과제 1-4

> **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.)**

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 데이터가 있을 때, 리뷰를 긍정 또는 부정으로 분류할 수 있는 모델을 만들어야 한다. 즉, **Input 데이터로 리뷰 데이터(N자 이하의 문자열)가 주어지면 이를 0 또는 1(0은 부정, 1은 긍정)으로 분류**하는 문제로 정의할 수 있다.
- 문제에서 수집된 리뷰 데이터 개수가 1000개 미만이라고 하였다. Text generation등의 방법으로 데이터 수를 늘리는 것도 좋지만, 그러기엔 처음 데이터 수가 너무 적기 때문에 생성한 데이터의 비율이 원래 데이터보다 많이 크게 되어야 할 것이기에 생성된 데이터에 너무 의존하게 될 것이다. 따라서 Pretrained model에 Transfer learning을 통해 학습을 진행하는 것이 바람직할 것이다.
- 데이터 전처리의 경우, 리뷰의 경우 보통 최대 길이가 정해져 있기 때문에, 모든 입력 데이터를 최대 길이가 되도록 패딩 토큰 <pad>을 넣어준다. 길이 제한이 없다면, 전체 길이 제한을 임의로 설정하고, 그 길이가 넘는 데이터는 제외하는 방법도 있을 것이다. 하지만 가능하다면 좋은 Text summarization 모델이 있다면, 적절하게 리뷰를 축약해서 데이터로 활용해도 좋을 것 같다.
- 긍정 리뷰와 부정 리뷰의 비율에 차이가 있을 경우, 적은 비율의 데이터를 생성하거나 (Oversampling) 많은 비율의 데이터를 줄여(Undersampling) 둘의 비율을 맞출 수 있다. 오버샘플링은 데이터 손실이 없고 데이터를 증가시킬 수 있지만 생성된 데이터가 아무래도 Noise있는 데이터가 많을 것이라 문제가 생길 수 있고, 언더샘플링은 데이터 품

질의 문제없이 label 비율을 맞출 수 있지만 학습 데이터 수가 줄어드는 문제가 있다. 일단 적절한 방법으로 Text generation을 통해 오버샘플링을 시도해볼 것이다.

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.

- 네이버 영화 리뷰 긍정 여부 데이터셋이 있다.
- 링크 : https://github.com/e9t/nsmc
- 약 20만 개의 영화 리뷰와 해당 리뷰의 negative, positive 여부를 0과 1로 나타낸 label 목록이 있다.
- positive와 negative label data는 거의 1:1 비율이다. 중립 리뷰는 없다.
- 15만 개의 train data, 5만 개의 test data로 나뉜 파일도 존재한다.

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.)

- SOTA 모델 중 ALBERT 모델을 골랐다. 이는 논문에서부터 모델 경량화에 초점을 둬서 기존 BERT 모델에 비해 속도가 빨라 서비스에 유리하고, 성능도 SOTA급으롭 잘 나오기 때문이다.
- ALBERT는 BERT에서 단어 임베딩 레이어 수를 Hidden 레이어 수보다 적게 하였다. (기존엔 Hidden 레이어와 Embedding 레이어 수가 동일하였다.) 이는 Embedding layer보다 Hidden layer에서 문맥이 반영이 되기에 더 많은 정보를 포함하여야 하기 때문이다.

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
	xlarge	1270M	24	2048	2048	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

- Parameter-sharing을 통해 파라미터 수를 획기적으로 줄였다.
- Sentence ordering prediction을 통해 두 문장의 순서를 맞히도록 학습하여, 기존의 다음 문장을 예측하는 형태보다 모델 성능 향상에 기여하게 하였다.

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 → …)

- 우리가 수집한 리뷰 데이터셋과 pretrain을 위한 데이터셋, 여기서는 위에서 언급한 네이버 영화 리뷰 긍정 여부 데이터셋을 준비한다.
- 두 데이터를 모두 전처리한다.(data imbalance 처리, padding, tokenize)
- 모델을 불러온다. ALBERT 모델을 불러온다 가정하고, 마지막 layer에 out dimension이 2인 linear layer를 추가한다.
- 네이버 영화 리뷰 데이터로 모델을 먼저 학습시킨다.
- 우리 학습 데이터로 모델을 학습시킨다.
- validation 및 test score에 따라, 위 과정에서 모델 구조나 데이터 형태, 하이퍼파라미터를 바꿔가며 재학습한다.

5. 평가 방식

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

- 긍부정 예측 task는 결국 classification task라고 할 수 있으므로, 해당 task에서 사용하던 metric인 Accuracy, Precision, Recall을 사용할 수 있다.

- Accuracy는 **예측에 성공한 데이터 수 / 전체 데이터 수** 를 의미한다.
- Precision은 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율이다.
- Recall은 실제로 True인 데이터 중에서 모델이 True라고 예측한 비율이다.
- 예측 label이 2개인 경우 위의 경우밖에 찾지 못했으나, multi-label인 경우 cross entropy loss를 활용한 F1-score를 metric으로 쓸 수 있다. 이는 전체 label 중 가장 정답일 확률이 높은 label을 맞춘 경우의 비율을 의미한다. 즉 **가장 확률이 높다고 예측한 label이 실제 정답인 데이터 수 / 전체 데이터 수**이다.