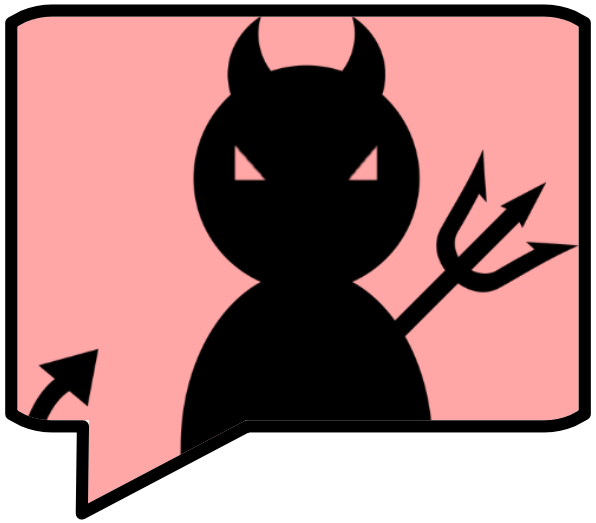
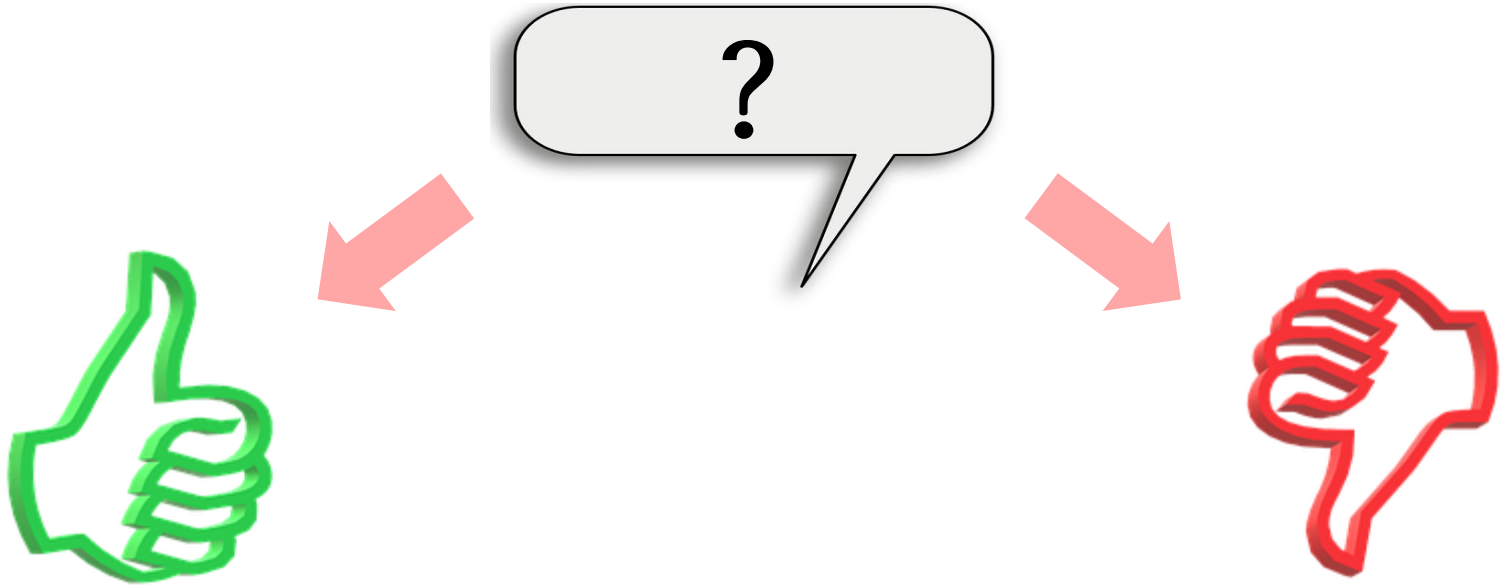


악플 분류 모델 만들기

소셜네트워크 2조
김하람, 박채은, 위효원, 이의동

주제 선정

sns 상에 달린 댓글이
악플인지 아닌지를 “분류”하는 모델 만들기

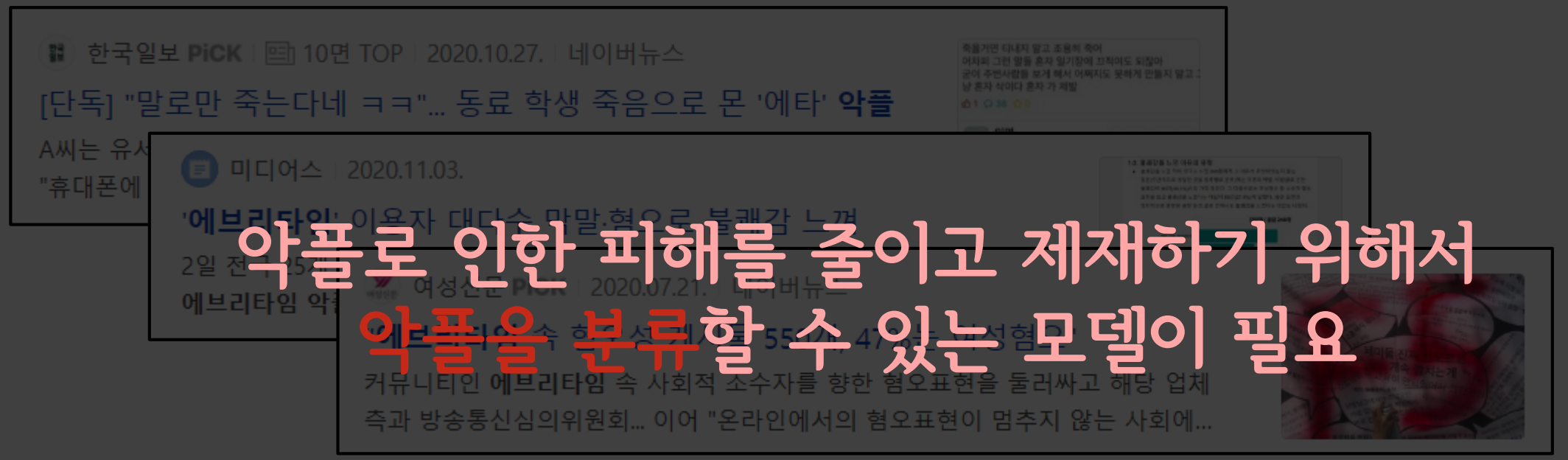


주제 선정 배경

대학교 익명 커뮤니티
에브리타임

악플로 인한 피해 심각

주제 선정 배경



악플로 인한 피해를 줄이고 제재하기 위해서
악플을 분류할 수 있는 모델이 필요

대학교 익명 커뮤니티
에브리타임

악플로 인한 피해 심각

사용한 데이터 소개

kaggle

Comments & ‘hate speech’ label

Comments 예

‘1,2화 어설펐는데 3,4화 지나서부터는 갈수록 너무 재밌던데’

‘hate speech’ label

hate	offensive	none

사용한 데이터 소개

kaggle

Comments & ‘hate speech’ label

기사 제목: """반드시 살려낸다"" \'골목식당\' 백종원, 약속 지켰다..성내동 大성공[어저께TV] "

hate

‘그입 닥쳐라 ..지금까지 골목시장에 늘어놓은 짜장면집 빽다방등...철수나 하셔....돈 많으넘이 돈 욕심낸다고’

offensive

‘지금이야 방송빨 타니깐 잘되지 몇년 지나봐야 안다’

none

‘옆동네는 눈물흘립니다’

사용한 데이터 소개

data set

train set

모델을 train 하는데 이용

dev set

성능 평가에 이용

데이터셋 처리

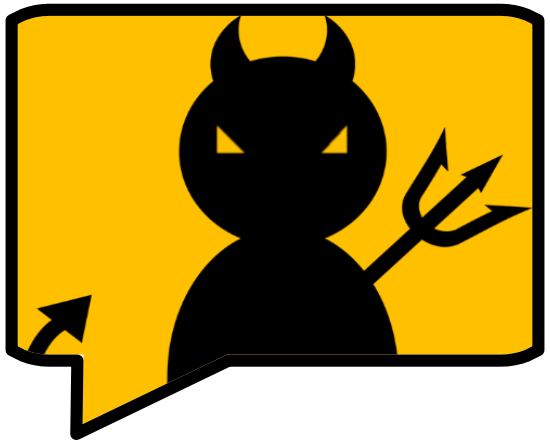
hate	offensive	none
1		0

공격적인 댓글도 악플로 분류

선정한 악플 분류 모델: SVM / BERT

SVM
1. 딥러닝 모델인 Bert와 비교해보고자 선택
2. SVM은 분류 목적으로 사용하는 모델
3. 신경망보다 사용하기 쉬움.
4. 딥러닝 이전에 많이 사용되었던 모델임.

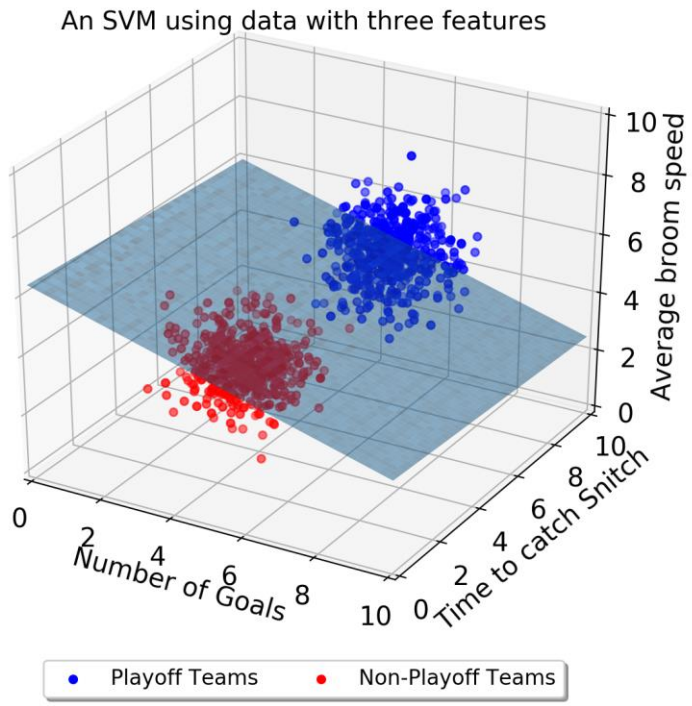
BERT(KOBERT)
1. 머신러닝 모델인 SVM과의 비교를 위해 선택
2. Google이 개발한 최신모델로 자연어처리 모델 중 가장 성능이 좋다고 알려짐.
3. 양방향성을 포함하여 문맥을 고려할 수 있음.



SVM 모델

머신러닝 모델

SVM이란



Support Vector Machine

- 결정 경계(분류를 위한 기준 선)을 정의하는 모델

- 즉, 최적의 결정 경계를 찾는 것
  -> 마진을 최대화하는 초평면을 찾기

Support Vectors: 결정 경계와 가까이 있는 데이터 포인터들

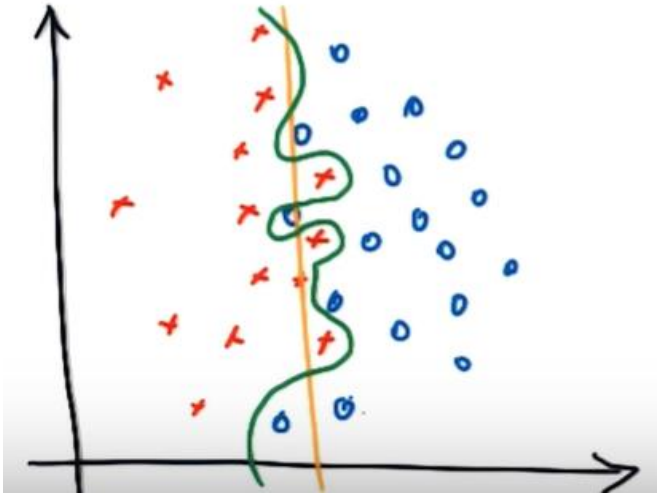
Margin: 결정 경계와 서포트 벡터 사이의 거리

Sklearn SVM 파라미터

C

Controls trade-off between smooth decision- boundary and classifying training points correctly

C값이 작을수록 오류를 더 많이 허용 -> 일반적인 결정 경계
C값이 클수록 오류를 덜 허용 -> 세심한 결정 경계



Sklearn SVM 파라미터

Kernel

선형적 분류가 되지 않는 저차원의 데이터를 고차원으로 매핑시키는 커널 함수를 결정

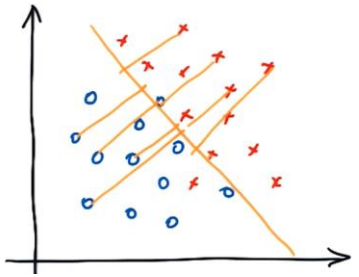
-> parameter로 linear, polynomial, sigmoid, rbf 등의 kernel을 선택

Sklearn SVM 파라미터

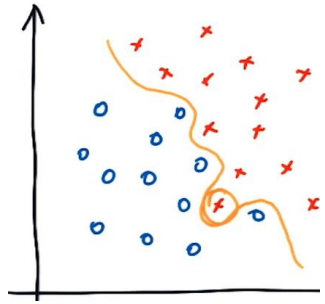
Gamma

Defines how far the influence of a single training point reaches

Decision boundary의 굴곡에 영향을 주는 데이터의 범위



Gamma가 작으면 reach가 멀다 -> 경계와 가까운 포인트의 영향이 상대적으로 적다
-> 경계가 직선에 가깝다



Gamma가 크면 reach가 가깝다 -> 경계와 가까운 포인트의 영향이 상대적으로 크다
-> 경계가 굴곡진다

SVM 모델

1. Comments 텍스트 전처리: mecab 사용해서 comments를 토큰화

정제(한글, 띄어쓰기만 남기기/ 불용어 제거), 품사 태깅, 중요 품사만 남기기

예) “10년만에 재미를 느끼는 프로였는데왜 니들때문에 폐지를해야되냐”
=> ['재미', '느끼', '프로', '왜', '폐지', '되']

2. Comments 임베딩: TFIDF 방식을 활용

TF, Word2vec, doc2vec보다 더 좋은 성능을 보임.

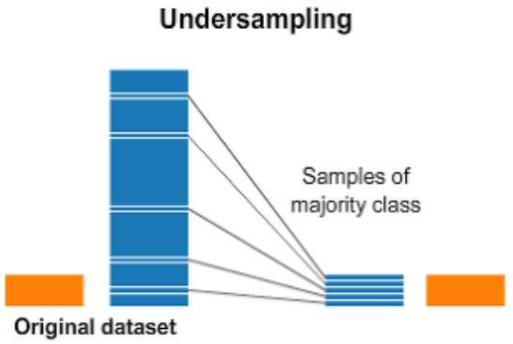
SVM 모델

3. Gridsearch로 hyperparameter 선정: C, kernel 등

C = 1, gamma = 1, kernel = 'rbf' 선정됨.

4. Best hyperparameter로 SVM 모델 적합

5. 성능 확인: dev set을 undersampling해서 accuracy와 f1 score 구함.

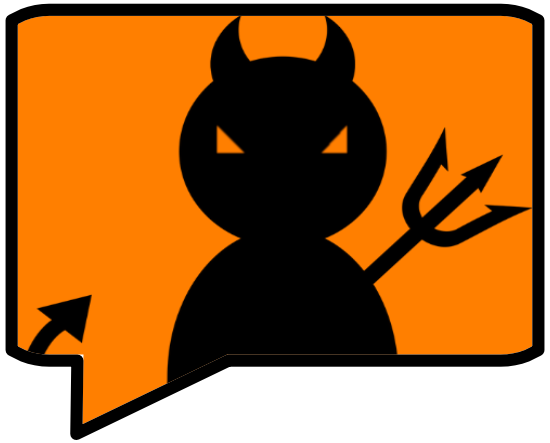


Confusion matrix

실제 \ 예측	악플X	악플
악플X	117	43
악플	35	125

Accuracy: 0.756

F1 score: 0.762



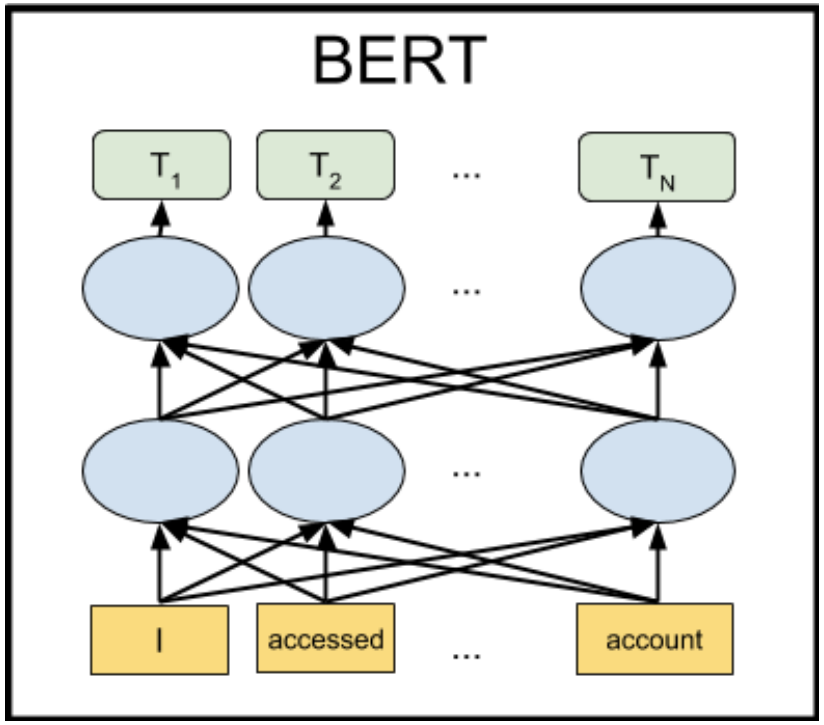
Bert 모델

딥러닝 모델

BERT by Devlin et al. 2018.

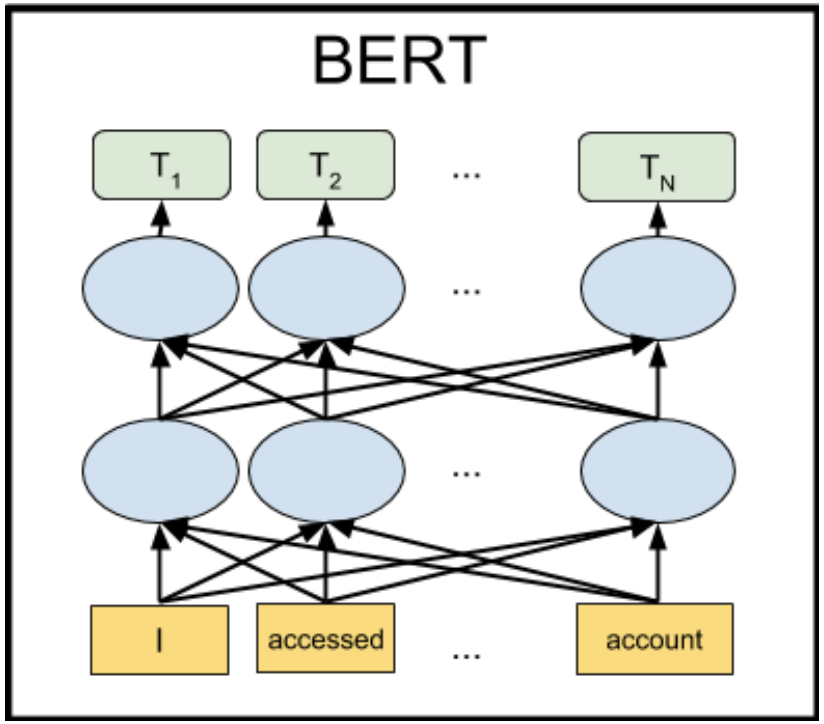
Pre-trained language model
- 문맥을 반영한 워드 임베딩

- Fine-tuning based model
Pre-trained된 parameter들이
downstream task 학습을 통해 fine-tuning됨



BERT by Devlin et al. 2018.

- Transformer(Vaswani et al. 2017)의 encoder 구조 사용
- Self-attention을 활용한 효율적인 학습



- Masked LM, Next sentence prediction 방법론을 사용

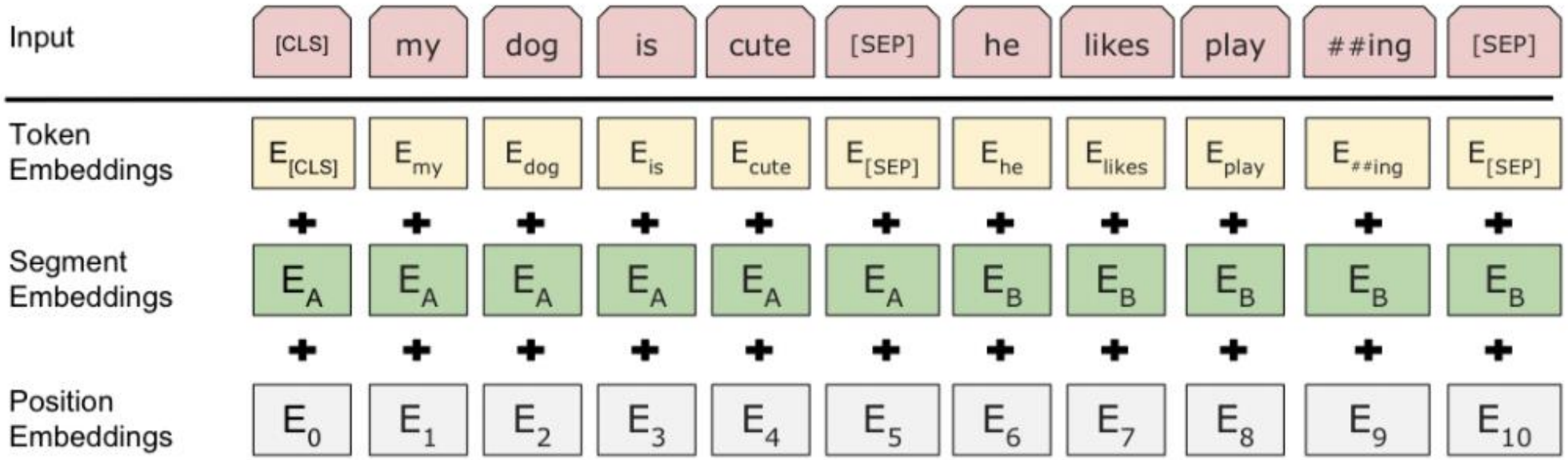
“Deep bidirectional”

Masked LM: 문장 내 Masked된 토큰을 예측

Next sentence prediction: 두 문장이 이어지는 문장인지 아닌지 판별

BERT Tokenize

Sentence -> Tokenize to tokens for BERT



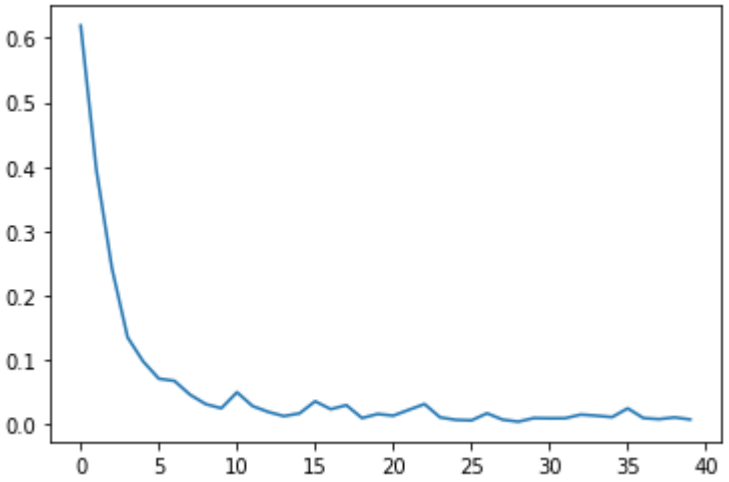
Bert detail

훈련 parameter 설정
batch_size = 64
dr_rate = 0.3
learning_rate = 5e-5

Loss function:
CrossEntropyloss

BERT result

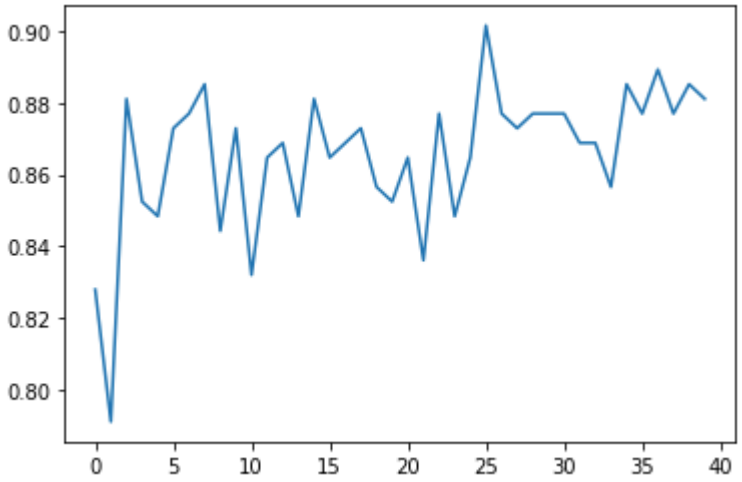
Train loss



최종 acc: 0.9979

최종 epoch: 40

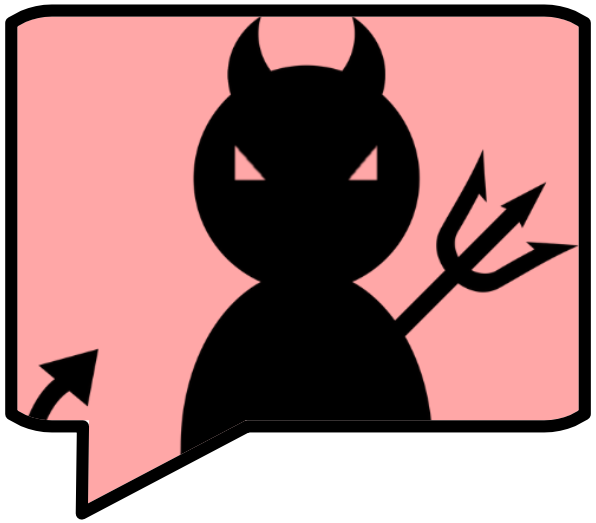
Test(Dev) acc



최종 acc: 0.8811

```
epoch 40 loss 0.008182252757251263 acc 0.9979008436203003
100% 4/4 [00:01<00:00, 2.28it/s]

epoch 40 dev acc 0.8811475038528442
```



악플 분류 모델 적용 - 에브리타임

에브리타임 게시판별 악플 비율 비교

데이터 소개

연세대학교 신촌캠 에브리타임 어플에 올라온 게시글 크롤링

데이터 소개

시간: 2021-03-17 기준 최근 100페이지의 데이터

게시판 종류:
- 자유게시판
- 새내기게시판
- 비밀게시판
- 시사, 이슈 게시판
- 정보게시판

SVM 모델 결과

게시판 종류	Hate(hate&offensive)	None
비밀 게시판	0.6677	0.3323
시사이슈 게시판	0.6064	0.3936
새내기 게시판	0.6063	0.3937
자유게시판	0.5619	0.4381
정보게시판	0.4051	0.5949

Bert 모델 결과

- ex) 시사, 이슈 게시판

Hate(hate & offensive)	None
"설탕세 뭐누??? 이제 더이상 세금 걷을 데가 없어서 설탕세를 걷누? 장난하나", “ㅇㅇㅇ 딸 홍대미대 입시청탁 의혹 뜨길래 조국처럼 터지나 하고 지켜봤는데 지원도 안했네ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 에라이~",	"혹시 업적, 잘한 일 궁금해서 알아보는 사람 있어? 어떤 사람들을 찾아봐?” "그만큼 지지하신다는 거지", "3자구조 ㅇㅇㅇ 1등이라고..? ㅋㅋㄱ 객관성 있는 자료인가",

Bert 모델 결과

게시판 종류	Hate(hate&offensive)	None
비밀 게시판	0.4341	0.5658
시사이슈 게시판	0.3381	0.6618
새내기 게시판	0.3084	0.6915
자유게시판	0.2236	0.7763
정보게시판	0.1172	0.8227

최종 결론

게시판 종류/ 악플 비율	SVM	BERT
비밀게시판	0.6677	0.4341
시사이슈 게시판	0.6064	0.3381
새내기 게시판	0.6063	0.3084
자유게시판	0.5619	0.2236
정보게시판	0.4051	0.1172

1. 공통점: 순위
에타 게시글이라는 new data에 대해
같은 순위를 매긴 것으로 보아
두 모델이 악플 분류를 잘 수행하고 있음을
알 수 있음.

2. 차이점: 비율
버츠가 SVM에 비해 악플을 더 보수적으로
판단했다. 절대적인 숫자 자체가 작다.

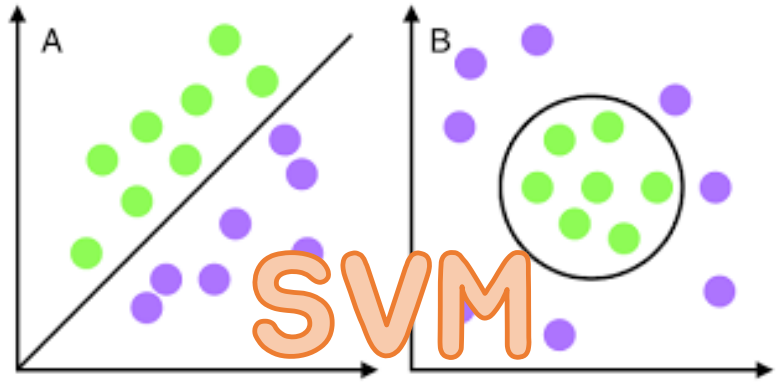
최종 결론

3. 모델의 accuracy에서 SVM이 0.76,
버츠가 0.88 정도로 버츠가 조금 더 좋은 성능을 보였다.

Google AI

NLP BERT

0.88



0.76