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?)

—-

OTT 서비스의 사용자 리뷰를 크롤링하여 긍부정 데이터셋을 생성해 리뷰 감정 판별 모델을 생성합니다. 데이터를 생성할 때 편향이 일어나지 않도록 넷플릭스, 왓챠, 티빙과 같은 스트리밍 서비스에서 특정 플랫폼, 장르, 국가에 치우치지 않게 크롤링 대상을 고르게 설정합니다. EDA(Easy Data Augumentation)을 사용하여 데이터를 증강해 과소 적합을 방지합니다. Synonym Replacement, Random Insertion을 사용할 때 문장의 긍부정 뉘앙스가 달라지지 않도록 데이터를 살펴보아야 합니다. Random Deletion을 사용할 때 감정을 나타내는 핵심 단어가 빠지지 않도록 이 또한 살펴보아야 하겠습니다.

긍정 리뷰가 부정 리뷰보다 많은 경우, 데이터 불균형을 고려해야 합니다. 데이터 불균형을 해소하는 방법은 크게 Weight Balancing과 Over& Under Sampling

으로 두 가지가 있습니다. Weight Balancing을 사용하여 학습할 때, 모델이 더 데이터가 적은 부정 리뷰에 민감하게 반응하도록 가중치를 줍니다. Over & Under sampling을 할 때, 상대적으로 수가 적은 부정 리뷰를 오버 샘플링한다면 부정에 과적합이 일어날 우려가 있고 상대적으로 수가 많은 긍정 리뷰를 언더 샘플링한다면 유의미한 데이터의 손실이 일어난다는 게 고려할 점입니다.

길이가 긴 리뷰의 전처리는 여러 방식으로 할 수 있습니다. 모델이 수용가능한 글자수에 맞춰서 텍스트를 자르는 방법, 형태소 분석기를 사용하여 부사와 형용사를 삭제하는 방법, Text Summarization 학습 모델을 사용하여 핵심 문장을 골라내어 사용하는 방법이 있습니다. 모델링을 수행할 시간에 따라 어떤 방식을 선택할지 달라집니다. 수행할 시간이 적다면 글자 수에 맞춰서 Trim 하는 방법, 시간이 많다면 Text Summarization을 사용하겠습니다. 부사와 형용사를 삭제하는 방법은 긍부정을 판단하기 위해 이러한 수식어가 크게 작용하기 때문에 위 모델에서는 사용하지 않을 방식이라고 생각합니다.

리뷰라는 자유로운 매체 특성상 줄임말이나 가벼운 욕설이 등장할 수 있습니다. 이러한 경우도 학습하기 위해 사전에 추가시키거나 자모음, 음절, 단어를 하나로 임베딩하는 방법을 고려해볼 것 같습니다. 또, 작품의 줄거리를 한 문단으로 작성하고 그 뒤에 자신의 의견을 작성하는 형태의 리뷰도 자주 등장합니다. 이러한 아웃라이어를 탐지하기 위해 데이터를 점검하는 일이 필수적이겠습니다.

### 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.

—-

대표적인 한국어 영상매체 리뷰 데이터셋으로는 NSMC(Naver Sentiment Movie Corpus)가 있습니다. 네이버가 제공하는 영화 리뷰 감정 데이터셋입니다. 모두 140자 미만의 길이로 된 한글 텍스트 데이터이며 0과 1로 긍부정 라벨링이 되어 있습니다. 0은 부정이고 1은 긍정을 의미합니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

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

가장 대표적인 영어 영화 데이터셋으로 IMDb가 있습니다. 50,000개 영화 후기를 포함한 데이터로 감상평과 제목, 언어, 장르 정보를 제공합니다. 0과 1로 긍부정이 태깅되어 있습니다. 한 영화당 30개 이하의 감상평만 포함되도록 제한했습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

추가로 감정이 태깅되어 있지 않은 데이터도 제공합니다. 직접 라벨링이 가능하다면 활용해서 사용할 수 있겠습니다. 라벨 미부착 데이터는 라벨에 -1로 표기되어 있습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

* https://www.imdb.com/interfaces/

### 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.)

—-

긍부정 감정분류라는 Sentiment Classification Task에서는 ‘NB-weighted-BON + dv-cosine’ 모델을 사용할 수 있습니다. Naïve Bayes weighted Bag of N-gram과 cosine similarity를 적용한 감정분류 모델입니다. 내적(Dot product) 대신 코사인 유사도를 사용한 문서 임베딩을 사용합니다. IMDb 데이터셋에 Naïve Bayes weighted Bag of N-gram과 cosine similarity방식을 적용하였을 때 위 자료를 보면 내적을 사용하거나 L2 Regularization을 사용하는 방식보다 더 정확도가 향상되었음을 알 수 있습니다. 하단 세 개의 자료를 보면 특히 Naïve Bayes weighted Bag of N-gram을 사용할 때에 일반적으로 더 높은 정확도를 보여줍니다.

테이블이(가) 표시된 사진

자동 생성된 설명

* https://aclanthology.org/P19-2057.pdf

### 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 -> 데이터 전처리 -> 토크나이저에 적용 -> 모델 정의 -> 사전 학습된 모델 로드 -> (제대로 로드 됐는지 infer를 통한 확인) -> 전처리한 데이터를 넣어 Transfer Learning -> 성능 평가 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 TF-IDF처리 -> 문서 임베딩-> 모델 정의 -> 임베딩 데이터를 넣어 Train -> 성능 평가 및 inference -> 모델 저장

### 평가 방식

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

—-

Accuracy, Recall, Precision, F1이 있습니다.

Accuracy는 전체 예측한 data의 개수에 대해 옳게 예측한 data 개수의 비율을 나타내는 평가 지표입니다.

Precision(정밀도)이란 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율입니다. Recall(재현율)은 모델이 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율입니다.

Recall과 Precision은 흔히 Trade-off 관계에 있습니다. 모델을 설계하는 엔지니어는 만들고자 하는 모델 목적에 맞게 정밀도에 더 중점을 두어 모델을 설계하거나 재현율에 더 중점을 두어 설계하는 방식을 고려해야 합니다.

F1은 Precision과 Recall의 조화평균입니다. F1-score는 조화평균을 이용하면 산술평균을 이용하는 것보다, 큰 비중이 끼치는 bias가 줄어들기 때문에 정밀도와 재현율 두 개의 성능이 조화롭게 성능을 내는지 판단하기 위해 위 연산하여 평가합니다.