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?)

—-

자연어 처리에서 감정분석은 어려운 부분이다. 슬픔, 두려움, 즐거움, 기쁨, 행복, 우울, 무서움 같은 감정은 문장의 순서에따라 자주 바뀌는 주관적 요소이기 때문에, 긍정, 부정 2가지로 나누는것을 목표로 한다.

주어진 비정형 텍스트에서 긍정이 70%, 부정이 30% 나왔다고 했을때, 왜그런가에 대해선 뚜렷하게 알수가 없다. 다만 **의사결정나무**를 통해 그의미를 유추해 볼수있다.

리뷰 댓글이 정말 긴경우 가령 **“올 봄에 벚꽃기간에 방문, 협재를 바라보는 바다뷰가 좋고 대로변이라 렌트해서 가기도 좋음. 조식은 이용안했는데 근처 옹포밥집까지 아침 산책겸 걸어가서 하고옴. 루프탑 수영장과 바가 있었는데 내가 갔을때는 밤에 비바람이 너무 불어서 이용못하고옴 ㅠㅠ 단점으로는 모 유명 여행블로거 리뷰처럼 화장실 물떄가... 그거빼곤 다 만족”** 과 같은 댓글이 있으면 소량의 "특수 문자"와 "모음"이 존재하기때문에 정규 표현식을 이용하여 제거 할수있다.

### 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.

—-

| 한국어 데이터셋 | 영어 데이터셋 |
| --- | --- |
| <https://github.com/e9t/nsmc>  네이버 영화리뷰  설명: 긍/부정 라벨 데이터  - 학습 15만건 / 테스트 5만건 | <https://www.kaggle.com/datasets>  캐글 영어 데이터셋  굉장히 많은 데이터셋이 존재하며  주로 긍정/부정 형식의 2진 답변을 반환하여 1/0 또는 1/-1의 값으로 긍정 부정을 판단한다. |
|

### 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.)

—-

“The appetizers are ok, but the service is slow.”

와 같은경우 taste에선 긍정, service와 같은 태도는 부정이다. 문장의 내용과 aspect에서 관계성을 탐구하는것이 가치가 있어 보인다.

Aspect-level sentiment classification을 위한an Attention-based Long Short-Term Memory Networkd에 대해서 설명을 하자면.

Aspect-level sentiment analysis은 세밀하고 깊이있는 분석을 제공하는데, LSTM에서

중간층을 softmax layer에 넣으면 결과가 긍정,부정, 중간 이렇게 나오게 된다.

### 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 → …)

**긍정부정 단어모음집 업로드 - > csv파일로 저장 ->각각 문장을 토큰화 ->토큰화된 단어를 컴퓨터가인식할수 있도록 정수인코딩-> y값을로 들어갈 label -1,0,1로 원핫인코딩 진행-> from keras.layers import Embedding, Dense, LSTM , from keras.models import Sequential, from keras.preprocessing.sequence import pad\_sequences 와 같은 필요한 것들 import -> LSTM, softmax,categorical\_crossentropy사용 -> training과 validation\_split진행 -> rmsprop를 모델로 지정및 학습진행 -> 데이터셋 평가 ->교차검증을 위해 optimizer를 adam으로 진행 및 학습-> 데이터셋 평가**

* (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.

—-

### 평가 방식

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

—-

성능 평가 지표(Evaluation Metrics)는 일반적으로 모델이 분류냐 회귀냐에 따라 여러 종류로 나뉜다.

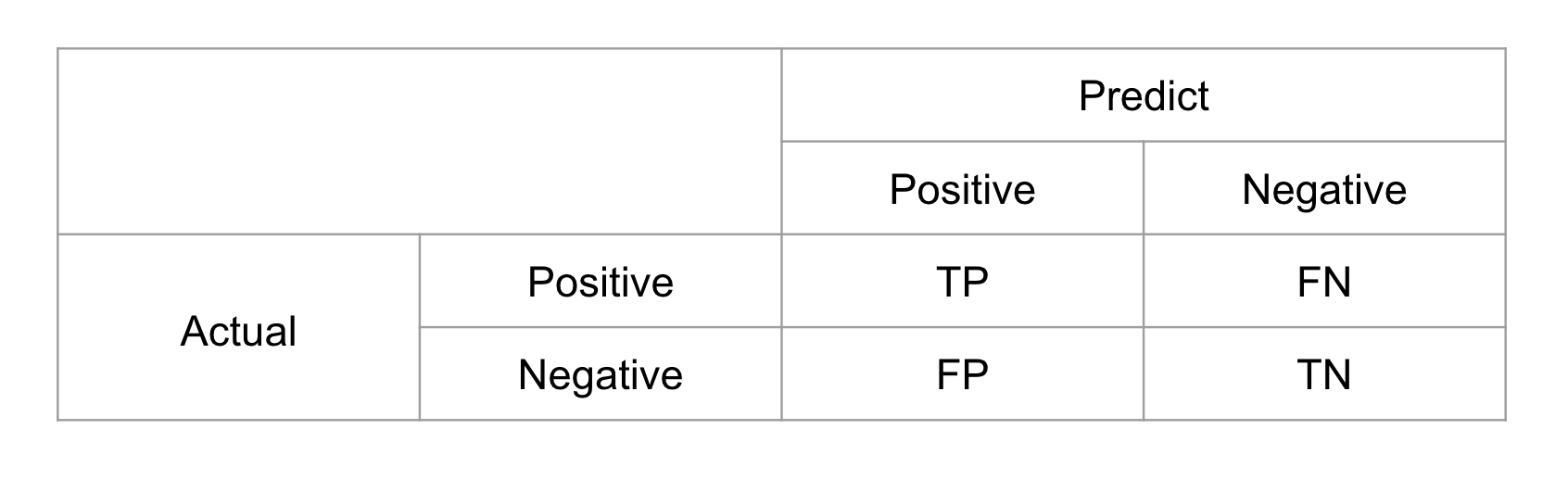
**분류의 성능 평가 지표로는**

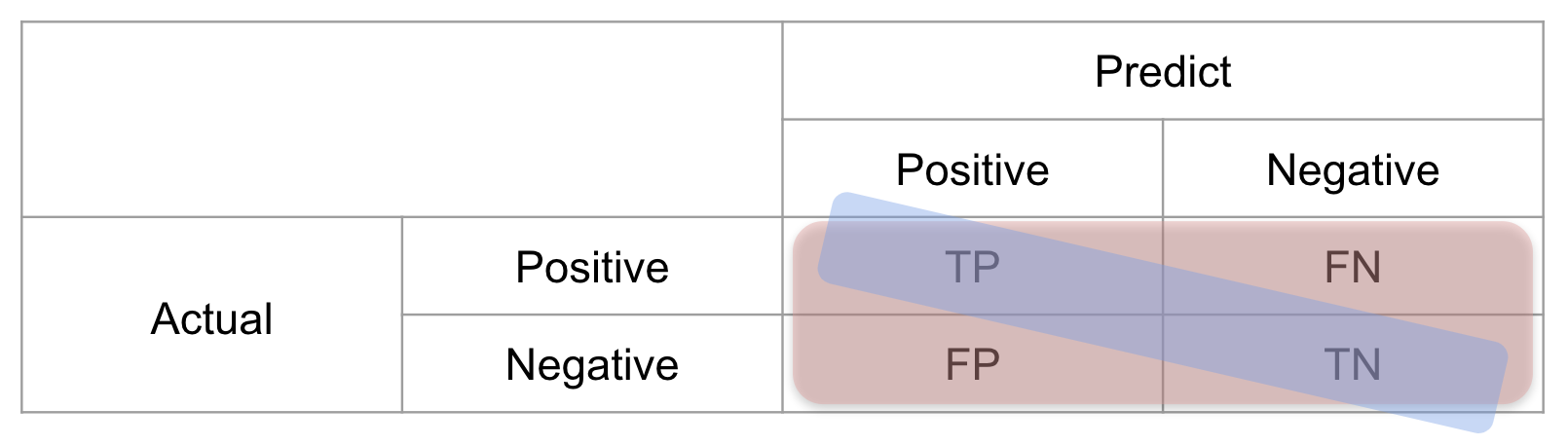
* **정확도(Accuracy)**
* **오차행렬(ConfusionMatrix)**
* **정밀도(Precision)**
* **재현율(Recall)**

정확도(Accuracy) = 예측 결과가 동일한 데이터 건수 / 전체 예측 데이터 건수

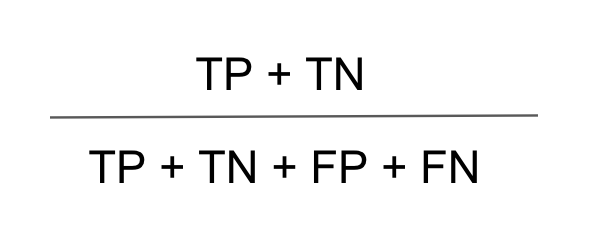
이진분류는 데이터의 구성에 따라 머신러닝모델 성능을 왜곡할 수 있어서 정확도 하나로 성능을 평가하지 않는다. 예를 들면 100대의 데이터에서 90개의 레이블이 0, 10개의 레이블이 1이면 0의 예측결과를 반환하여 정확도가 90%가 되기 때문이다. 이 한계점을 극복하기 위해 여러 분류지표를 같이 사용해야 한다.

Confusion Matrix





**Accuracy**는 정확도로 전체 예측한 것 중에 올바른 예측을 얼마나 했는지를 지표로 구한다.

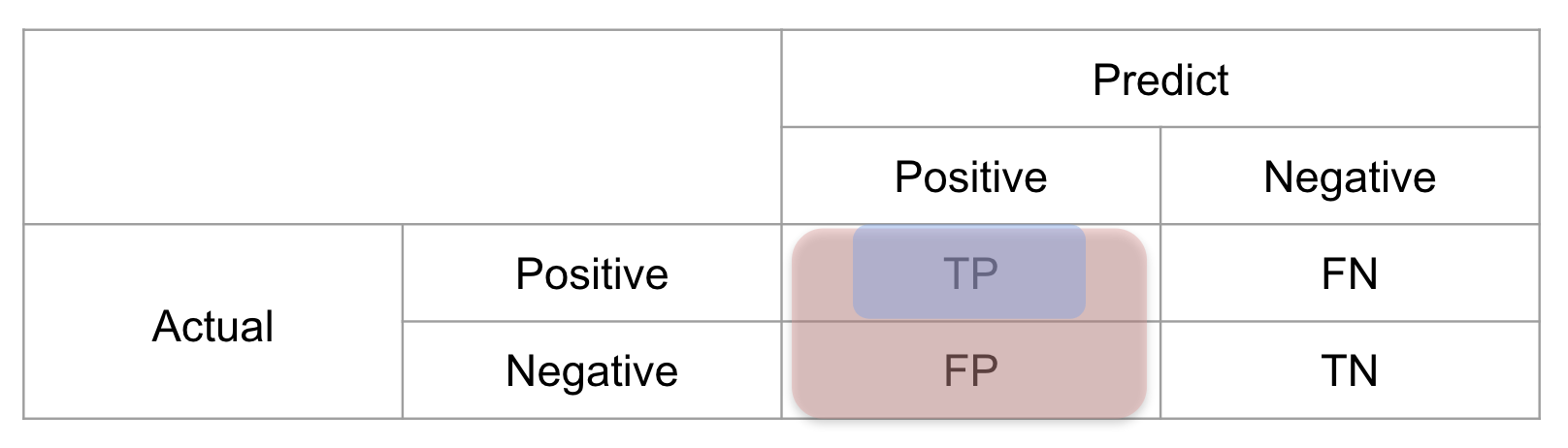


Accuracy값이 높을수록 예측 정확도가 높다고 할수있다.

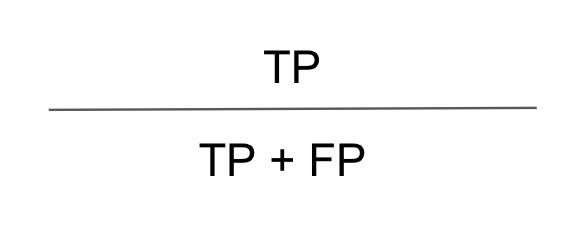
에러 도출은 Error Rate = 1 - Accuracy 로 나타 낼수있는데, 데이터가 균형잡혔을때는 성능이 좋으니 불균형 할때는 부정한 예측을 잘 하지 못한다.

그렇기 때문에 Accuracy는 데이터가 균형 잡혔을때 써야 한다.

**Precision**

****

***Precision****은 정밀도로 긍정으로 예측한 것 중 실제로 맞춘 비율*이다.



정밀도는 Positive라고 예측한 것중 실제로 positive인지를 보여주는 지표로, 얼마나 정밀한지 나타낼수있다.

Recall

**재현율 = TP / (FN+TP)**

재현율은는 실제 Positive중 예측과 실제 값이 Positive로 일치한 데이터의 비율을 뜻한다.

공식의 분모인 FN+TP는 실제 값이 Positive인 모든 데이터 건수이며 분자인 TP는 예측과 실제 값이 Positive로 일치한 데이터 건수이며, 이를 민감도(Sensitivity) 또는 TPR(Tru Positive Rate)라고도 부른다.