니다. 긍정적, 부정적, 중립적 클래스에 대한 AUC는 각각 0.95, 0.85, 0.90으로 측정되었습니다.