

Week1-4 과제

한국 스트리밍 서비스 (왓*, 쿠*플레이, 티*)에서 시청자가 영화를 보고 남긴 리뷰를 긍정과 부정으로 나누어 볼 수 있는 대시보드를 만들려고 한다. 리뷰 긍부정 판별 모델을 만들려고 할 때, NLP 리서처/엔지니어로서 어떤 의사 결정을 할 것인지 각 단계에 맞춰 작성해보자. (단, 수집된 리뷰 데이터의 개수가 1,000개 미만이라고 가정하자.)

대시 보드 예시.

긍정	부정
ID: REVIEW:	ID: REVIEW:
ID: REVIEW:	ID: REVIEW:

1. 문제 정의

풀고자 하는 문제를 정의하세요. 또한 데이터 생성 시 고려해야 할 사항이 있다면 무엇인지 설명하세요. (예, 만약 긍정 리뷰가 부정 리뷰보다 많은 경우 어떻게 해야 할까?, 길이가 정말 긴 리뷰는 어떻게 전처리 해야 할까?)

영화에 대한 만족도 분석을 진행하려고 한다. 어떤 부분에서 만족했는지, 어느 부분이 아쉬운지 키워드를 알아 낼 수 있다.

별점이 있는 리뷰라면 4,5점을 긍정으로 1,2점을 부정리뷰로 분리하고 3점인 리뷰는 테스트데이터로 사용 해 확인을 해본다.

4점과 5점은 확실한 긍정이지만 3점은 사람에 따라 긍정일수도 부정일수도 있다. 따라 테스트를 진행해보고 이후 사람의 판단하에 직접 라벨링을 하는 것이 적절할 것이다.

긍정 / 부정리뷰의 길이의 개수가 다르다면 적은 쪽을 복제하거나 많은 쪽을 삭제하는 방법을 택해 샘플링을 진행하는 것이 좋다.

불균형한 데이터로 학습을 진행하면 편향된 결과를 야기할 가능성이 크기 때문이다. 현재 수집된 리뷰 데이터가 1,000개 미만으로 많지 않은 개수이기에 적은 쪽을 늘리는 오버샘플링을 진행하는 방향이 좋다.

2. 오픈 데이터셋 및 벤치 마크 조사

리뷰 긍부정 판별 모델에 사용할 수 있는 한국어 데이터셋이 무엇이 있는지 찾아보고, 데이터셋에 대한 설명과 링크를 정리하세요. 추가적으로 영어 데이터셋도 있다면 정리하세요.

앞에서 찾아본 영화데이터로 ‘네이버 영화 리뷰’가 있다.
그리고 리뷰분석에서 가장 유명한 영어자료인 IMDb도 있다.

- 네이버 영화리뷰
 - 네이버 영화에서 스크랩된 총 20만개의 리뷰데이터
 - 모든 리뷰의 길이는 140자 미만
 - <https://github.com/e9t/nsmc>
- IMDb
 - 해외영어로 된 영화 리뷰데이터셋
 - 훈련용 데이터, 테스트 데이터가 각 25000개로 총 5만개의 데이터셋
 - <https://www.imdb.com/interfaces/>

3. 모델 조사

Paperswithcode(<https://paperswithcode.com/>)에서 리뷰 긍부정 판별 모델로 사용할 수 있는 SOTA 모델을 찾아보고 SOTA 모델의 구조에 대해 간략하게 설명하세요.
(모델 논문을 자세히 읽지 않아도 괜찮습니다. 키워드 중심으로 설명해 주세요.)

IMDb 를 데이터셋으로 이용한 논문에서 제안한 모델

- [NB-weighted-BON + dv=cosin](#)
- 코사인 유사도를 통해 임베딩 및 학습을 진행함

4. 학습 방식

- 딥러닝 (Transfer Learning)
사전 학습된 모델을 활용하는 (transfer - learning)방식으로 학습하려고 합니다. 이 때 학습 과정을 간략하게 서술해주세요. (예. 데이터 전처리 → 사전 학습된 모델을 00에서 가져옴 → ...)

데이터 전처리 - 사전 학습된 모델을 불러오기 - 모델에 적용하기 위해 데이터를 가공
- 데이터학습 - 테스트 - 평가

- (Optional, 점수에 반영 X) 전통적인 방식
Transfer Learning 이전에 사용했던 방식 중 TF-IDF를 이용한 방법이 있습니다.
TF-IDF를 이용한다고 했을 때, 학습 과정을 간략하게 서술해주세요.

5. 평가 방식

긍부정 예측 task에서 주로 사용하는 평가 지표를 최소 4개 조사하고 설명하세요.

긍부정 예측은 긍정/부정 2가지의 클래스를 가지는 이진분류 모델이기에 아래의 지표에 따라 평가될 수 있다.

1. 오차행렬

		True class		Measures
		Positive	Negative	
Predicted class	Positive	True positive TP	False positive FP	Positive predictive value (PPV) $\frac{TP}{TP+FP}$
	Negative	False negative FN	True negative TN	Negative predictive value (NPV) $\frac{TN}{FN+TN}$
Measures	Sensitivity	$\frac{TP}{TP+FN}$	Specificity	Accuracy $\frac{TP+TN}{TP+FP+FN+TN}$

<https://nicola-ml.tistory.com/41?category=806541>

위 지표에서 파생된 평가지표들을 주로 사용한다. 그 중 긍부정 예측에서는 정확도(**Accuracy**)를 가장 많이 이용한다.

2. F1 score

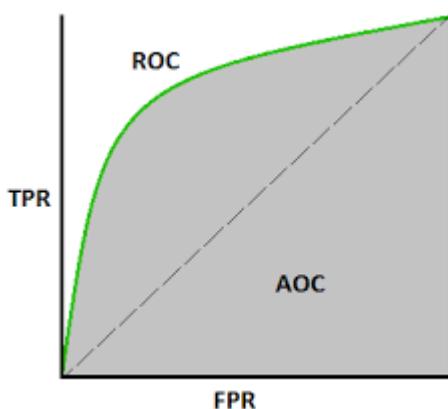
이진분류에서는 Precision과 Recall의 조화평균

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

<https://inside-machinelearning.com/en/recall-precision-f1-score-simple-metric-explanation-machine-learning/>

3. ROC

4. AUC



1-특이도(False Positive Rate)와 민감도(True Positive Rate)를 이용한 ROC곡선으로 위 좌표는 [0,0]에서 [1,1]이므로 총 면적이 1이다.

이 곡선의 아래범위를 AUC(Area Under the Curve)라고 이야기한다.