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?)

—-

※ 본 문서는 수집 데이터 관련 피드백 전 작성되어 이해의 배경이 다를 수 있습니다.

- **감성 분석** (Sentiment analysis)

  감성 분석은 엔터티가 지닌 주관성을 식별하는 태스크이다.

  현 분석의 목표는 사용자가 생성한 텍스트를 긍정(y=1), 부정(y=0) 의견으로 이진 분류하는 것이다. 따라서 감성 분석 기법을 제시된 과제인 '리뷰 긍부정 판별 모델'의 주요 프로세스로 삼을 것이다.

  예시 :

"스토리가 놀라울 정도로 흥미진진하다"	1
"배우의 연기력이 너무 처참했음…"	0

- **분석 시의 고려사항**
    - **리뷰 텍스트의 길이와 가독성** - 어떤 텍스트를 배울 것인가
        - 학습 데이터의 품질은 감성 분류 성능에 영향을 준다.
        - Li et al.(2018)[1]는 감성 분류의 정확도에 '가독성'과 '길이'가 영향을 미침을 강조, 일반적으로 더 높은 가독성과 짧은 텍스트로 이루어진 데이터셋이 좋은 품질의 분류 결과를 가져오는 것으로 알려져있다.
        - 따라서, 수집 대상으로 삼을 (1) 텍스트의 글자수 제한과 (2) 적절한 불용어 제어가 필요하다.
    - **리뷰 텍스트의 크기** - 데이터가 적을 때 필요한 학습방법
        - 수집 데이터의 개수가 1,000개로 적으므로 전이 학습(TL)을 통한 사전학습을 고려해야한다.
    - **리뷰 텍스트의 불균형** - 긍정 리뷰가 부정 리뷰보다 너무 많다면
        - 한쪽 레이블이 다른 한쪽에 비해 너무 적거나 많다면 데이터 불균형에 의해 감성 분류기의 성능(Accuracy가 편향의 함정에 빠짐)이 떨어진다.
        - 특히 추천 등 해당 과제의 향후 활용 지점을 생각해봤을 때, 사람은 대상의 다양한 측면을 평가할 때 긍정적인 것보다 부정적인 것을 더 중요시하는 만큼 핵심 과제로 데이터 불균형을 다뤄야한다.
        - 충분한 시간이 주어진다면 오버샘플링(랜덤, SMOTE,..), 언더샘플링 등의 기법을 비교 선택해야한다.
    - **리뷰 텍스트의 모호함** - 어느 수준에서 판별할 것인가
        - 대시보드는 긍정/부정만 고려하지만, 실제 텍스트는 본문 안에서 긍부정이 혼재되어있을 수 있다.
        - 따라서, 레이블링 시에 문서 / 문장 / 특징(feature or aspect) 수준 중 적합한 추출 수준이 무엇인지 고려해야한다.

    예시 :

[1] Li, L.; Goh, T.-T.; Jin, D. How textual quality of online reviews affect classification performance: A case of deep learning sentiment analysis. Neural Comput. Appl. 2018, 1–29.

“영화는 전반적으로 아름다운 풍경을 보여준다.”	1
“카메라 워크는 가히 예술의 경지에 있다.”	1
“하지만 그게 이 영화의 전부였다.”	0

- **중립적인 리뷰 데이터** - 긍정, 부정, 중립의 임계치
  - 긍부정에 속하지 않은 중립된 리뷰 텍스트도 존재한다.
  - 사용자들은 리뷰에 콘텐츠에 관련된 평론만 생성하지 않는다. 리뷰는 종종 일기, 영화를 보고 떠오른 저녁 메뉴나 의미없는 단순 문자의 반복 등을 포함한다.
  - 먼저 대시보드에 모든 리뷰 텍스트의 분석 결과를 실을 것인지, 긍/부정이 명확한 텍스트만을 한정할 것인지의 결정이 선정되어야한다.
  - 후자라면 중립 레이블을 생성해 최종 단계에서 긍/부정만을 디스플레이한다.

예시 :

“2007년 어느 겨울, 눈이 내리던 그때의 내가 지금 이렇게 되리라곤 상상도 못했지… 아스라하다.”	1 ?
“카모메 식당 완주. 오늘은 오니기리 먹어야지.”	1..?

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.

—-

- **NAVER Sentiment Movie Corpus (NSMC)**

라이센스	도메인	스타일	윤리적 리스크	크기
CC0 1.0	Review	Colloquial	Medium	Large(~50k)

**위 표는 〈KLUE: Korean Language Understanding Evaluation〉을 참조*

NSMC는 네이버 영화 서비스에서 스크랩한 영화 리뷰 데이터셋. 사용자가 직접 생성한 리뷰로 이루어져있으며, 리뷰 원문과 이진 감성 레이블이 제공된다.

레이블의 각 클래스 데이터는 양적 균형을 이루고 있다.

- **데이터 구성**

  3개의 컬럼 id, document(리뷰 원문), label(P - 1/N - 0)로 구성

- **데이터 크기**

  총 200k의 리뷰 데이터 중 학습용 150k, 테스트용 50k

- **더 알아보기**

  NSMC : https://github.com/e9t/nsmc

  Toxic comment data (NSMC 레이블 상세화) :

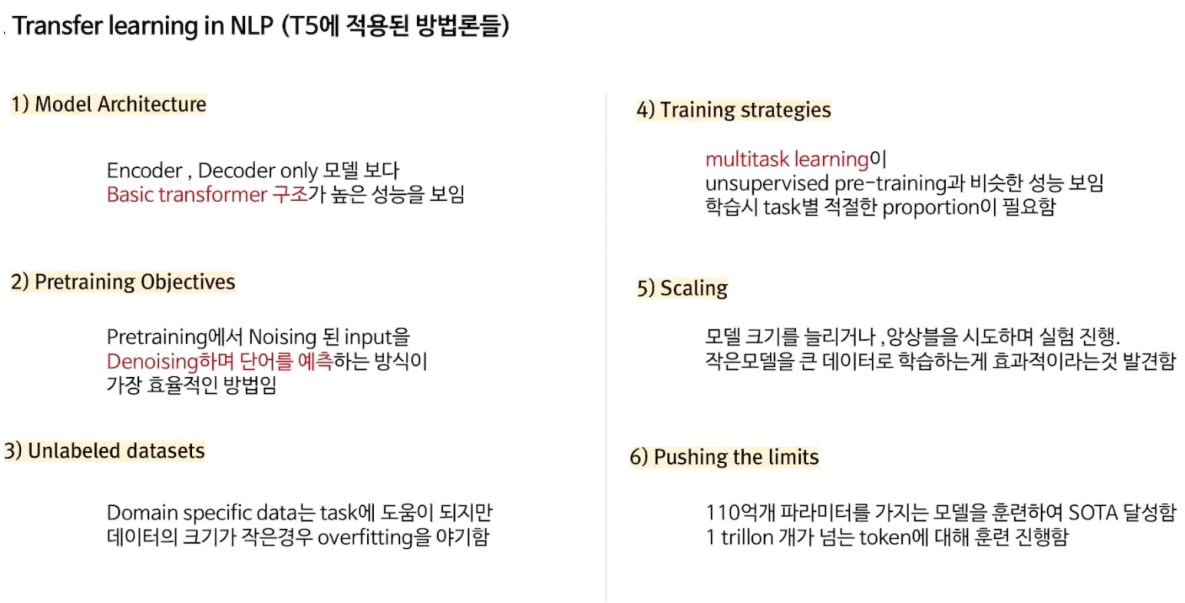
  https://github.com/songys/Toxic_comment_data

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.)

—-

- **T5-3B**



- T5(Text-To-Text Transfer Transformers)는 **인코더-디코더 트랜스포머**

- **모든 NLP 태스크에 통합된 프레임워크**이며 입력-출력 포맷이 자유롭다.
- **3B 변형** : 32-headed attention, dff = 16,384로 약 28억 개의 파라미터를 가짐
- **pre-training dataset** : Colossal Clean Crawled Corpus (C4)
- **vocab** : SentencePiece (Kudo and Richardson, 2018)

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 → …)

  - 전이학습(TL)은 레이블되어 있지 않은 풍부한 텍스트 데이터에서 self-supervised 태스크(언어 모델링이나 누락된 단어 채우기)를 미리 학습하여(Pre-training) 모델을 만듬. 그리고 적은 레이블된 텍스트 데이터를 이용하여 모델을 튜닝(fine-tuning)함.

  —-

5. 평가 방식

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

—-

- **Accuracy** (정확도)

  $$\frac{TP+TN}{P+N} \times 100$$

  - Accuracy는 테스트의 정확도로 전체 예측한 것 중에 올바른 예측을 얼마나 했는지를 나타낸다.
  - 양성으로 예측하거나 음성으로 예측한 것 중 실제로 양성과 음성을 모두 얼만큼 맞췄는지를 알 수 있다.
  - 이에 따라 $Error\ Rate\ =\ 1\ -\ Accuracy$를 구할 수 있다.

- **Precision (정밀도)**

  $$\frac{TP}{TP+FP}$$

- 분류기의 정확성, 즉 양성을 예측하거나 음성을 예측했을 때 얼마만큼 잘 맞췄는지를 측정한다.
- 정밀도가 높을수록 위양성이 적음을 의미하고 정밀도가 낮을수록 위양성이 많음을 의미한다.

- **Recall (재현률)**

$$\frac{TP}{TP+FN}$$

  - Recall은 분류기의 완전성 또는 민감도를 측정한다.
  - 재현율이 높을수록 위음성이 적다는 것을 의미하고, 재현율이 낮을수록 위음성이 더 많다는 것을 의미한다.

- **F1 score**

$$2 * \frac{Recall * Precision}{Recall + Precision}$$

  - F-점수는 정밀도와 재현율의 조화 평균으로 분류기의 성능을 측정한다.
  - 일반적으로 불균형 데이터의 평가 지표로 사용된다.