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.)

테이블이(가) 표시된 사진

자동 생성된 설명

1. 문제 정의

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

—-

**Answer)**

감정 분석은 주어진 텍스트의 극성을 구분하는 작업입니다. 모델의 목표에 따라 긍정, 부정, 중립, 객관 등의 여러가지 분류로 라벨링을 할 수 있습니다. 주어진 텍스트와 라벨들을 이용해, 모델이 올바른 감정을 예측할 수 있도록 하는 것 입니다. 감정 분석 기술은 머신 러닝 접근, 렉시콘 기반 접근 등이 존재합니다. 하위 카테고리로는 multimodal sentiment analysis, aspect-based sentiment analysis, fine-grained opinion analysis, language specific sentiment analysis 가 존재합니다.

데이터를 생성할 시에, 긍정 리뷰가 부정 리뷰보다 많은 경우 혹은 그의 역 관계라면, 오버샘플링이나 언더샘플링을 통해 해결할 수 있습니다. 다수 혹은 소수 범주의 데이터를 그 역 범주의 데이터 수에 맞게 늘리거나 줄이는 방안이 있습니다. 전이학습을 이용하는 것도 방안이 될 수 있습니다.

또한 데이터 중 정말 길이가 긴 리뷰 같은 경우, 결국 Outlier이기 때문에, 가중치 조정으로 해결할 것 같습니다.

2. 오픈 데이터 셋 및 벤치 마크 조사

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

**Answer)**

<https://github.com/e9t/nsmc/>

네이버 영화리뷰 데이터 셋이 있습니다.

2011년에 만들어진 데이터셋이고, 모든 리뷰는 140자보다 짧습니다. 실재 리뷰는 1부터 10까지 점수가 존재했습니다. 데이터를 만든 제작자분께서 1-4까지는 부정, 5-8은 중립, 9-10은 긍정을 의미한다고 판단하셨고 이에 따라, 중립을 제외하고 긍정은 1, 부정은 0으로 라벨링을 하셨습니다.

해당 데이터는 id, document, label 총 3개의 열로 구성되어 있습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

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.)

**Answer)**

2022년 1월 쓰여진 Dual Contrastive Learning: Text Classification via Label-Aware Data Augmentation 논문에 소개된 RoBERTa + DualCL 모델이 Accuracy를 94.39를 기록하였습니다. Contrastive learning은 비지도 환경에서 self-supervision을 통한 representation learning에서 눈에 띄는 성공을 거두었지만, supervised learning 작업에 효과적으로 적용시키는 것이 실무 과제로 남아있습니다. 이 모델은 입력 샘플들의 특징과 분류기의 parameters를 동일한 공간에서 동시에 학습하는 DualCL framework을 이용합니다. DualCL framework은 분류기의 parameter를 다른 라벨에 연관이 된 증강 샘플로 간주한 후, 입력 샘플과 증강된 샘플 간의 대조 학습을 사용하는 방식입니다.

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 → …)

**Answer)**

데이터 불러오기(정규화) → 데이터 전처리 → 모델 정의(및 사전 학습된 모델을 불러옴) → Fine-tuning(손실함수 및 Optimizer 정의) → Train dataset을 통해 학습 후 Test dataset을 통해 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

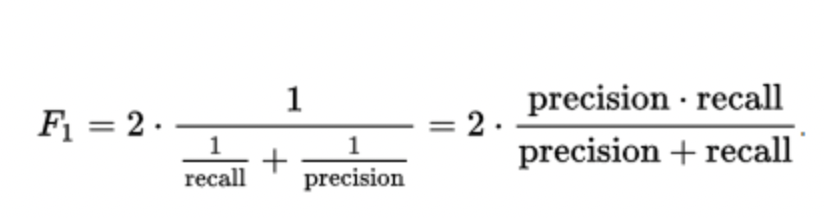
긍부정 예측 task에서 주로 사용하는 평가 지표를 최소 4개 조사하고 설명하세요.

**Answer)**

**F1 score**

Recall(재현율)과 Precision(정밀도)의 조화 평균값이다.

보통 F1 Score 값이 높으면 성능이 높다고 할 수 있다.



**Recall**

재현율은 실제로 True인 데이터를 모델이 True라고 인식한 데이터의 수

Precision 과 서로 상반되는 관계에 놓여있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

**Precision**

정밀도는 모델이 True로 예측한 데이터 중 실제로 Ture인 데이터의 수

Recall 과 서로 상반되는 관계에 놓여있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

**Accuracy**

올바르게 예측된 데이터의 수를 전체 데이터의 수로 나눈 값

텍스트이(가) 표시된 사진

자동 생성된 설명