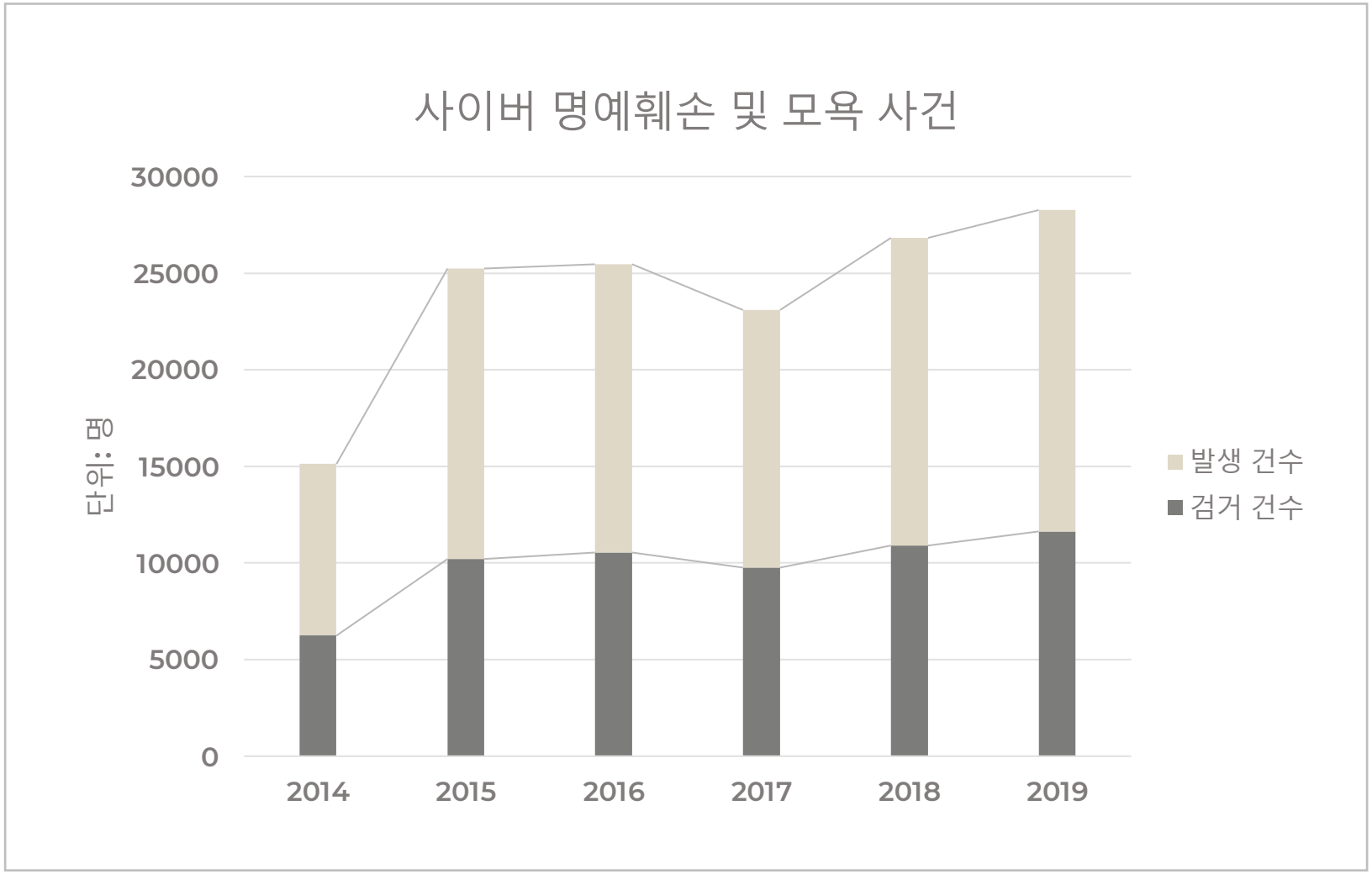
BERT 활용 악성 댓글 감지 시스템

AI 07 한 다 운

Contents

Part 1

프로젝트 목적



출처: 2018년 사이버폭력 실태조사 (방통위, 한국정보화진흥원)

비방/비난 댓글에 대한
솜방망이 법적 처벌

개인 SNS 댓글 테러 문화

악성 댓글 피해자의 고통

Part 1

프로젝트 목적

저것들이 가수냐 댄서지 댄서들 끝은 99프로 새드앤딩 그리고 철이없어도 너무없다

10+8 진짜 비교된다

인생패배자들 지금와서 삥뜯어볼려고 에라이 ..캬~~~악~~~~퉤슈

히트곡이 없는데 추억팔이가되는 놀라운 조선식 방송국연예매니징ㅋㅋ

배우로서 명성보단 ***로 더 유명한 대표적인 **거품배우**..

흑발미녀???**장난똥때리냐**?????마녀란 말도 아깝다!!!

자연어 처리 기술로 **비방 댓글을 감지**하여 **악플 사전 방지**

Part 1

가설 설정

가설 1. Pre-trained model 중 한국어로 학습시킨
bert-kor-base **모델의 정확도**가 가장 높을 것이다.

가설 2. 모델의 정확도는 0.5 이상일 것이다.

Part 2

Model Processing

데이터 전처리 | Pre-trained Model | 모델 학습 결과

데이터셋 (출처 :Koco, Korean Hate Speech Dataset)

comments	contain_gender_bias	bias	hate
송중기 시대극은 믿고본다. 첫회 신선하고 좋았다.	False	none	none
지** 나쁜놈	False	none	offensive
알바쓰고많이만들면되지 돈욕심없으면골목식당왜나온겨 기댕기게나하고	False	none	hate

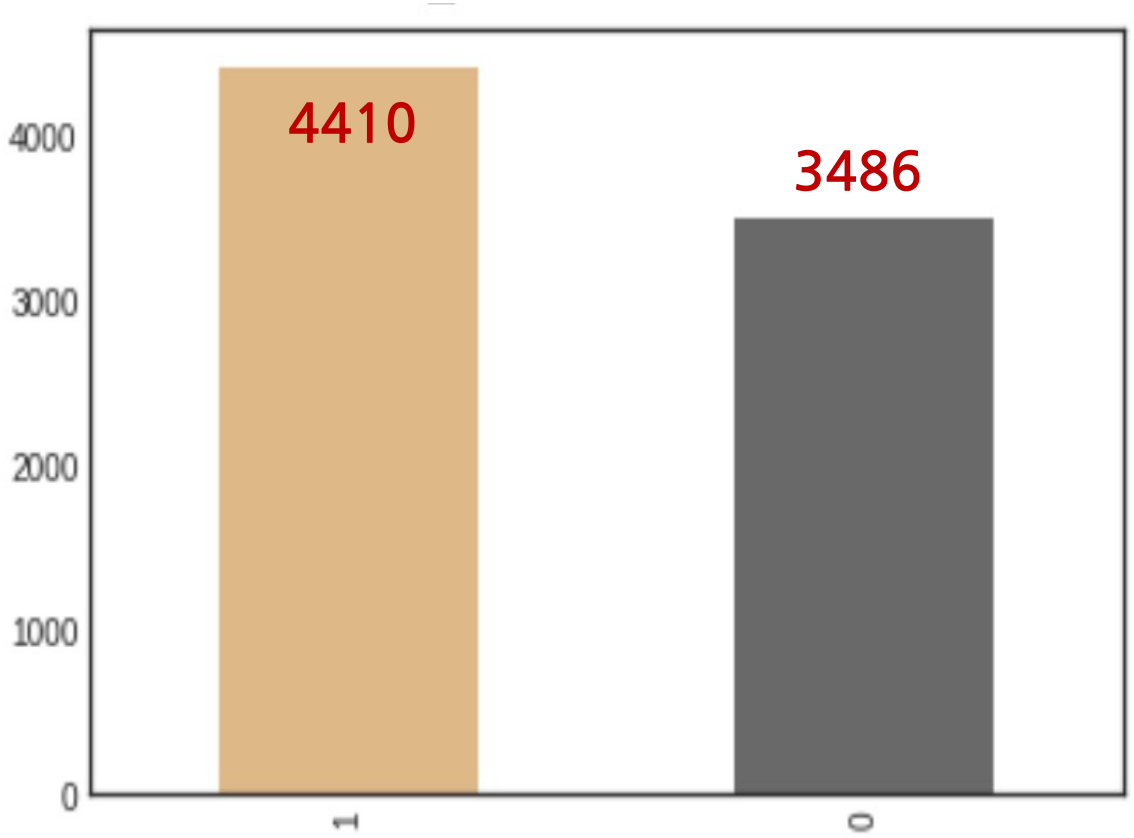
comments	hate
송중기 시대극은 믿고본다. 첫회 신선하고 좋았다.	0
지** 나쁜놈	1
알바쓰고많이만들면되지 돈욕심없으면골목식당왜나온겨 기댕기게나하고	1

이진 분류로 전환

Part 2

Model Processing

데이터 전처리 | Pre-trained Model | 모델 학습 결과



총 9,381 데이터.

4410 / (4410+3486) =0.558

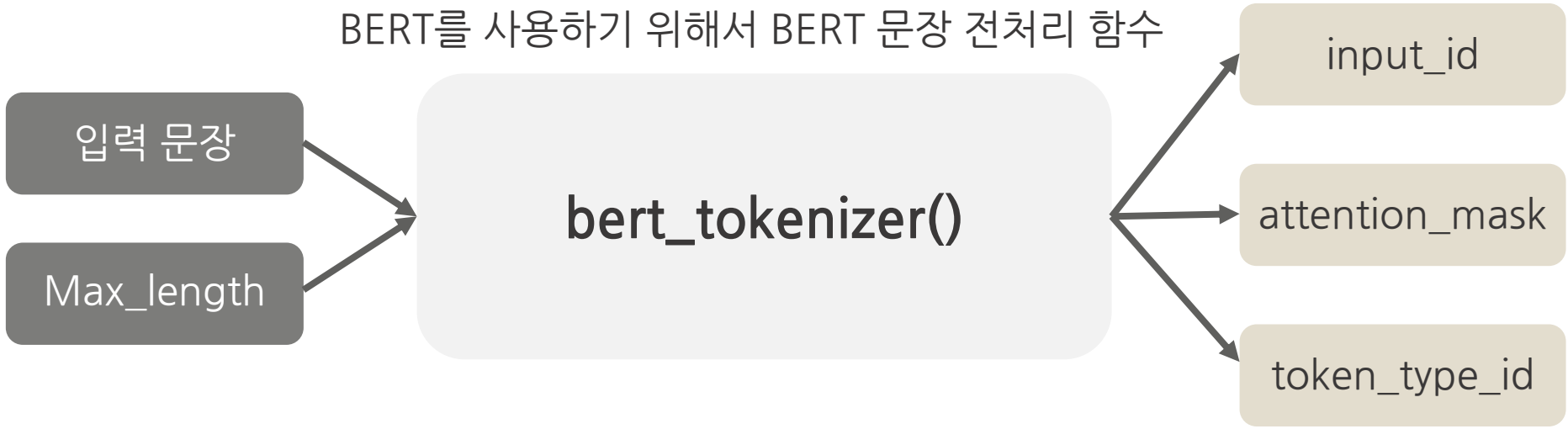
즉, 모델의 정확도가 최소 0.558이상 되어야 유의미하다.

Part 2

Model Processing

데이터 전처리 | Pre-trained Model | 모델 학습 결과

BERT를 사용하기 위해서 BERT 문장 전처리 함수



input_ids : 문장을 인덱스 값으로 변환. **[101 10150 116 0 0]**
attention_mask : 패딩된 부분이 학습에 영향 주지 않도록 처리. **[1 1 1 0 0]**
token_type_ids : 여러 문장이 들어 갔을 때 이를 구분. **[11110000]**

Part 2

Model Processing

데이터 전처리 | Pre-trained Model | 모델 학습 결과

STEP 1

Pre-trained model BERT base multilingual cased.
- 총 104개의 언어로 학습
- 대문자/소문자에 따라 다른 단어로 인식하는 점 주의.

TFBertClassifier.from_pretrained('bert-base-multilingual-cased')
TFBertClassifier.from_pretrained('kykim/bert-kor-base')

STEP 2

Dropout
- 과적합 방지

이진 분류 문제
- 해당 댓글이 비방인가 아닌가 분류하는 문제.

Tf.keras.layers.Dropout(self.bert.config.hidden_dropout_prob)
Num_class=2

STEP 3

Optimizer
- Adam : Momentum과 RMSprop의 장점 결합

Early stopping
- val_accuracy가 5번이 지나도 성능 개선이 되지 않으면 early stop 적용.

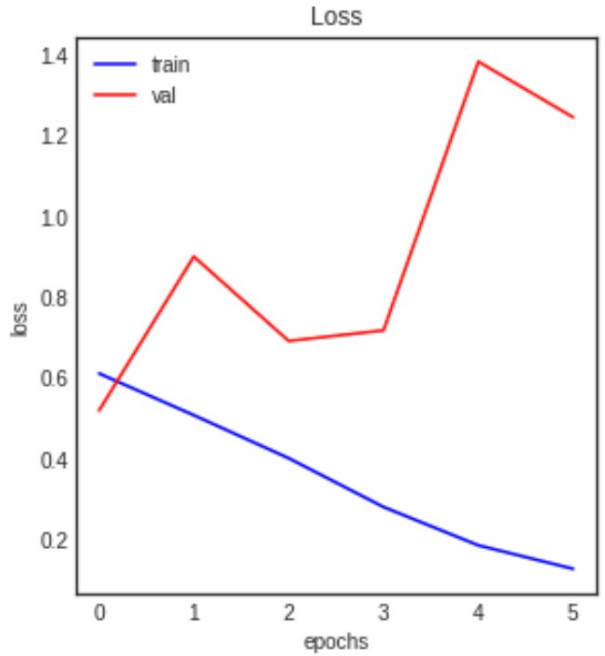
```
es_callback = EarlyStopping(monitor='val_accuracy', min_delta=0.0001
, patience=5)
```

Part 2

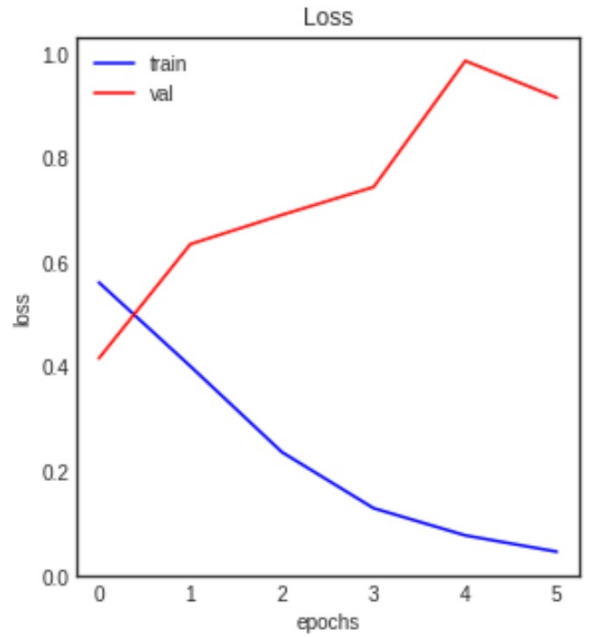
Model Processing

데이터 전처리 | Pre-trained Model | **모델 학습 결과**

bert-base-multilingual-cased



kykim/bert-kor-base



	Accuracy	Val_accuary	F1_Score
bert-base-multilingual-cased	0.952	0.711	0.750
kykim/bert-kor-base	0.984	0.804	0.839

Part 3

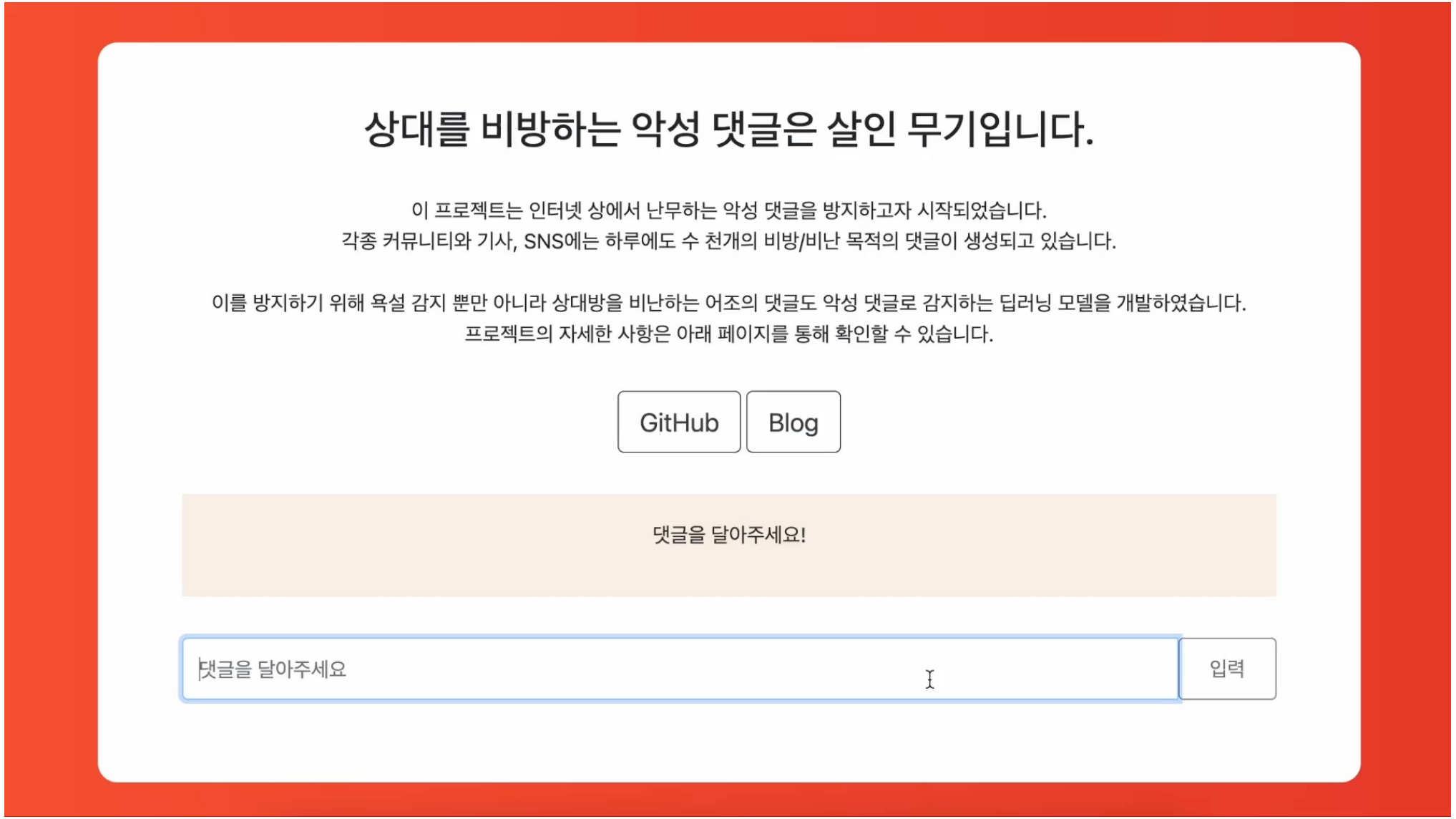
Results

비난/비방 의미

	Comment	bert-base-multilingual-cased	kykim/bert-kor-base
1	장** 얘도 참 이젠 짠하다...	1	1
2	누군데 얘네?	1	1
3	발연기 ** 국밥장인 팬들만 물고 빠는 ㅋㅋㅋ	1	1
4	참나 **아 말 함부로 하는거 아니다 니네 체인점 개판인데가 한두군데가 아니다가격은 드릅게 비싸요	1	1
5	김** 겁나 비호감인데 잘나가네. 방송을 잘하나베	1	1
6	슬금슬금 기어나올 생각말고 하던대로 그냥 조용히 살어라! 잠재적 살인마.	1	1
7	안타깝네요ㅜㅜ좋아하던 배우인데ㅜ	0	0
8	축하합니다! 예쁘게 잘~사세요~	1	1
9	신선하게 웃긴다ㅋㅋㅋ역시 동엽신~~!! 장소연님은 진짜 조선족인가 착각할정도로 말투가 리얼하네요	1	1

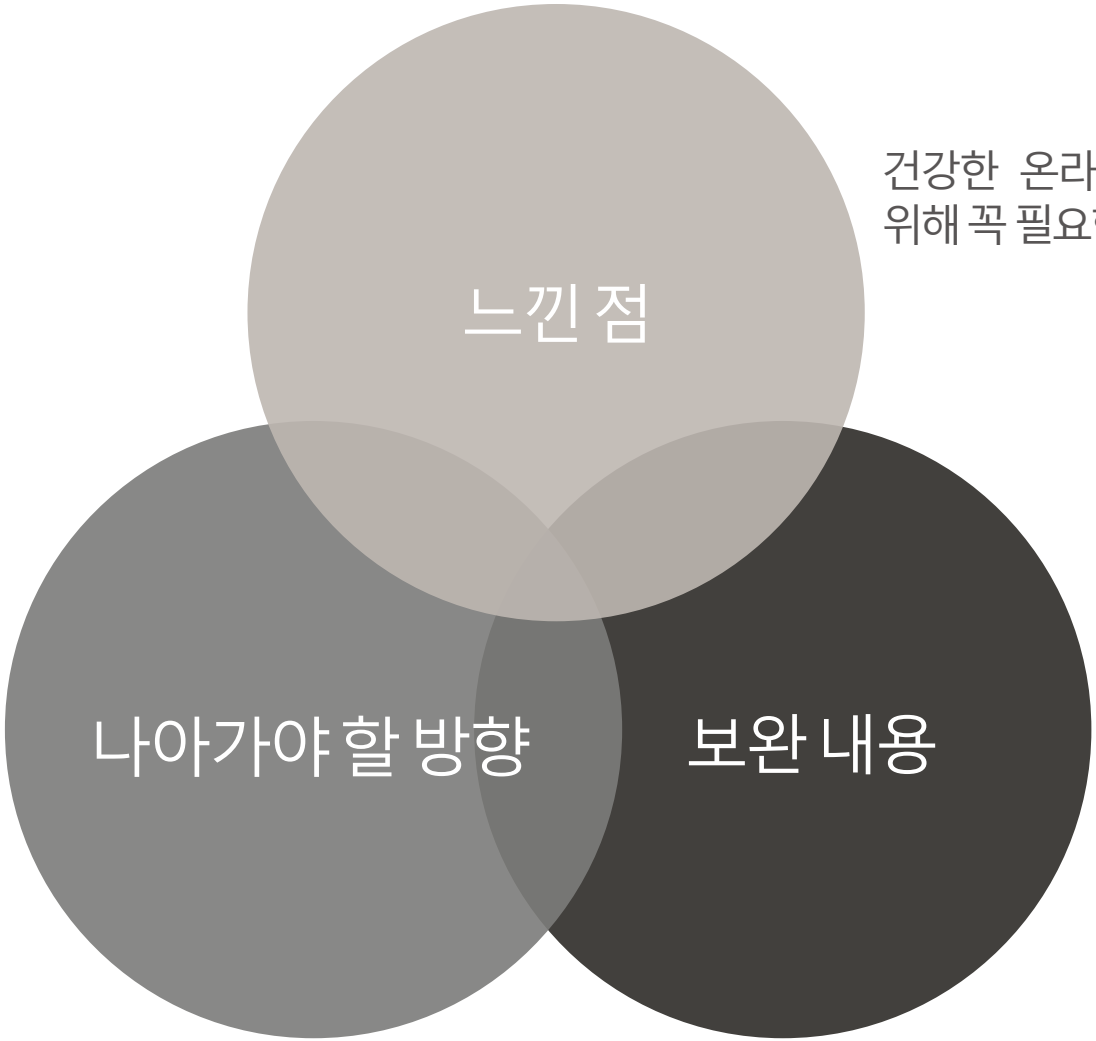
Part 3

Results



Part 3

Retrospect



건강한 온라인 문화를 만들기 위해 꼭 필요한 기술입니다.

댓글 기능을 모두 막는 것은 오히려 자유 발언 기회를 막는 행위가 될 수 있습니다.
따라서, 유저가 악플을 올리기 전, 해당 댓글이 비방이 될 수 있음을 인지시키는 시스템이나 페널티를 부과하는 시스템이 필요합니다.

데이터셋의 부족
댓글은 많지만, 비방 여부를 라벨링한 데이터는 적습니다.

감사합니다

AI 07 한 다 운
dawun.han@gmail.com