**제품리뷰 데이터에 대한 감성분석 보고서**



201820789 정재현

**목차**

1. 서론

A. 보고서 개요

2. 본론

A. 연구 과정

가) 데이터 전처리

나) 모델학습 및 튜닝

다) 모델 평가 및 시각화 보고

3. 결론

A. 요약 및 개선점

**1. 서론**

A. 보고서 개요

본 연구는 아마존 UK 신발에 대한 리뷰에 대해 감성분석을 시행하고 있으며 해당 과정에 대해 기술하고 감성분석 데이터의 전처리와 varder 모델의 구현, 결과 분석, 그리고 모델에 대한 고찰을 통해 연구의 한계점을 회고해보고자 한다.

**2. 본론**

A. 연구 과정

가) 데이터 전처리

우선 리뷰데이터의 경우 모두 텍스트로 이루어져 있다. 하지만 리뷰의 언어가 영어로 모두 통일되어 있지 않고 비영어권 리뷰도 있기 때문에 해당 프로젝트에서 사용하고자 하는 감성분석 모델인 varder의 사용에 문제가 있을 수 있기 때문에 비영어 데이터 사용여부에 대해 고민해본 결과 googletrans모듈을 통해 비영어 데이터를 번역하여 감성분석데이터 표본에 채용하고자 하였다. 또한 리뷰 데이터를 보면 특수문자가 많이 들어있어서 특수문자를 제거하는 과정을 거쳤으며, 텍스트를 소문자로 변환, 불용어 제거 결측치로 된 데이터나 공백 문자열이 존재하는 데이터를 결측치로 바꾸기 위한 작업 또한 실시하였다. 이후 리뷰의 제목과 리뷰의 본문에 해당하는 컬럼을 병합하여 데이터 학습과 처리과정에서의 효율성을 높였다. 이후 리뷰 데이터를 스테밍하고 표제어를 추출하는 작업을 거쳐 데이터 전처리를 마무리하였다.

나) 모델 학습

모델 같은 경우 위에서 기술했듯이 varder를 통해 하였으며 쓰이는 데이터는 리뷰의 제목과 본문을 병합한 processed\_review 데이터 프레임 통해 학습시켰다. 도움이 된 리뷰 데이터나 리뷰 평점 데이터에 가중치를 더하여 계산하는 방식으로 모델을 짜도 되었지만 데이터 편향이나 모델의 정확성에 부정적인 영향을 끼칠 수 있기에 해당 연구에서는 라뷰 데이터만으로 진행하였다.

또한 그리드 서치를 통해 최적의 파라미터를 파악하여 하이퍼 파라미터를 통한 튜닝을 진행하였다.

다) 모델 평가 및 시각화

모델의 경우 먼저 train 데이터와 test 데이터 간의 차이를 확인해보았다. 정확도, 정밀도, 재현율, f1 점수에서 train 데이터와 test 데이터 간의 차이는 보이지 않았으나 재현율이 높다는 점에서 오버피팅의 가능성 또한 보여 이후 ROC와 혼동 행렬 및 모델 예측 분포 검사를 시행하였다.

텍스트, 도표, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

그림 1. 혼동 행렬 및 ROC 커브 시각화

True Positives (TP): 실제 긍정적 샘플 중 모델이 올바르게 예측한 수 (1250)

True Negatives (TN): 실제 부정적 샘플 중 모델이 올바르게 예측한 수 (25)

False Positives (FP): 실제 부정적 샘플 중 모델이 잘못 예측한 수 (88)

False Negatives (FN): 실제 긍정적 샘플 중 모델이 잘못 예측한 수 (2)

현재 위의 수치만 보면 문제점이 당장은 없어 보인다. 하지만 ROC 모양이 기형적이기에어디서 문제인지 확인 중 모델 예측 분포를 검사를 시행하였다.

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

그림 3. 예측 분포 검사 시각화

시행결과 부정적인 리뷰의 표본이 긍정적인 리뷰가 압도적으로 많은 것으로 관측되었다. 즉 데이터가 과도하게 편향되어 이로 인한 오버피팅이 발생하였음을 알 수 있다. 따라서 오버 샘플링을 하여 부정적인 데이터의 비율을 늘려 모델의 오버피팅을 완화하고자 하였다.

텍스트, 도표, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그림 4. 오버샘플링 이후 혼동행렬과 ROC 커브 시각화

오버샘플링 이후 데이터를 다시 확인해본 결과 ROC의 모습이 좌측상단으로 이동하였고 AOC의 면적이 확대되었음을 확인되었기에 오버샘플링이 효과적으로 적용되었음을 확인할 수 있다. 하지만 아직 좌측상단에 밀착되어 있지 않기에 개선의 여지는 남았다고 볼 수 있다.

**3. 결론**

가) 요약

해당 연구는 리뷰데이터를 전처리하여 varder 모델에 다른 가중치를 넣지 않고 리뷰에 대한 긍정과 부정적인 감성을 분석하도록 하였다. 하지만 데이터의 대부분이 긍정적인 데이터에 편향되어 있고 ROC의 모양이 매우 기형적인 모습을 보이는 오버피팅 문제가 발생되었다. 이에 따라 오버샘플링을 진행하여 데이터의 편향을 줄였으며 ROC가 좌측 상단으로 이동시키며 데이터 클래스의 불균형 문제를 완화하였다.

나) 개선점

위에 기술했듯 해당 모델은 순수하게 리뷰의 텍스트만으로만 평가하였으며 다른 감성분석에 대한 척도는 넣지 않았다. 평점이나 도움이 된 리뷰라는 척도를 통하여 가중치를 조절하는 모델에 대한 설계를 통해 학습을 더욱 정교화 시킬 수 있을 것으로 예측된다.

따라서 해당 모델을 개선하기 위해서는 다양한 피쳐가 존재해야 한다고 정리하며 글을 마친다.