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 모델**을 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: |

### 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?)

—-

NLU - Sentiment analysis task

일단 리뷰를 긍정과 부정의 의미로 나누는 task. 길이에 대한 threshold를 설정하고 너무 긴 리뷰는 제외하고 빈 공간엔 padding을 주어야 한다. (사실 입력단에서 글자수를 제한하면 이러한 걱정은 필요 없을 것 같다.) 또, 문장부호를 제거하고 불용어(조사),희귀단어(빈도가 극도히 낮은 단어)를 제거하는 전처리를 거쳐야한다. 과도한 혐오표현이 포함된 리뷰는 아예 모델에 넣지 않거나 그 표현을 수정/제외 해야한다. 대시보드에 엄청나게 많은 양의 데이터를 표시하지 않을 예정이면 굳이 긍정과 부정의 비율을 맞추어 보여줄 이유는 없을 것 같다.

### 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.

—-

Korean)

네이버 영화 리뷰 감성 분류 데이터셋(https://github.com/e9t/nsmc/)

네이버의 영화 리뷰에서 샘플링되어서 만들어졌다.

개수 : 20만개(긍정/10만, 부정/10만)

140자 미만의 리뷰들로 이루어져있고 id/리뷰/label 로 구성되어 있다. 영화 리뷰 기준 rating 1-4 는 부정, rating 9-10은 긍정으로 판단하여 label 을 부정은 0 긍정은 1로 분류하였다. 중립적인 의견은 제거되었다.

English)

SST-2 Binary classification 데이터셋(https://huggingface.co/datasets/gpt3mix/sst2)

스탠포드 감성 트리뱅크 SST-2 데이터 세트에는 영화 리뷰의 11,855개 문장의 구문 분석 트리에 미세한 감정 레이블이 있는 215,154개의 구문이 포함되어 있다. 0은 긍정 , 1은 부정으로 표현되었다.

### 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.)

—-

SMART-RoBERTa Large

- 사전 훈련된 언어 모델에 대해 강력하고 효율적인 미세 조정을 위한 새로운 계산 프레임워크 제안

1. 모델의 용량을 효과적으로 관리하는 Smoothness-inducing regularization

2. Bregman proximal point optimization

### 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 로드 → 중복 및 결측치 처리 → 데이터 정제 → 토큰화 및 불용어 제거 → 빈도수가 맞은 단어 제거 → 패딩 → hugging face 에서 불러온 모델 구축 및 학습 → inference 로 성능 확인

### 평가 방식

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

—-

- Binary classification

① Accuracy(%)

모든 테스트 데이터 중 실제로 긍정 또는 부정 레이블을 맞춘 경우를 백분율로 표현

- STS linear regression metrics

② Pearson Correlation Coefficient

모델의 예측 값()과 실제 값(y)의 pearson correlation coefficient (r)는 -1과 1사이의 값으로, 평가 지표는 pearson correlation coefficient에 100을 곱한 백분율로 표현

③

④