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 1,000 개 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?)

   —-

   문제정의: 각각의 리뷰에서 긍정과 부정의 문맥을 이해하고 분류하여야 한다.

   고려사항:

   1. 한국어의 특성상 단어하나하나보다 문맥의 흐름을 이해해야 하는데 어떤 식으로 분류를 해야 할까?
   2. 명사를 추출하여 제거하고 사용하면 어떨까?
   3. 각 리뷰에서 토픽 별로 분류하고 감성단어를 분류하면 어떨까?

      토픽 예시:

      배우(***배우님, ***역),

      패션(셔츠, 드레스, 반지),

      스토리(캐릭터, 영화, 스토리)

4. 의미적 분석으로 각각의 리뷰에서 긍정과 부정의 단어마다 점수를 부여하여 전체 점수로 분류하는 것을 어떨까?
   긍정 : +1 점 / 부정 : -1 점

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.

—-한국어

네이버 영화 리뷰 https://github.com/e9t/nsmc

네이버 영화 리뷰에 대한 자료

korean-hate-speech https://github.com/kocohub/korean-hate-speech

한국 연예뉴스의 비방, 등의 hate-speech 를 담은 자료

AI HUB 감성 대화 말뭉치 https://aihub.or.kr/aidata/7978

일반인 1,500 명을 대상으로 하여 음성 10,000 문장 및 코퍼스 27 만 문장 구축 및 세대별 감성 대화 텍스트 구축

AI HUB 한국어 감정 정보가 포함된 연속적 대화 데이터 셋

https://aihub.or.kr/keti_data_board/language_intelligence

- A 열 : 발화 시작 구분자
- B 열 : 발화 본문
- C 열 : 해당 문장의 감정 정보 (행복/중립/슬픔/공포/혐오/분노/놀람)

각 대화문의 시작은 파란색 음영 및 A 열의 S 표시로 구분되어 있음

모든 대화문은 두 사람의 대화 내용이며 행이 바뀌면 발화자가 바뀜

AI HUB 한국어 감정 정보가 포함된 단발성 대화 데이터 셋

https://aihub.or.kr/keti_data_board/language_intelligence

SNS 글 및 온라인 댓글에 대한 웹 크롤링을 실시하여 문장을 선정함

7 개 감정(기쁨, 슬픔, 놀람, 분노, 공포, 혐오, 중립) 레이블링 수행

AI HUB 감정 분류용 데이터 셋

https://aihub.or.kr/opendata/keti-data/recognition-visual/KETI-01-001

감정 유추가 가능한 대화 데이터를 사람이 연기하여 결과를 저장하고, 동시에 해당 데이터의 감정 상태와 감정 주체 부여

영어

Twitter Sentiment Analysis https://www.kaggle.com/jp797498e/twitter-entity-sentiment-analysis

트위터의 엔터티 레벨 감성 분석 데이터셋

긍정, 부정 및 중립의 세 가지 클래스

AI HUB 트위터에서 수집 및 정제한 대화 시나리오 데이터 셋

https://aihub.or.kr/keti_data_board/language_intelligence

다수의 turn(질문, 답변 1 회)으로 구성된 연속 대화 시나리오

IMDb Movie Reviews https://paperswithcode.com/dataset/imdb-movie-reviews

긍정적 또는 부정적으로 분류된 IMDb(Internet Movie Database)의 50,000 개 리뷰

부정적인 리뷰는 10 점 만점에 4 점 이하, 긍정적인 리뷰는 10 점 만점에 7 점 이상

영화당 30 개 이하의 리뷰가 포함

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.)

—-

Sentiment Classification Using Document Embeddings Trained with Cosine Similarity

Task	Dataset	Model	Metric Name	Metric Value	Global Rank	Uses Extra Training Data	Benchmark
Sentiment Analysis	IMDb	NB-weighted-BON + dv-cosine	Accuracy	97.4	# 1	✓	Compare

cosine similarity 를 사용하여 training document embeddings 함.

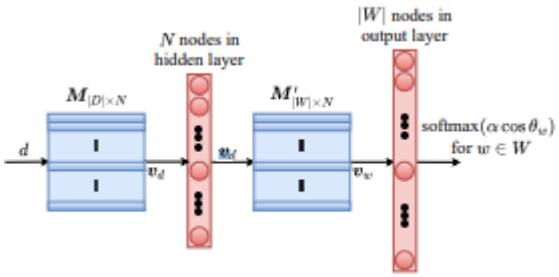

Figure 1: Proposed Architecture.

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 00 에서 가져옴 → …)

-

  데이터 불러오기 -> 불러온 데이터의 review 부분을 tokenize -> 데이터 전처리 및 데이터 탐색 -> 사전 학습된 모델 NB-weighted-BON + dv-cosine 에서 가져옴 -> 새로운 데이터로 결과 예측 및 평가지표확인 -> 결과 저장

- (Optional, 점수에 반영 X) 전통적인 방식

  Transfer Learning 이전에 사용했던 방식 중 TF-IDF 를 이용한 방법이 있습니다. TF-IDF 를 이용한다고 했을 때, 학습 과정을 간략하게 서술해주세요.

  —-

  데이터 불러오기 -> 데이터 전처리 후 말뭉치 생성/ BoW 벡터/단어 분포 탐색 -> 단어 별 빈도 분석 및 상위 빈도수 단어 출력 -> TF-IDF 변환/벡터-단어 맵핑/ 중요 단어 추출 Top 3 TF-IDF

5. 평가 방식

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

—-

True Positive(TP) : 실제 True 인 정답이 True 라고 예측(정답)

False Positive(FP) : 실제 False 인 정답이 True 라고 예측(오답)

False Negative(FN) : 실제 True 인 정답이 False 라고 예측(오답)

True Negative(TN) : 실제 False 인 정답이 False 라고 예측(정답)

1. 정확도 Accuracy: 전체 사례 중 예측이 맞은 비율

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

2. 정밀도 Precision : 모델이 True 라고 분류한 것 중에서 실제 True 인 것의 비율

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

3. 재현율 Recall : 실제 True 인 것 중에서 모델이 True 라고 예측한 것의 비율

$$Recall = \frac{TP}{(TP + FN)}$$

4. F1 스코어 F-measure :

보통가중치를 가진 조화평균

Precision 과 Recal 의 트레이드 오프를 잘 통합하여 정확성을 한번에 나타내는 지표

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$