202020986 이승진

1. 감정 분석의 기본 이해

감성 분석은 텍스트 데이터를 분석하여 그 안에 담긴 감정을 파악하는 자연어 처리(NLP) 기술입니다. 주로 긍정, 부정, 중립과 같은 감정 범주로 분류합니다.

자연어 처리를 이용한 감정 분석은 다양한 분야에서 활용될 수 있습니다.

마케팅: 고객 리뷰나 눈을 분석하여 제품에 대한 반응 파악

금융: 금융 뉴스나 댓글을 분석하여 시장 동향 예측

1. 데이터 전처리

데이터 전처리는 머신 러닝 모델의 성능에 큰 영향을 미칩니다. 텍스트 데이터의 경우, 전처리를 통해 노이즈를 제거하고, 모델이 학습할 수 있는 형태로 변환해야 합니다.

1. 모델 선택 및 학습

VADER는 소셜 미디어 텍스트와 같이 짧은 텍스트의 감성 분석에 특히 유용한 도구입니다.

짧은 텍스트를 분석하기 위한 도구이기 때문에 복잡한 모델 학습 과정이 필요 없고 빠르고 효율적이며 결과를 쉽게 해결할 수 있다는 장점이 있습니다.

하지만 같은 이유로 세밀한 감정 변화나 복합 감정을 정확하게 잡아내기 어렵습니다. 또 복잡한 문장 구조나 영어가 아닌 데이터에 대해서는 쉽게 사용하기 어렵다는 단점이 있습니다.

1. 평가

정확도: 0.9067524115755627

정밀도: 0.9061488673139159

재현율: 1.0

F1 점수: 0.9507640067911715

ROC-AUC: 0.532258064516129

정확도, 정밀도, 재현율, F1 점수 모두 높은 값을 가지고 있지만

ROC-AUC 값이 낮습니다.

이렇게 낮은 이유는 데이터셋에 양성 클래스가 매우 많고 그에 비해 음성 클래스는 매우 적습니다.

1. 시각화 및 해석

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

모델이 긍정적인 것은 예측은 잘 하지만 부정적인 것은 개선이 필요합니다.   
데이터셋에 긍정적인 클래스가 매우 많고 그에 비해 부정적인 클래스는 매우 적습니다.

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

ROC 곡선

모델의 성능이 그렇게 좋다고 볼 수 없습니다.

이는 오버피팅이 있기 때문입니다. 모델이 학습 데이터에 너무 잘 맞추어져서 새로운 데이터에 대해 일반화 능력이 떨어지는 현상입니다. 이로 인해 모델의 성능이 학습 데이터에서는 매우 높지만, 실제 사용 시에는 성능이 저하됩니다.

오버피팅이 발생한 이유는 여럿 있을 수 있으나 저의 모델에서는 데이터의 양이 부족하고 불균형한 클래스 분포가 원인입니다.