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:

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?)

리뷰 긍부정 판별 모델을 만들기 위해서 NLU의 subtask 중 하나인 감성분석(sentiment analysis) 모델을 활용해야 한다. 감성분석은 주어진 텍스트의 polarity(극성)을 분류하는 태스크다. 예를 들어, 텍스트 기반의 트윗은 '긍정', '부정', 또는 '중립'으로 분류될 수 있다. 주어진 텍스트와 매칭하는 라벨과 함께 모델은 알맞은 감정을 예측하도록 학습한다.

<u>리뷰 데이터 특성상 고려할 점</u>

1. 비구조적 특성에 따른 복잡한 전처리 방식
   리뷰 데이터는 인터넷에 사람이 직접 생성하는 데이터이기 때문에 제각각이고, 비구조적인 특성을 갖고 있다. 따라서 다양한 전처리가 필요하다. 특수문자, 줄임말, 초성, 신조어 등을 고려하여 한국어의 인터넷 단어 사전을 제작하고 사전학습 하는 것 또한 고려할만하다.
2. 긍정, 부정 라벨링의 불균형
   학습을 위한 데이터셋 생성시 한 쪽 극단으로만 쏠려있지 않은, 균형 잡힌 데이터를 수집하는 것이 중요하다. 리뷰를 남기는 유저 특성상 부정적인 리뷰가 많을 가능성이 높다. 실제로 제품에 그럭저럭 만족하는 상당수의 유저가 리뷰를 남기는 적극적인 활동까지 이어지지 않는 경우가 많다.
3. 리뷰 데이터 생성 대상의 다양성 확보
   유저 데이터를 확보할 수 있다면, 성별, 나이 등 인구학적인 특성이 균등하게 수집되어 작성자 특성에 따른 편향을 제어한다.
4. 감성 외의 필요없는 정보를 담고 있는 데이터의 길이
   긍정, 부정으로 구분하는 것에 도움이 되는 단어 이외에 영화의 줄거리, 배우 설명 등 복잡한 정보를 길게 갖고 있는 것이 리뷰 데이터의 특성 중 하나다. 감성 점수가 높지 않은 문장의 경우 걸러내는 것 또한 도움이 될 수 있다.
5. 학습 데이터 개수의 적은 양
   리뷰 데이터가 1,000개로 이는 학습 데이터로 사용하기에 굉장히 적은 개수이다. 따라서 transfer learning 등 모델을 미리 대규모 말뭉치로 사전학습 하여 활용하는 것도 좋은 방법이다.

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.

---

1. NSMC – Naver Sentiment Movie corpus (한국어)
   - 출처 : https://github.com/e9t/nsmc
   - 소개 : 네이버 영화에서 크롤링한 영화 리뷰 데이터셋

- ID, 리뷰 텍스트, 라벨의 3개 칼럼으로 구성되어 있으며 라벨은 0(부정) 또는 1(긍정)의 값을 가짐
- 모든 리뷰 데이터는 140 글자 미만
- 학습 데이터 : 150,000
- 테스트 데이터 : 50,000

2. 네이버 쇼핑 리뷰 데이터 (한국어)
    - 출처 : https://github.com/bab2min/corpus/tree/master/sentiment
    - 소개 : 네이버 쇼핑에서 제품별 후기를 별점과 함께 개인이 수집한 데이터셋
    - 별점, 리뷰 데이터 2개의 칼럼으로 구성되어 있으며 별점 1,2점은 부정으로, 별점 4,5점은 긍정으로 판별함
    - 총 리뷰 데이터 갯수는 200,000개
3. 스팀 리뷰 데이터 (한국어)
    - 출처 : https://github.com/bab2min/corpus/tree/master/sentiment
    - 소개 : 게임 유통 서비스 스팀이 판매하는 게임 페이지에 작성된 게임 리뷰 데이터셋으로, 게임 커뮤니티 특성상 비속어 및 은어가 많이 포함되어 있음
    - 라벨링은 0(부정)과 1(긍정) 두 가지 값을 갖고 있으며, 긍정과 부정의 비율이 1:1에 가깝도록 샘플링됨
    - 총 리뷰 데이터 갯수는 100,000개
4. SST (Stanford Sentiment Treebank)(영어)
    - 출처 : https://nlp.stanford.edu/sentiment/
    - 소개 : 완벽히 라벨링 되어 있는 parse trees 말뭉치로 영화 리뷰에서 추출된 11,855개의 문장들로 구성되어 있음. Stanford parser로 파싱되었으며 총 215,154개의 고유한 구문이 포함됨.
    - SST-5, SST fine-grained : 긍정, 어느 정도 긍정, 중립, 어느 정도 부정, 부정의 5단계로 라벨링이 되어 있음
    - SST-2, SST binary : 위의 5단계에서 긍정, 부정만 남은 2가지 값을 라벨로 가짐

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.)

—-

참고 논문 : SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization

- 모델 이름 : SMART-RoBERTa Large
- 모델 등장 배경 : 기존의 많은 최신 모델들은 먼저 대규모 말뭉치(corpus)로 사전학습 한 뒤 다운스트림 태스크를 위한 fine-tuning을 함
    - 하지만 한정된 데이터 리소스와 사전 학습 모델의 매우 큰 용량으로 인해 과도한 fine-tuning을 하게 됨 > 다운스트림 태스크의 데이터에 오버피팅되고 사전 학습을 잊게됨
    - 위에 제시된 한계를 개선하기 위해 사전학습 모델을 위한 효율적인 fine-tuning 프레임워크를 제시함
- 특징
    1. Smoothness-inducing regularization
        - 모델 용량 제어
    2. Bregman proximal point optimization
        - 사전학습 된 모델이 새로운 데이터에 오버피팅 되는 것을 방지

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 → …)

  Transfer Learning이란 특정 태스크를 학습한 모델을 다른 태스크 수행에 재사용하는 방식을 말한다. 트랜스퍼 러닝을 적용하면 모델의 학습 속도가 빨라지고 새로운 태스크를 수행할 때 성능이 향상되는 경향이 있다. 사전학습 할 때의 태스크를 업스트림(upstream) 태스크라 부르고, 그 후 해결해야 하는 태스크를 다운스트림(downstream) 태스크라고 한다.

- 데이터 로드 -> 데이터 전처리 -> 사전학습 모델의 토크나이저 로드 -> 전처리 된 데이터에 토크나이저 적용 -> 사전 학습된 모델 로드 -> 토큰화 된 데이터 입력하여 Transfer Learning -> 성능 compute 및 Inference -> 새롭게 학습된 모델 저장

- (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.

- 데이터 로드 -> 데이터 전처리 -> 데이터 TF-IDF처리 -> 문서 임베딩(벡터화) -> 모델 정의 -> 임베딩 된 데이터 입력하여 모델 학습 -> 성능 compute 및 Inference -> 모델 저장

5. 평가 방식

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

감정 분석 task를 해결하기 위한 딥러닝 모델에는 RoBERTa, T5 등이 있는데 F1, 재현율(recall), 정밀도(precision)과 같은 분류 성능 평가 지표를 기준으로 평가한다.

1. recall : 재현율, 실제 true 인 것중에서 모델이 true라고 예측한 것의 비율.

$$(Recall) = \frac{TP}{TP + FN}$$

2. precision : 정밀도, 모델이 True 라고 분류한 것 중에서 실제 true 인 것의 비율.

$$(Precision) = \frac{TP}{TP + FP}$$

3. f1-score : precision과 recall의 조화 평균. 산술 평균을 이용하는 것보다 major label이 끼치는 bias가 줄어들기 때문에, 데이터 label이 불균형 구조일때 특히 자주 사용한다.

$$(F1\text{-}score) = 2\times\frac{1}{\frac{1}{Precision}+\frac{1}{Recall}} = 2\times\frac{Precision \times Recall}{Precision + Recall}$$

4. ROC curve : Recall-Fallout의 변화를 시각화 한 것으로, Receiver Operating Characteristic 의 줄임말이다. Fallout은 실제 False인 데이터 중에서 모델이 True로 분류한, Recall은 실제 True인 데이터 중에서 모델이 True로 분류한 비율을 나타낸 지표로써, 두 지표를 각각 x, y의 축으로 놓고 그려지는 그래프를 해석하는 것이다. 커브가 왼쪽 위 모서리에 가까울수록 모델의 성능이 좋다고 평가한다. 이 때 그래프 아래의 면적값을 수치로 이용하는데, 이 수치를 AUC (Area Under Curve)로 지칭한다.