

# 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 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.)

대시 보드 예시.

긍정	부정
ID: REVIEW:	ID: REVIEW:
ID: REVIEW:	ID: REVIEW:

## 1. 문제 정의

풀고자 하는 문제를 정의하세요. 또한 데이터 생성 시 고려해야 할 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 통해 시청자들의 여론을 파악하고자 한다.

### [문제 사항]

- 모델링을 위해 ‘데이터부족’ 문제가 존재
  - Transfer Learning으로 사전 학습된 모델을 활용할 수 있다  
TL은 학습데이터가 많이 부족한 경우, 다른 분야 또는 비슷한 분야의 데이터로 학습된 모델을 사용하여 학습하는 것을 가능하게 한다.
- 오입력 또는 일상어 문제 (\*개인적 추측입니다)  
인터넷 리뷰의 특성 상 맞춤법이 맞지 않는 문장이 많다  
맞춤법이 틀린 단어나 일상어를 형태소분석기가 분류하지 못해 잘못 토큰화 되는 경우 해당 리뷰는 정확한 판별을 하기 어렵다
  - 일상어 소스를 학습, 평가 데이터에 추가할 수 있다
  - ‘재밌다’처럼 문법적으로 틀린 예외를 수정해준다

ex) train[""].apply(lambda x : x.str.replace('재밌', '재밌').replace('체고','최고'))

### [기대 사항]

리뷰 공부정 판별 모델을 통해, 고객관리, 컨텐츠 수요/선호도 파악 등을 통해 플랫폼의 서비스 만족도를 높일 수 있습니다.

더 나아가 리뷰 공부정을 점수화 해 개인화 추천시스템에 활용할 수 있는 데이터로 만들 수 있습니다.

---

## 2. 오픈 데이터셋 및 벤치 마크 조사

리뷰 공부정 판별 모델에 사용할 수 있는 한국어 데이터셋이 무엇이 있는지 찾아보고, 데이터셋에 대한 설명과 링크를 정리하세요. 추가적으로 영어 데이터셋도 있다면 정리하세요.

---

- 감성분석 한국어 데이터셋

데이터명	출처	세부사항	총 건수	링크
NSMC	개인 수집 데이터	긍정(100k), 부정(100k)	200k	<a href="#">링크</a>
감성 분석 말뭉치 2020	모두의 말뭉치	-	2,081 (문서 기준)	<a href="#">링크</a>
소상공인 고객 주문 질의-응답	AI Hub	발화자, 상점카테고리, Q/A구분, 감성, 인텐트, 개체명 등의 태깅 존재	5,000k	<a href="#">링크</a>
네이버 쇼핑 리뷰	개인 수집 데이터	긍정(99,963), 부정(100,037)	200k	<a href="#">링크</a>
steam 게임 리뷰	개인 수집 데이터	긍정(29,996), 부정(50,004)	100k	<a href="#">링크</a>

### 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.)

---

#### KorBERT

- 한국어는 교착어로 명사/동사와 같은 내용어와 조사/어미와 같은 기능어가 결합하여 하나의 어절을 구성하는 언어이기 때문에, 올바른 문맥표현을 위해서는 내용어와 기능어를 구분하는 단계가 필요
- 따라서 KorBERT는 입력 문장에 대해 형태소 분석을 수행하고, 형태소 분석된 결과에 기반하여 각 토큰 간의 문맥표현을 학습한다

구조	KorBERT-Morphology 모델과 형태소 분석 수행
모델 사이즈	[Morphology] vocab = 30,349  [WordPiece] vocab = 30,797  - 12-layers
학습 corpus	기사, 백과사전 - 23GB - 4.7B 형태소
벤치마크 성능	[Morphology] - 기계독해: KorQuAD 1.0 EM 86.40%, F1 94.18% - 의미역결정: Korean Propbank F1 85.77%  [WordPiece] - 기계독해: KorQuAD 1.0 EM 80.70%, F1 91.94% - 의미역결정: Korean Propbank F1 85.10%

#### 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 → ...)

1. 데이터셋 불러오기
2. transformers 모듈로 huggingface에서 pretrained model & tokenizer 불러오기  
\*호출 메소드
  - AutoTokenizer.from\_pretrained
  - AutoModel.from\_pretrained
3. 데이터 전처리
  - NaN 및 중복제거 처리
  - 입력데이터셋 토큰화(tokenization, padding, int encoding)
4. Fine-tuning
  - 모델에 학습 대상 train data의 input data로 label 예측
5. 정확도 평가
  - test data의 input data로 predict
  - predict 값과 test data label 간의 Accuracy 측정
6. 정확도 개선
7. best model 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.

—

1. 데이터셋 불러오기
2. 데이터 전처리
3. 데이터를 TF-IDF Vectorizing
4. 학습, 검증 데이터셋 분리
5. 모델 선정 및 학습
6. 평가 후 모델 저장

## 5. 평가 방식

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

—

아래 평가 지표들은 confusion matrix의 2진 분류 결과에 기반합니다

- True Positive (TP) - 모델이 정답(Positive)을 맞추었을 때
- True Negative (TN) - 모델이 오답(Negative)을 맞추었을 때
- False Positive (FP) - 모델이 오답(Negative)을 정답(Positive)으로 잘못 예측했을 때
- False Negative (FN) - 모델이 정답(Positive)을 오답(Negative)으로 잘못 예측했을 때

### 1. Accuracy (정확도)

전체 중 모델이 바르게 분류한 비율 (전체 표본 중 정확히 분류된 표본의 수)  
데이터에 불균형이 있는 경우 정확도가 높더라도 신뢰할 수 있는 모델인지  
확인할 수 없는 문제가 있다

#### \* Accuracy (정확도)

$$\text{정확도} = \frac{TP + TN}{TP + FN + FP + TN}$$

### 2. Precision(정밀도)

모델이 Positive라 분류한 것 중 실제값이 Positive인 비율  
precision이 높다는 것은 관련있는 결과를 그렇지 않은 것 대비 잘 맞추는  
모델임을 알 수 있다

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

### 3. Recall(재현도)

실제값이 Positive인 것 중 모델이 Positive로 분류한 비율  
Recall이 높은 경우 모델이 관련있는 결과를 최대한 많이 가져옴을 알 수 있다

\* Recall (재현도)

$$\text{재현도} = \frac{TP}{TP + FN}$$

### 4. F1 Score

Precision과 Recall의 조화 평균  
Precision과 Recall은 trade-off 관계에 있어, 이 2가지가 모두 중요한 경우 F1 Score를 지표로 활용해, 최적의 Precision과 Recall을 기준으로 평가한다

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * (precision * recall)}{precision + recall}$$

- [평가지표 이미지 출처](#)
- [평가지표 이미지 출처2](#)