DEVIEW 2020

조금 더 아름다운 댓글 경험을 위해서 Task 정의 부터 서비스까지 Cleanbot 2 개발기

최경호 미디어인텔리전스
김성민 미디어인텔리전스

CONTENTS

1. 우리가 고민한 악플

사람마다 다른 악플의 기준

1.1 표현의 자유 vs. 누군가의 상처

DEVIEW 2020

표현의 자유

세계인권선언 제19조
https://www.un.org/en/universal-declaration-human-rights/index.html

Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

모든 사람은 의견과 표현의 자유를 가질 권리가 있다. 이 권리에는 간섭 없이 의견을 가질 수 있는 자유가 포함되며, 어떤 미디어를 통해서든 정보와 사상을 찾고, 받고, 전달할 수 있는 자유가 포함된다.
(파파고 고마워요)

1.1 표현의 자유 vs. 누군가의 상처

악플로 인해 다치는 사람이 없게 하자는 궁극적인 목표를 위해,
악플이 미치는 피해에 대한 메커니즘을 확인하고자,
악플 피해에 대한 심리상담학적 선행 조사나, 연구를 통해 실마리를
얻고자 싶었으나…

1.2 법적, 학술적 배경

RISS 처음 방문이세요? 로그인 회원가입 MyRISS 내

검색키워드 악플 (검색결과 261 건)

학위논문 32

제한적 본인확인제의 효과성에 관한 연구 :악플 피해 사건과 이명박 정부에 대한 악플 게재 여부를 중심으로
이광원 | 韓國外國語大學校 大學院 | 2008 | 국내석사
원문보기 | 목차검색조회 ▾ | 음성듣기

인터넷 뉴스 댓글이 원문과 태도에 미치는영향 :정치적·윤리적 요인의 선플, 악플 유형을 중심으로
정영라 | 성균관대학교 | 2014 | 국내석사
원문보기 | 목차검색조회 ▾ | 음성듣기

악플 경험자와 비경험자의 특성 비교연구 :인구사회학적, 심리적, 환경적 요인을 중심으로
권미애 | 부경대학교 대학원 | 2008 | 국내석사
원문보기 | 음성듣기

국내학술논문 44

F 무료 M 기관 내 무료 W 유료 | 기관별 국내

KCI 등재
인터넷의 공공성과 '악플'에 대한 대처 방안에 관한 연구
김민기,이진로 | 한국정치커뮤니케이션학회 | 2008 | 정치커뮤니케이션 연구 | Vol.9 No.-
W 원문보기

KCI 등재후보
인터넷의 공공성과'악플'에 대한 대처 방안에 관한 연구
김민기,이진로 | 한국정치커뮤니케이션학회 | 2008 | 정치커뮤니케이션 연구 | Vol.0 No.9
F 원문보기

인터넷상의 야누스, 악플의 사회심리학
나은영 | 언론중재위원회 | 2015 | 言論仲裁 | Vol.13 No.-
W 원문보기

인터넷상의 야누스, 악플의 사회심리학

저자	나은영
발행기관	언론중재위원회
학술지명	言論仲裁(Press arbitration quarterly)
권호사항	Vol.13 No.- [2015]
발행연도	2015
작성언어	Korean
자료형태	학술저널
수록면	16-27(12쪽)
제공처	W eArticle

W 원문보기 인용하기

부가정보

목차 (Table of Contents)

1. 악플, 그 정체는?
2. 악플의 일반적 원인
(1) 익명성으로 인한 탈억제
(2) 순간적인 감정의 배설과 자존감 유지
(3) 스트레스의 배출구

1.2 법적, 학술적 배경

제 검색 내공이 부족하여

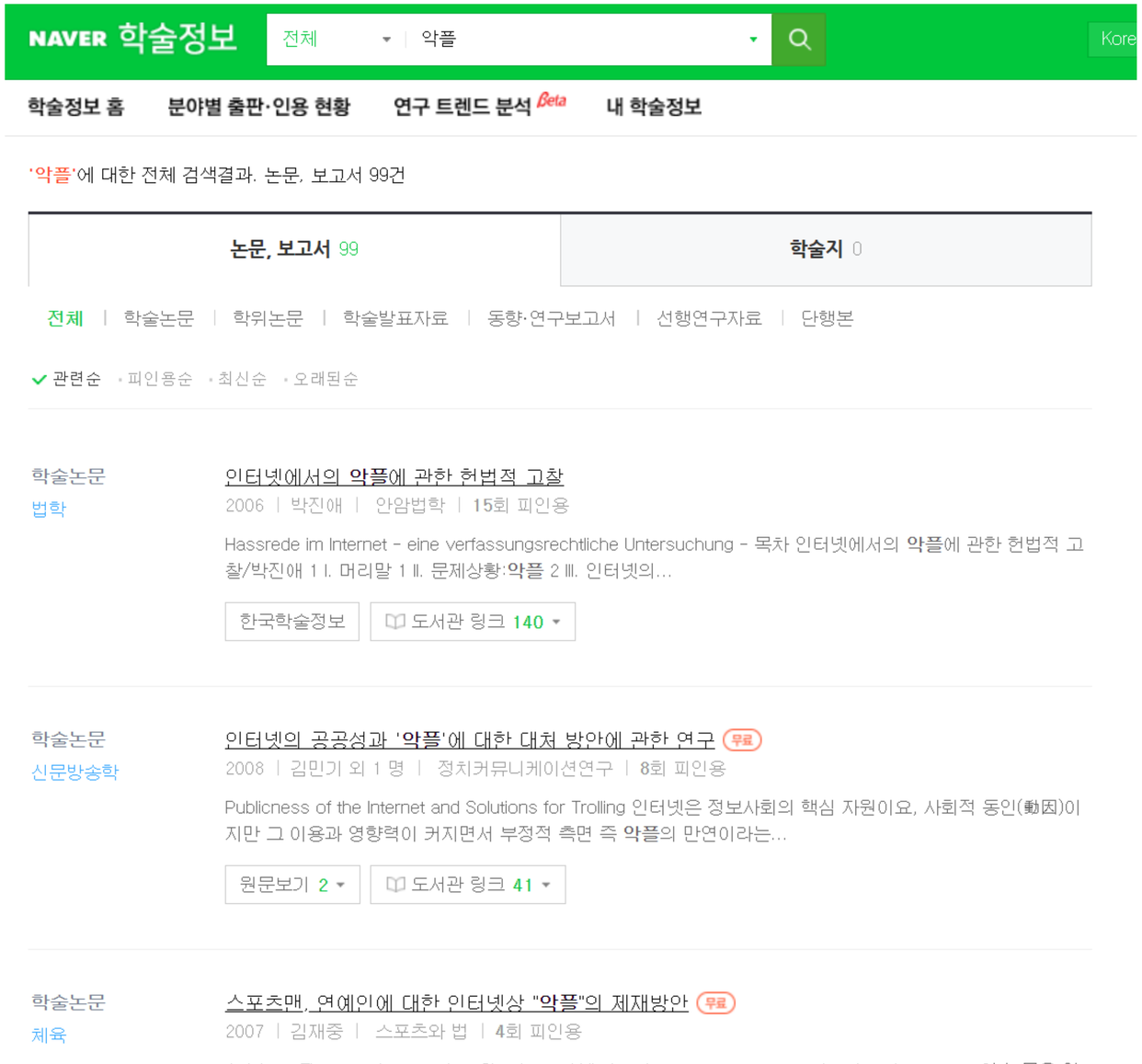
- 왜 악플이 발생하는지
- 법적으로 어떤 지

에 대한 연구 자료밖에 찾지 못했습니다.

ㅠㅠ

그럼에도…

데이터 냄새를 잘 맡는 ML Engineer다 보니



1.2 법적, 학술적 배경

데이터와 판례를 찾아보게 됩니다.

케글 Toxic Comment Classification Challenge

- https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview/evaluation

https://www.lawtimes.co.kr/Case-Curation

1.2 법적, 학술적 배경

실제 판례를 공학적으로 분석하여 확인한 결과

판결에는

- TEXT 자체의 의미와 악의성 뿐만 아니라
- 사건의 연속성
- 원고/피고의 상태와 의도
- 변론의 유무
- 형사와 민사
- 원심과 상고

등의 다양한 독립변인이 고려되어 있었습니다.

1.2 법적, 학술적 배경

DEVIEW 2020

뇌절했나보네ㅋㅋㅋ @ [부고] 하타케 가문의 카X시씨 벤트 앞에서 변사체로 발견… → ?

뇌절했나보네ㅋㅋㅋ @ [속보] 하타케 가문의 카X시씨 벤트타는 민트를 목격하여… → ?

판례를 이용하여 데이터를 만들게 되면 오히려

Task def에 없는 feature(독립변인)들이 더 큰 영향력을 가져 모델에게 도움이 되지 않은 것이고,

동시에 판례들은 데이터로 쓰기에 너무나 적고 구하기 힘들었습니다.

1.3 사회적 합의와 개인의 감성

사회적 합의의 기준은 지상파 방송에서 사용되는 언어 수준이라 생각하여
방통위 심의 기준과 심의 자료를 찾아 보았으나…

<방송통신심의위원회 SafeNet 등급기준>

	노출	성행위	폭력	언어	기타
4등급	성기노출	성범죄 또는 노골적인 성행위	잔인한 살해	노골적이고 외설적인 비속어	1. -마약사용조장 -무기사용조장 -도박 2. -음주조장 -흡연조장
3등급	전신노출	노골적이지 않은 성행위	살해	심한 비속어	
2등급	부분노출	착의상태의 성적접촉	상해	거친 비속어	
1등급	노출복장	격렬한 키스	격투	일상 비속어	
0등급	노출없음	성행위없음	폭력없음	비속어없음	

http://www.safenet.ne.kr/dstandard.do

1.3 사회적 합의와 개인의 감성

“착짱죽짱” 은 악플인가요?

“자낳괴ㅋㅋㅋㅋㅋ”는 악플인가요?

“Tlqkf ㅋㅋㅋㅋㅋㅋ” 는요?

자신이 아끼는 사람이 중국인 이라고 하면요.

자신이 유튜버 라면요.

그 댓글이 자신의 업적이 기록된 기사 있다면요.

1. 결론

"악플"은 선행 연구가 많지 않은 새롭고 민감한 주제이기에

또 다음과 같은 특징 때문에

기술적으로 사용하기 위한 요구사항 명세 수준의 정의는 어려웠습니다.

- 모든 세대의 모든 이용자를 만족하는 악플에 대한 정의는 없음
- 아무에게도 상처주지 않는다는 것을 보장 불가
- 시간이 지남에 따라 같은 글이 갖는 의미가 끼치는 악영향의 정도 변함
  e.g. 존맛, 존멋, 존예, 탈룰라, 뇌절, 위키너드, 넌씨눈

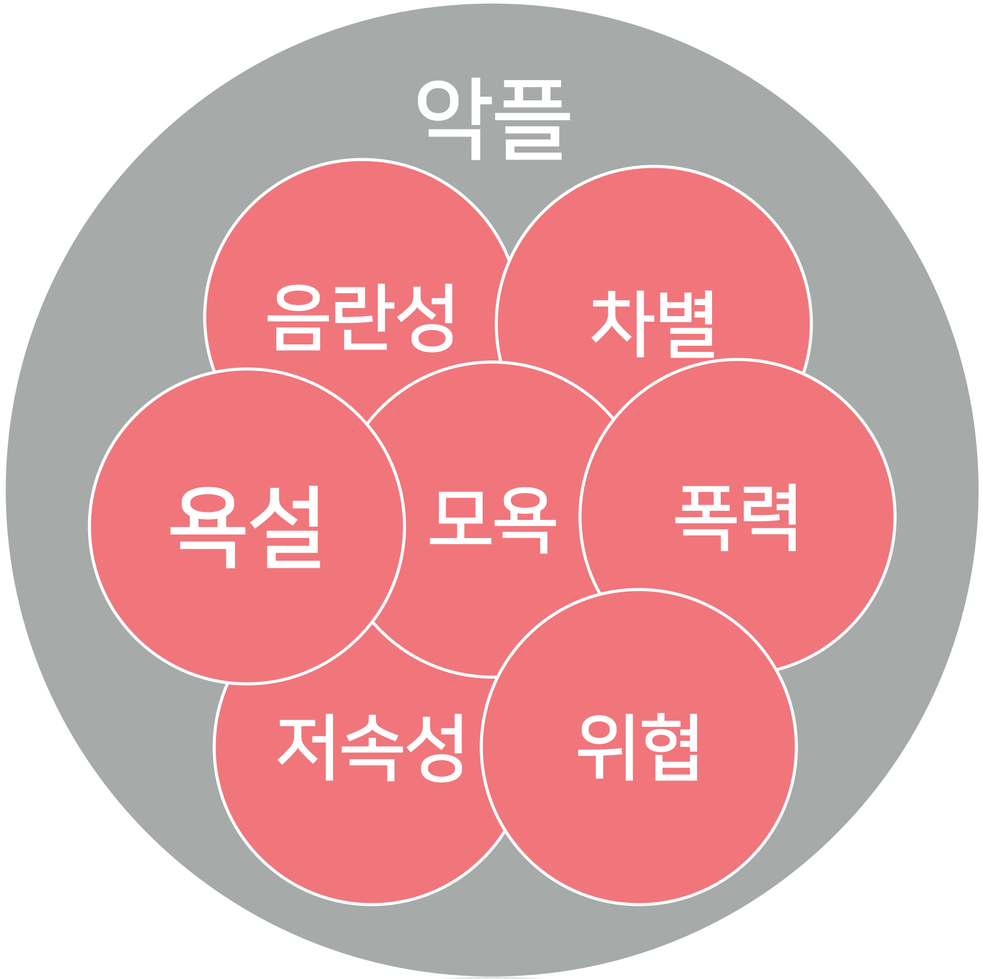
선행 연구와, 가공하여 사용할 수 있는 데이터도 구할 수 없었습니다.

때문에 우리는 공학적으로 접근을 하게 되었습니다.

1. 결론

때문에 우리는 클린봇에게 학습시킬 악플의 범위를
오른쪽과 같이 정의해 보았습니다.

또 각 항목에 대한 개인간의 의견차이는
통계로 극복 한다라는 전략을 취하기로 하였습니다.



2. 데이터 만들기

둘이 태깅하다 하나가 죽어도 모르게....

2.1 태깅 규칙 정하기

같은 댓글을 보더라도 각 개인이 느끼는 불쾌함과, 악성의 정도는 다릅니다.

데이터의 일관성이 있어야, 해당 분포를 모델이 잘 학습할 수 있으며, 결과물로 만들어진 서비스도 일관성을 가질 수 있습니다.

태깅 대상에 대한 정의를 **구체적**으로 할 수록 일관된 태깅을 할 수 있게 되나,
태깅 규칙이 복잡할 수록, **태깅 비용**이 비싸지며,
작업자가 **실수할 여지**가 커집니다.

또 훈련되지 않은 작업자의 경우 개인의 의견이 데이터에 반영될 여지가 있습니다.

2.1 태깅 규칙 정하기

쉽고 정확한 태깅을 위하여 인지 심리학 논문을 참고하였습니다.

가능하면 문화에 따라 형성되는 보편적인 어휘의 인지 경계에 포함되는 태깅 규칙은 피합니다.

- Ref. 홍기선. (2003). 언어와 사고: 보편적 인지와 개별언어의 상호작용. 담화·인지언어학회 학술대회 발표논문집, (), 3-13.

- 태깅 규칙을 바탕으로 같은 데이터를 다수의 연구자가 태깅하여 결과를 비교합니다.

- “태깅 규칙”을 결과물이 같아질 때까지 수정합니다.

태깅 대상이 태깅 규칙의 경계에 있는 경우 데이터의 일관성이 낮아질 여지가 높습니다.

- Ref. 이경수,송현주,손영우,황명진,and 박영실. "통계 조사 방법에 대한 인지심리학적 접근: 선택지의 개수와 순서에 따른 응답의 변화." 통계연구 13.2 (2008): 149-181.

- 어떤 레이블로 태깅할 지 고민이 든 경우 해당 댓글을 데이터에서 제외합니다.

- 즉 해당 댓글은 다른 태깅된 데이터를 통해 AI가 학습한 분포를 기준으로 평가하게 될 것이라 판단하였습니다.

2.2 데이터 구축하기

태깅되지 않은 댓글 데이터는 많습니다.

우선 빠르게 소수의 데이터를 태깅합니다.

소수의 데이터로 학습된 모델로 샘플링된 데이터를 돌립니다.

세상 사람들의 평균적인 표현이 그렇게 악랄하지 않고, 악의의 표현 방식이 다양하므로 스코어가 높은 댓글은 얼마 되지 않습니다.

- 악랄하지 않다. ➔ POS set 이 적다 ➔ 데이터 불균형
- 악의의 표현이 다양하다. ➔ NEG가 된 데이터에 악의의 표현이 많을 수 있다.

때문에 스코어 구간별로 무작위 추출을 하여 태깅할 댓글을 정합니다.

이를 반복하여, 원하는 수준의 F1 성능이 나올 때까지 반복합니다.

- 이때 test set은 고정합니다.

2.3 데이터 검증

앞서 밝힌 바 대로 악플의 정의는 사람마다, 그리고 같은 사람조차도 시간과 상황에 따라 다르게 판단합니다.

때문에 아무리 task def를 잘 했더라도 데이터에 오류가 녹아들 수 있습니다.

먼저 test set에서 FP, FN 케이스를 확인합니다.

- 데이터에 오류가 있다면 수정합니다.

Train set으로 train set을 평가하여 오류케이스를 확인합니다.

- 이때 모델이 실수하였을 수도 있고, 사람이 잘못 태깅했을 수도 있습니다. 더 이상 잘못된 태깅이 없을 때까지 고칩니다.

아직 끝이 아닙니다. Train set을 fold 하여 학습하고 infer하여 오류케이스를 확인합니다.

- 마찬가지로 잘못된 태깅이 확인되면 수정하고 반복합니다.

2.3 데이터 검증

레이블 노이즈 제거 전/후 모델 테스트셋 정확도 비교

Model	Acc. / F1 (Before)	Acc. / F1 (After)
CNN	0.9258 / 0.8847	**0.9355 / 0.9012**
BiLSTM	0.9350 / 0.8991	**0.9450 / 0.9129**
BiLSTM + LSTM	0.9392 / 0.9048	**0.9456 / 0.9159**
CNN + BiLSTM + LSTM	0.9397 / 0.9070	**0.9480 / 0.9178**

3. 모델링 하기

SOTA와 Serviceability

3. 모델링 하기

- 범용 장비에서 inference 에 대한
  높은 qps와 낮은 latency 보장(infer 비용 고려 필요)
- 데이터 확장 시 빠른 재학습 및
  AB test에서 적은 diff 보장
- 요구사항 변경 시 빠른 재학습 및
  최대한 목표로 설정한 diff만 발생

- 최고의 metric 상의 성능
- train 과 infer에 필요한 비용 고려하지 않음
- 모델과 기술에 대한 홍보 효과

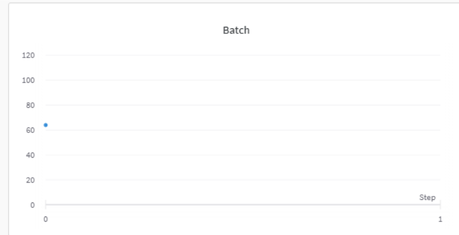
3.1 Test models

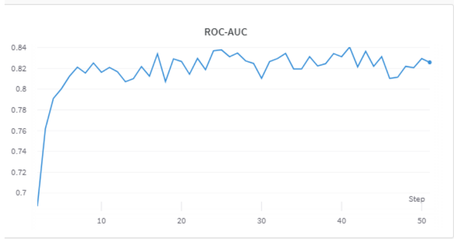
DEVIEW 2020

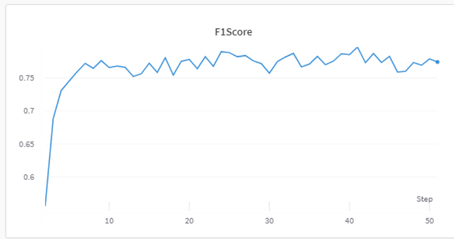
과제를 해결하기 위하여 모델은 어느 정도의 해상력을 가져야 할 지 확인해 봅니다..

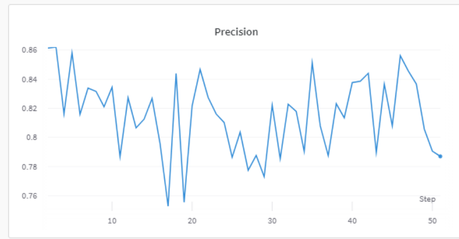
※ 이후 등장할 도표에서 각 테이블마다 사용한 데이터는 다릅니다.

Model	Accuracy	F1 Score
CNN	0.9355	0.9012
BiLSTM	0.9450	0.9129
BiLSTM + LSTM	0.9456	0.9159
CNN + BiLSTM + LSTM	0.9480	0.9178
CNN + BiLSTM + LSTM (Clova AutoML)	0.9496	0.9214









3.1 Test models

비용을 고려하지 않기도 해보고요.

Model	Accuracy	Precision	Recall	F1 Score
BERT 6-layer	0.9692	0.7597	0.6348	0.6917
BERT 12-layer	0.971	0.7629	0.6769	0.7173
BERT 12-layer + Dialog	0.9714	0.7683	0.6787	0.7208
RoBERT 24-Layer	0.9728	0.7649	0.7419	0.7483
CNN_LSTM(x2)	0.9672	0.6877	0.7284	0.7075
CNN_LSTM(x2) + PerEmb	0.9686	0.6969	0.9476	0.7214

3.1 Test models

모델을 찾기 위한 실험 결과 0.4% 의 acc 그리고 2.7% F1을 위해 24-layer의 RoBERT는 조금 비싸다는 판단을 하였습니다.

Model	Accuracy	Precision	Recall	F1 Score
BERT 6-layer	0.9692	0.7597	0.6348	0.6917
BERT 12-layer	0.971	0.7629	0.6769	0.7173
BERT 12-layer + Dialog	0.9714	0.7683	0.6787	0.7208
RoBERT 24-Layer	0.9728	0.7649	0.7419	0.7483
CNN_LSTM(x2)	0.9672	0.6877	0.7284	0.7075
CNN_LSTM(x2) + PerEmb	0.9686	0.6969	0.9476	0.7214

때문에 BERT 보다는 조금 낮은 정확도를 보이지만 CNN_LSTM(x2) + PerEmb를 사용하기로 하였습니다.

PerEmb(Persona Embedding)은 이어서 설명드립니다.

3.2 Persona Embedding

Persona Embedding이란 댓글 분류 모델의 퍼포먼스를 향상 시키기 위해 자체적으로 고안한 전이 학습 (Transfer Learning) 방법론으로, 두 댓글 간의 유사도를 학습함으로써 댓글에 대한 Representation Learning을 수행하도록 합니다.

두 댓글 간의 유사도는 일반적으로 토큰화 과정 이후 코사인 유사도를 통해 기계적인 산출이 가능한데, 이는 댓글 간의 의미적인 유사성을 고려하지 못합니다. Persona Embedding 과정에서는 두 댓글 간에 산출된 코사인 유사도에 대하여 작성자 동일성 여부에 따른 비선형 변환을 수행하여 새로운 유사도를 얻어내어 이를 정답 스코어로 간주하는 것이 특징입니다.

3.2 Persona Embedding

작성자가 동일한 두 댓글과 코사인 유사도

댓글1: 드마커스 커즌스 사주셔서 감사합니다!!!

댓글2: 커즌스 론도면 그래도 나름 잘잡았다 론도가 클러갔음 대박이였을것같은데

Cosine Similarity(댓글1, 댓글2) =

0.123

3.2 Persona Embedding

작성자가 다른 임의 두 댓글과 코사인 유사도

댓글1: 와 음원 0점 실화냐? ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ

댓글2: 사진 표정 진짜 웃기다ㅋㅋㅋㅋㅋ

Cosine Similarity(댓글1, 댓글2) =

0.701

3.2 Persona Embedding

데이터 형태: (x1, x2, x3, y1, y2)

x1, x2, x3: 댓글 내용, x1과 x2는 작성자가 서로 동일, x3는 작성자 다른 임의 댓글

y1: 동일 사용자가 작성한 두 댓글간의 코사인 유사도를 산출하고 이를 과소평가된 것으로 보고 그보다 적절하게 상향된 값을 두 댓글간의 의미적 유사도로 간주

y2: 임의 두 댓글간의 코사인을 산출하고 이를 과대평가된 것으로 보고 그보다 적절하게 하향된 값을 두 댓글간의 의미적 유사도로 간주

3.2 Persona Embedding

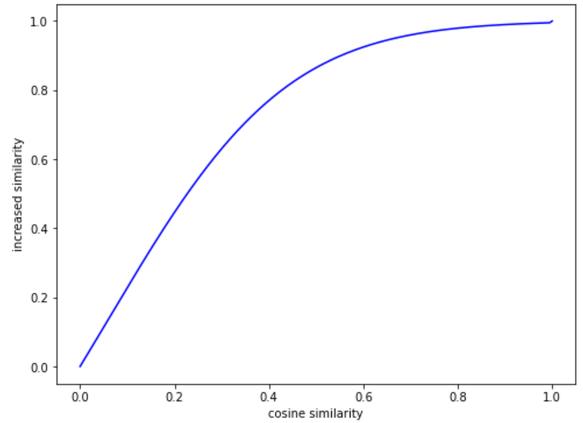
$$s\_up(c) = max(c, (\sigma[(\alpha - (\beta - \beta c))c] - 0.5) * 2.0)$$

$$s\_down(c) = min(c, \sigma(\alpha c - \alpha) * 2.0)$$

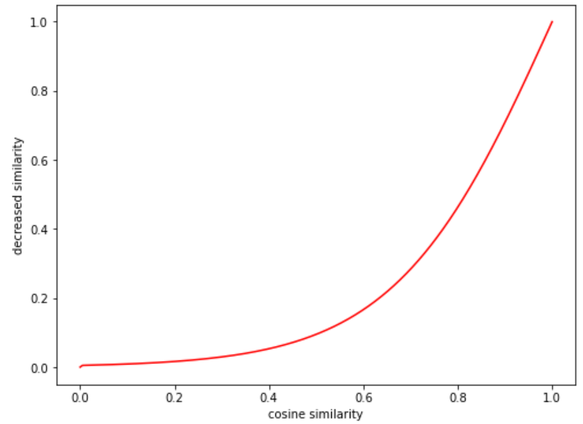
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$c = cos(\theta),\ 0 \leq \theta \leq \frac{\pi}{2}$$

$$\alpha, \beta > 0$$



유사도 상향 함수



유사도 하향 함수

3.2 Persona Embedding

정답 유사도 계산의 예 (작성자가 동일한 두 댓글)

댓글1: 드마커스 커즌스 사주셔서 감사합니다!!!

댓글2: 커즌스 론도면 그래도 나름 잘잡았다 론도가 클러갔음 대박이였을것같은데

Cosine Similarity(댓글1, 댓글2) =

0.123

0.28

유사도 상향 함수

3.2 Persona Embedding

정답 유사도 계산의 예 (임의 두 댓글)

댓글1: 와 음원 0점 실화냐? ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ

댓글2: 사진 표정 진짜 웃기다ㅋㅋㅋㅋㅋ

Cosine Similarity(댓글1, 댓글2) =

0.701

0.46

4

유사도 하향 함수

3.2 Persona Embedding

DEVIEW 2020

유사도 계산 함수

$$s(x1, x2) = e^{|F(x1)-F(x2)|}$$

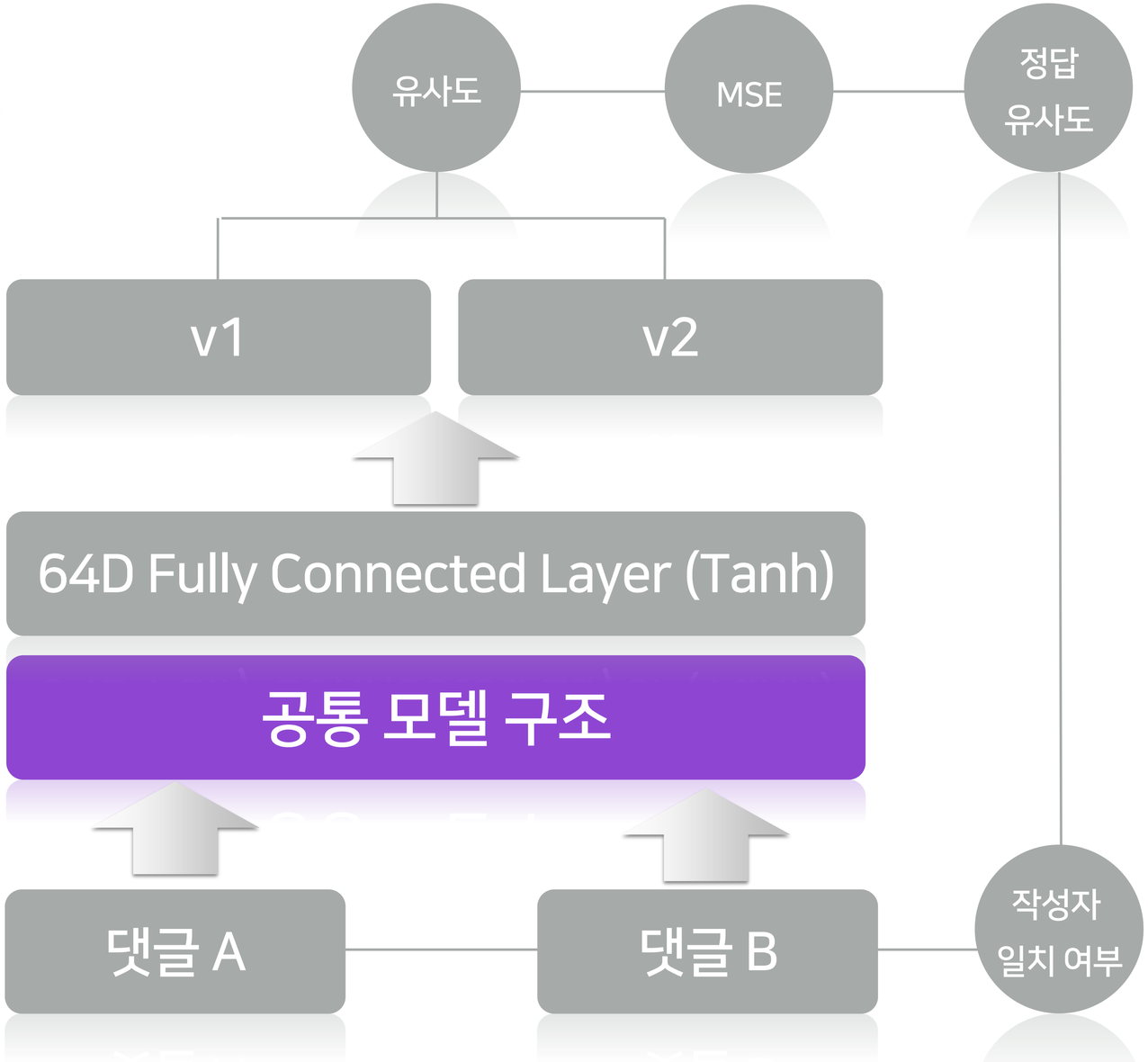
손실 함수

$$L(x1, x2, y) = yMSE(s\_up(c), s(x1, x2)) + (1 - y)MSE(s\_down(c), s(x1, x2))$$

$$MSE(a, b) = \sqrt{(a - b)^2}$$

3.2 Persona Embedding

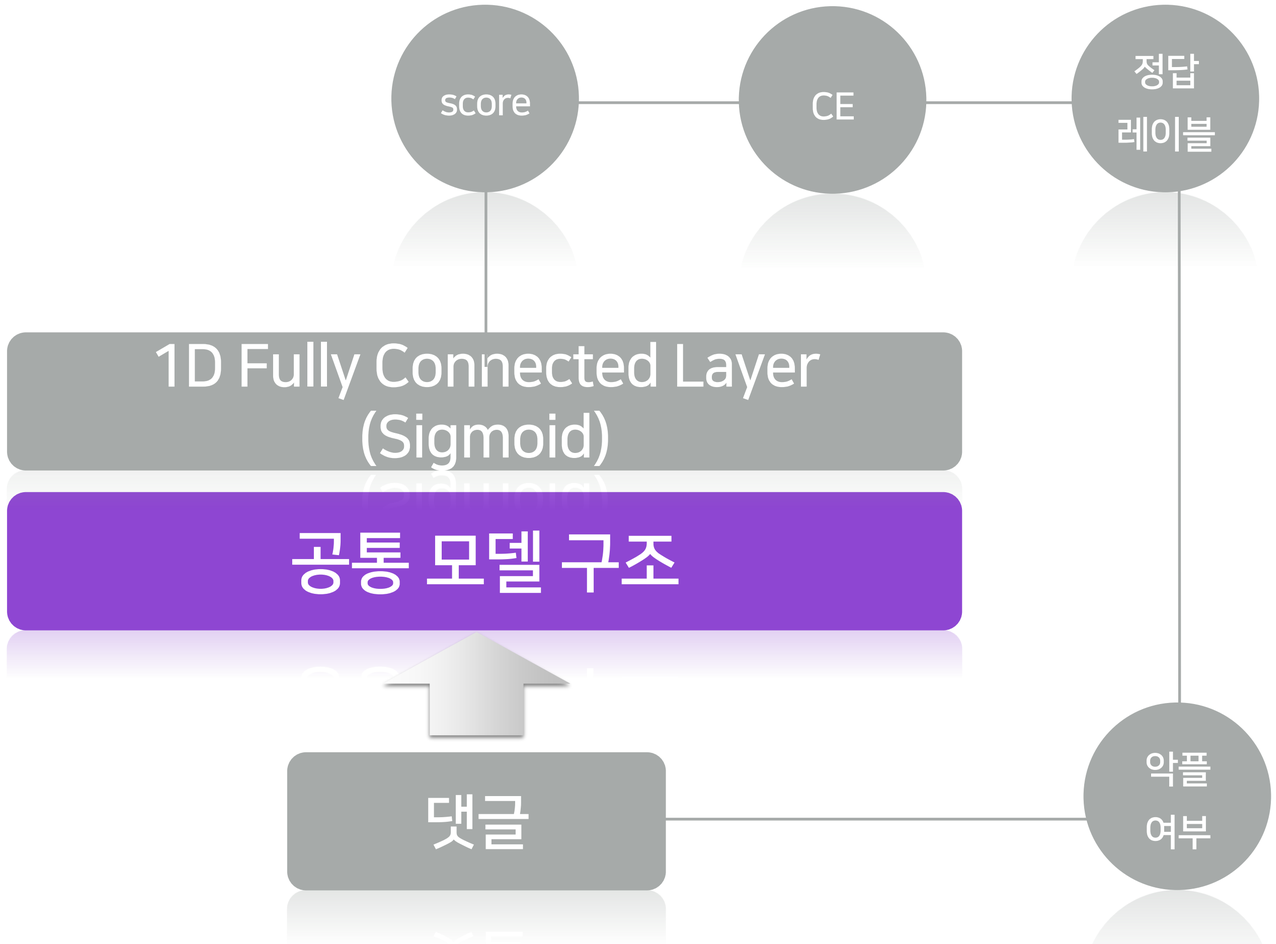
Transfer Learning Stage



3.2 Persona Embedding

DEVIEW 2020

Fine Tuning Stage



3.2 Persona Embedding

소규모 데이터 셋에서 Persona Embedding에 의한 분류 정확도 향상 효과를 검증해본 결과 Persona Embedding을 적용하는 것이 도움이 되며, 투입된 데이터 규모가 클수록 성능향상폭이 큰 것으로 나타났습니다.

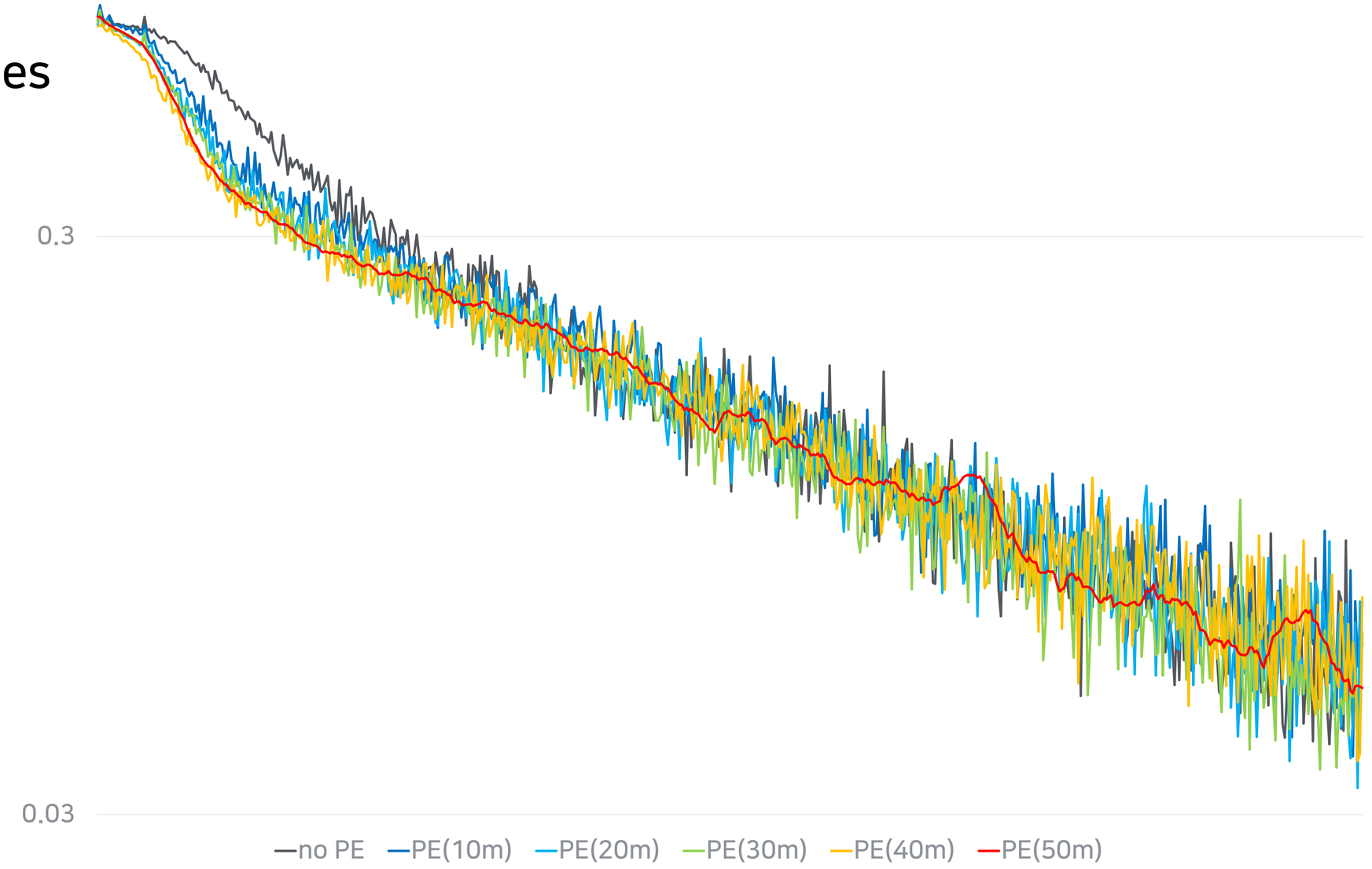
Model	Dev Loss (avg.)	Test Loss (avg.)	Test Accuracy (avg.)	Test F1 Score (avg.)
no P.E.	0.26253	0.279985	0.88538	0.849465
P.E. with 10m triplets	0.243064	0.259716	0.89178	0.857804
P.E. with 20m triplets	0.235027	0.24904	0.89844	0.868904
P.E. with 30m triplets	0.232956	0.246303	0.89808	0.867391
P.E. with 40m triplets	0.227463	0.241328	0.90214	0.873137
P.E. with 50m triplets	0.222996	0.237938	0.90322	0.873347

3.2 Persona Embedding

DEVIEW 2020

Training Losses

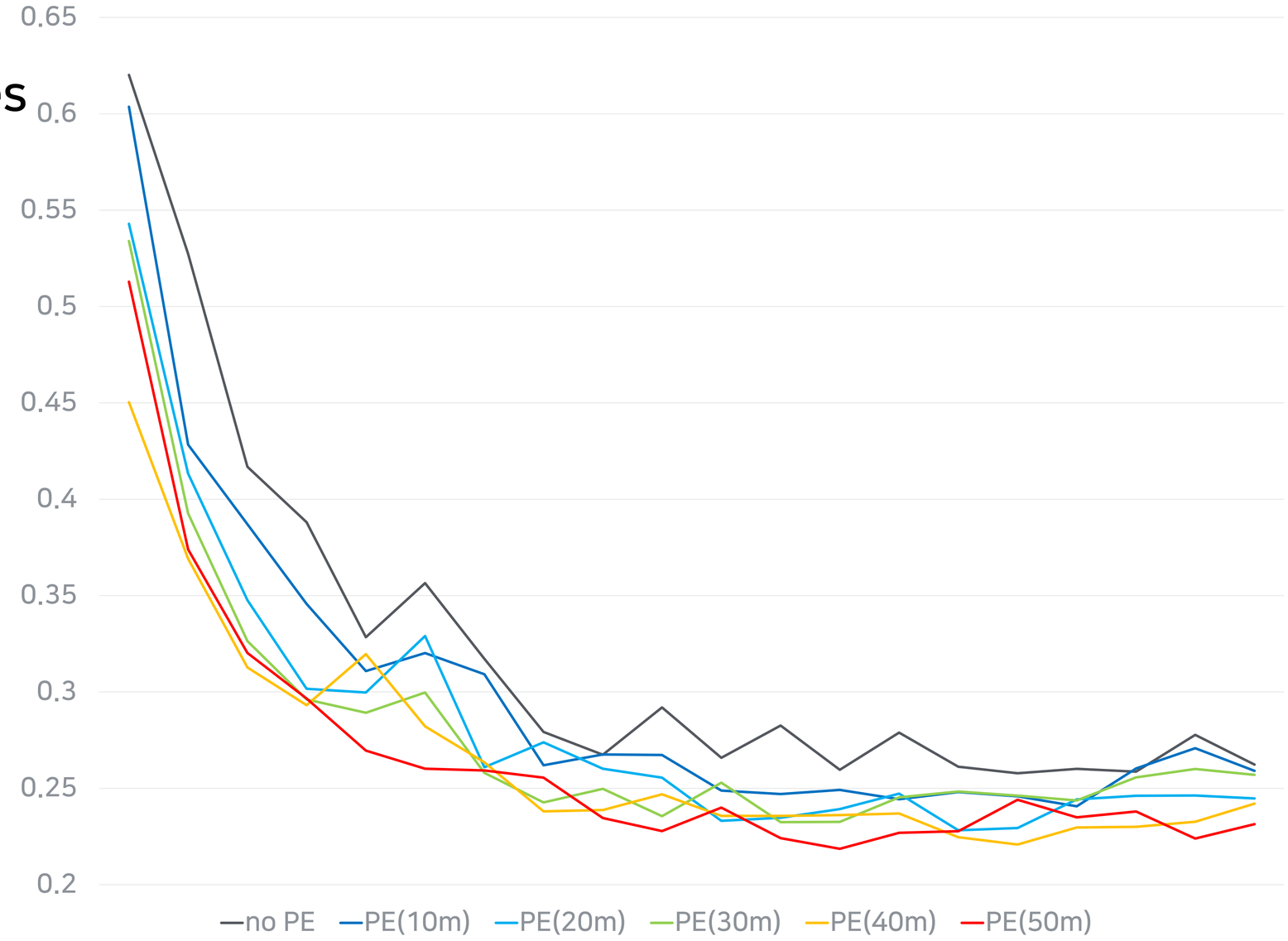
(Log Scale)



3.2 Persona Embedding

DEVIEW 2020

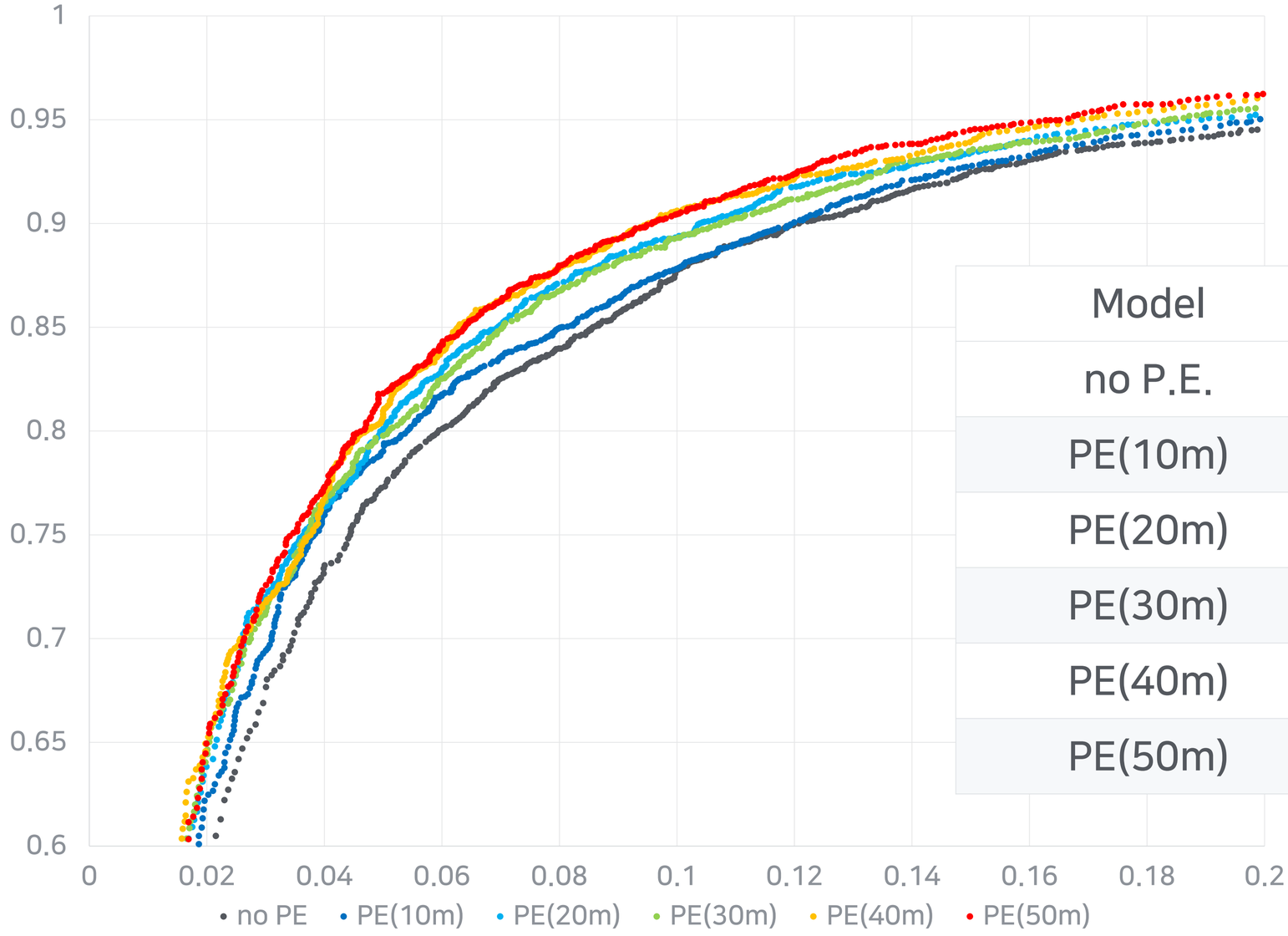
Validation Losses



3.2 Persona Embedding

DEVIEW 2020

Test ROC Curves



Model	AUC
no P.E.	0.9517
PE(10m)	0.9559
PE(20m)	0.9592
PE(30m)	0.9598
PE(40m)	0.9615
PE(50m)	0.9622

3.2 Persona Embedding

최종 배포 모델 선택을 위한 각 학습 조건별 테스트 결과

Persona Embedding과 AutoML의 조합이 가장 좋은 성능을 보여주었습니다.

학습 방법론	Accuracy	Precision	Recall	F1 Score
AutoML	0.9605	0.9337	0.9561	0.9447
ELMO	0.9448	0.8835	0.9719	0.9256
ELMO + AutoML	0.9619	0.9385	0.9547	0.9465
P.E.	0.9644	0.9534	0.9454	0.9494
P.E. + AutoML	0.9648	0.9556	0.9442	0.9499

3.3 Serviceability vs. SOTA

- 범용 장비에서 inference 에 대한
  높은 qps와 낮은 latency 보장(infer 비용 고려 필요)
- 데이터 확장 시 빠른 재학습 및
  AB test에서 적은 diff 보장
- 요구사항 변경 시 빠른 재학습 및
  최대한 목표로 설정한 diff만 발생

- 최고의 metric 상의 성능
- train 과 infer에 필요한 비용 고려하지 않음
- 모델과 기술에 대한 홍보 효과

3.3 Serviceability vs. SOTA

3.3 Serviceability vs. SOTA

4. 출시 준비하기

4.1 정량적 평가의 한계

F1 이라는 precision과 recall을 모두 고려하는 metric이 있음에도, 실제 모델을 운용할 때의 사람의 반응은 F1과 매우 다를 것입니다.

“이 댓글도 잡을 수 있을까?”
“이게 왜 악플이야!”

즉 precision에 더 민감하게 반응 합니다.
알파고 이후로 우리들은 경험에 의해 AI 모델에 대한 낮은 기대치가 보편화 되었음에도, precision에 대해서는 극도로 민감한 반응을 보일 수 있습니다.
때문에 기존 모델링의 사용한 F1-score와 acc 만으로는 서비스 가능 여부를 판단하기 어렵습니다.

4.1 정량적 평가의 한계

동시에 조금 거칠더라도 “표현의 자유”를 최대한 보장하기 위해.
더 다양한 이야기를 할 수 있는 서비스를 제공하고자,
AI가 가리게 될 악플의 기준을 조금 더 상향하여야 했습니다.

4.2 Parameter tuning with AutoML

DEVIEW 2020

모든 모델의 결과물을 사람이 매번 전수검수 하기는 어려우니,

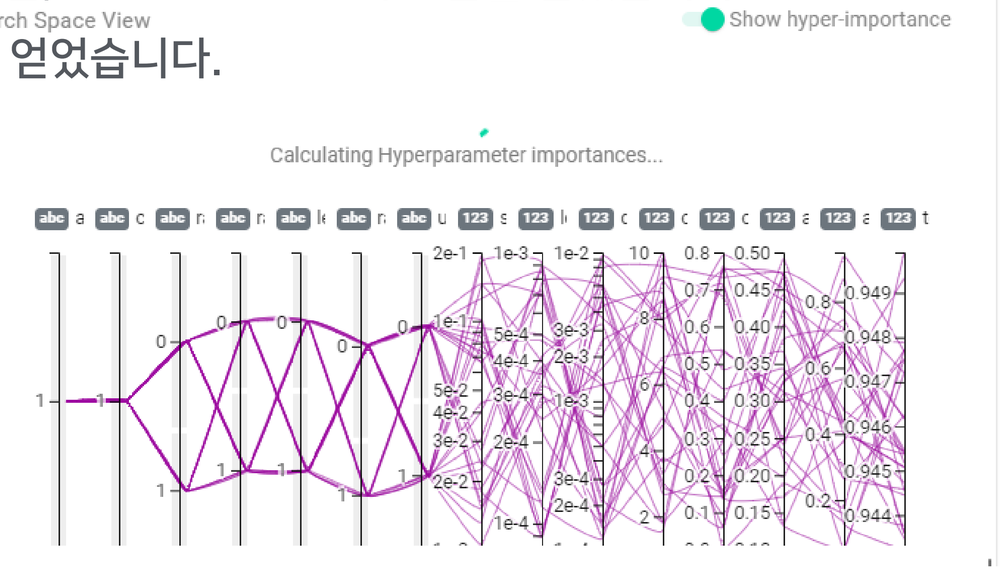
정량적 평가의 한계가 있음에도

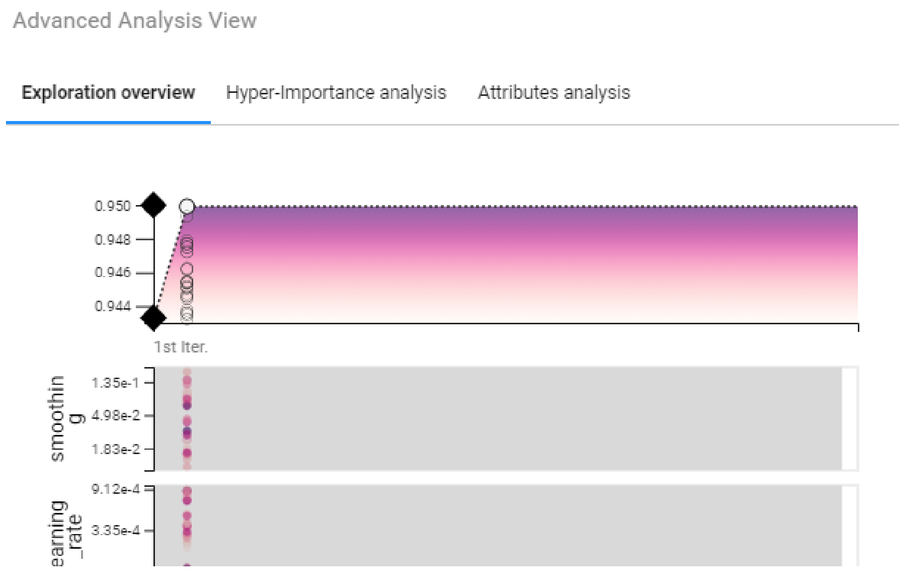
최고의 정량적 성능을 보인 모델들을 비교하여야 할 것입니다.

때문에 동일 모델베이스 동일 데이터에서도 정량적 성능 자체를 올리려는 노력이 필요합니다.

앞선 표에서 확인하셨겠지만, Clova AutoML의 도움을 받았습니다.

F1에서 약 0.44의 이득을 얻었습니다.





4.3 Scalable design for General and Public benefits

솔직히 네이버에서만 클린봇을 운영하려면,
SOTA 모델을 GPU에 얹어서 서비스하면 크게 문제가 없을 것입니다.

하지만

조금 슬픈 이유지만 새로운 방법으로 소통하는 모두에게
공공의 목적으로 필요하다면
누구든 **클린봇을 채용**하여 사용하여
자신의 **서비스 생태계**를
조금 더 **아름답게** 만들기를 바랬습니다.

4.4 Test, test, and test

ML 서비스의 테스트 방법은 기존 서비스의 테스트 방법과는 주된 대상이 조금 다릅니다.

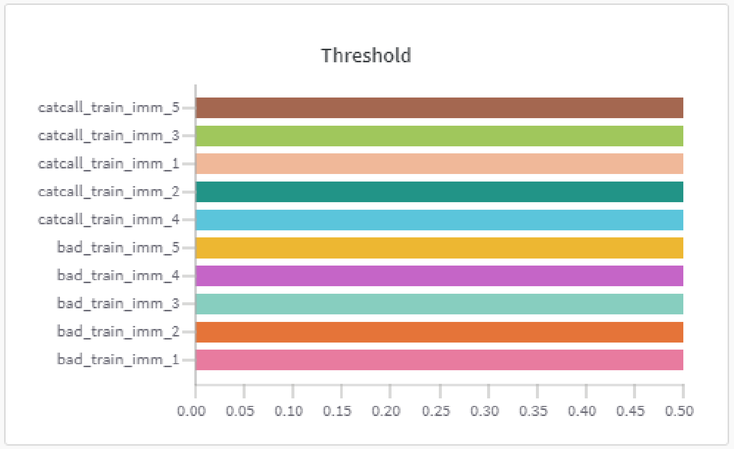
- https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning
- https://martinfowler.com/articles/cd4ml.html

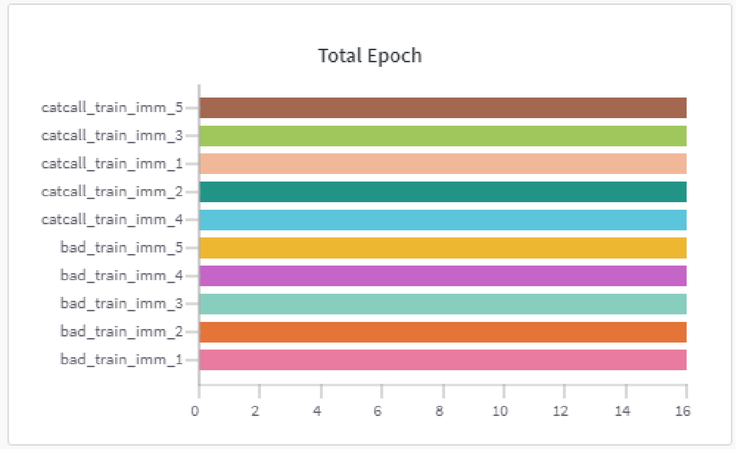
또한 운영중인 서비스의 특성상, 조금만 실수하여도 많은 책임이 따르기에
구글등에서 제안한 위 파이프라인 보다도 혹독한 검증 과정을 필요로 했습니다.

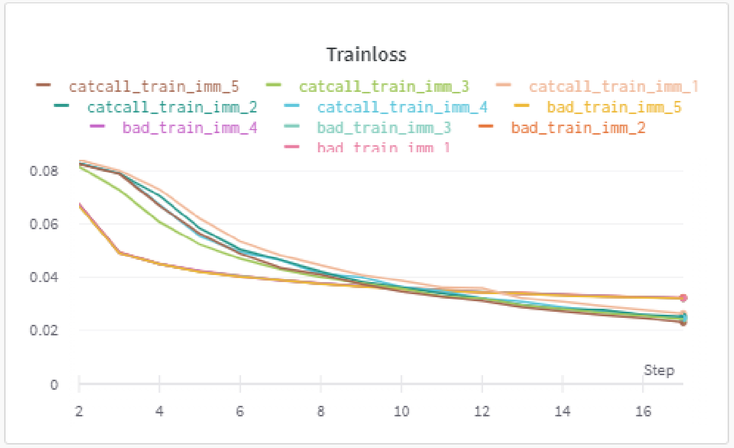
최종적으로 서비스에 투입될 모델을 선정하기 위해
특정 날짜의 로그를 추출하여, 준비된 모델들로 infer 하여 각 댓글 별 모델별로 스코어 기록하고,
스코어를 담당자와 연구자가 확인하여, 정성적으로 어떤 모델이 더 서비스와, 사회적 합의에 부합하는지 논의하였습니다.

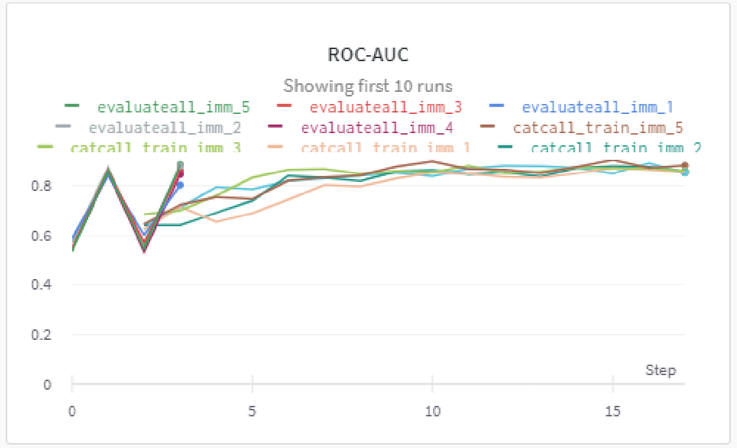
- 10만 건이 넘는 댓글과 악플의 스코어를 3명의 개발자와 다수의 운영자가 직접 눈으로 보고 모델을 검토하였습니다. ㅠㅠ
- (태현님 감사합니다!)

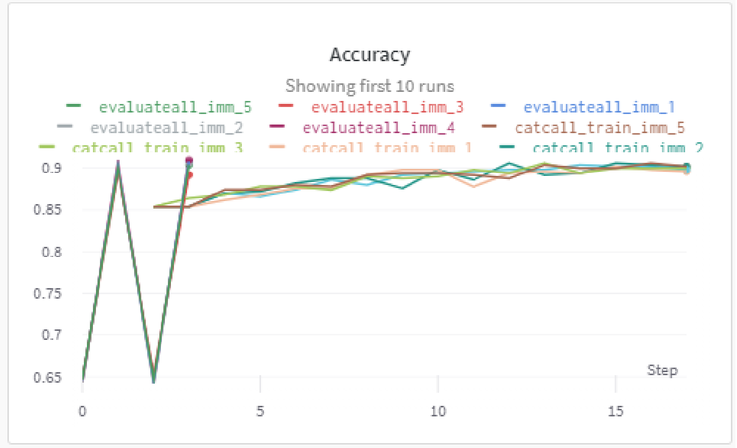
4.4 Test, test, and test











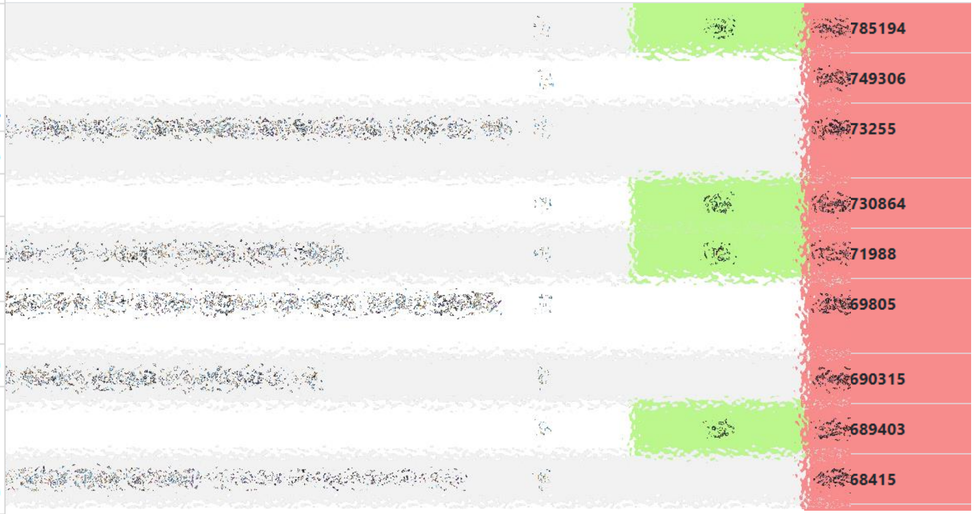
4.4 Test, test, and test

DEVIEW 2020

4.4 Test, test, and test

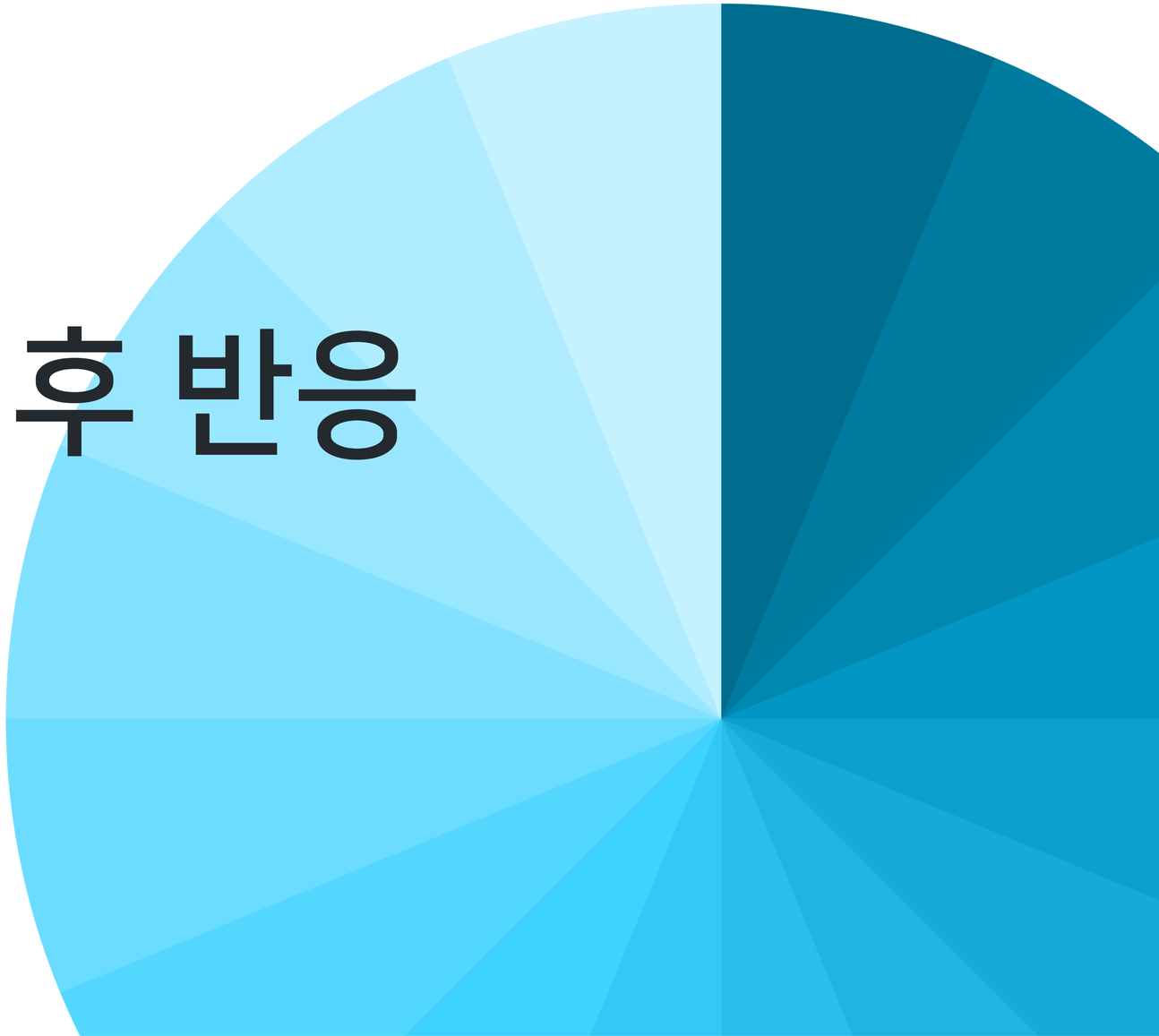
DEVIEW 2020

contents	cb-v2
안될만	7966
제목 뵈	7573
역시 좃	6876
기껏한	6517
참 법	6225
못었죠	0.975
wndu	4997
ㄴ 이햏	5417
기각이	5415
꼭 그떤	
차라리	
이래도	5284
마음만	5228
지랄하	3643
ㅋㅋㅋ	3631
이게 ㄴ	3612
이미 늦	7361
[redacted]	3599
지랄하	
모든 걸	
민들이	3599
지랄이	3598
문제 생	3594



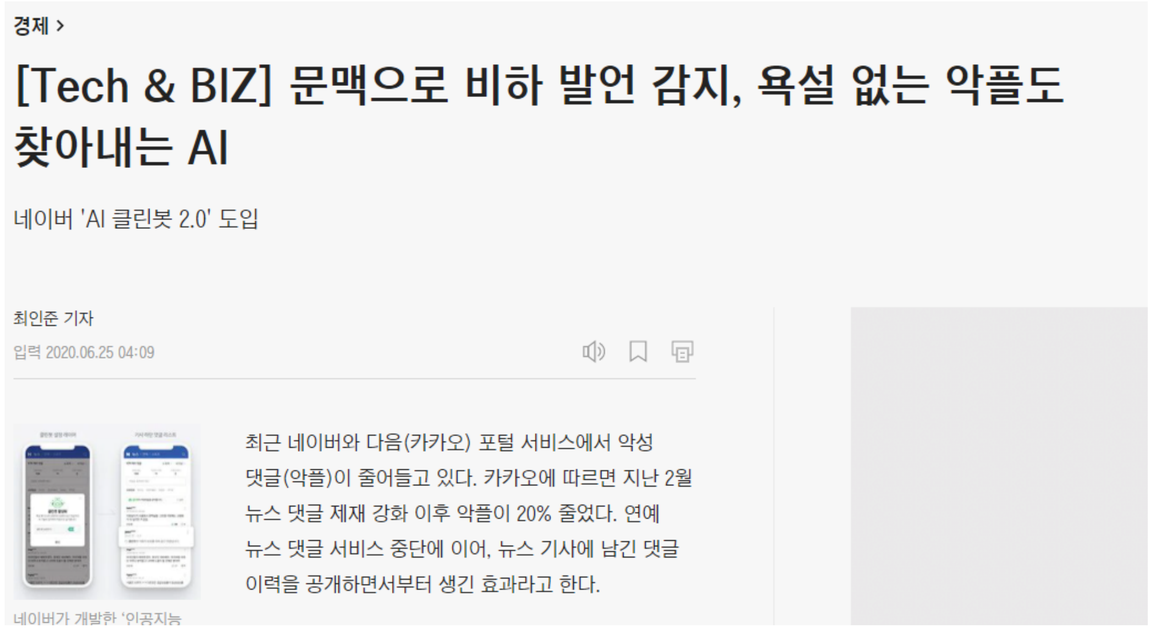
4.4 Test, test, and test

5. 출시 후 반응



5. 출시 후 반응

https://www.chosun.com/site/data/html_dir/2020/06/25/2020062500399.html?utm_source=naver&utm_medium=original&utm_campaign=news



https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=102&oid=023&aid=0003561288

5. 출시 후 반응

https://www.chosun.com/site/data/html_dir/2020/06/25/2020062500399.html?utm_source=naver&utm_medium=original&utm_campaign=news

경제 >

[Tech & BIZ] 문맥으로 비하 발언 감지, 욕설 없는 악플도 찾아내는 AI

네이버 'AI 클린봇 2.0' 도입

최인준 기자

입력 2020.06.25 04:09

최근 네이버와 ㄷ
댓글(악플)이 줄
뉴스 댓글 제재
뉴스 댓글 서비스
이력을 공개하면

네이버가 개발한 '인공지능

朝鮮日報

[아무튼, 주말] 하루에 악플만 수십만 개… 댓글 청소부는 바쁘다

B11면 1단 | 기사입력 2020.09.12. 오전 3:09 최종수정 2020.09.13. 오전 7:36 기사원문 스크랩 본문듣기 • 설정

13 8 요약봇 가

클린봇에 비친 한국 사회

일러스트=안병현

'클린봇이 부적절한 표현을 감지한 댓글입니다.'

https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=102&oid=023&aid=0003561288

5. 출시 후 반응

https://www.chosun.com/site/data/html_dir/2020/06/25/2020062500399.html?utm_source=naver&utm_medium=original&utm_campaign=news

경제 >

[Tech & BIZ] 문맥으로 비하 발언 감지, 욕설 없는 악플도 찾아내는 AI

네이버 'AI 클린봇 2.0' 도입

최인준 기자

입력 2020.06.25 04:09

최근 네이버와 ㄷ
댓글(악플)이 줄
뉴스 댓글 제재
뉴스 댓글 서비스
이력을 공개하면

네이버가 개발한 '인공지능

朝鮮日報

[아무튼, 주말] 하루에 악플만 수십만 개… 댓글 청 쁘다

B11면 1단 | 기사입력 2020.09.12. 오전 3:09 최종수정 2020.09.13. 오전 7:36 기사원문 스크랩 본문

13 8 요약

클린봇에 비친 한국 사회

일러스트=안병현

'클린봇이 부적절한 표현을 감지한 댓글입니다.'

https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=102&oid=023&aid=0003561288

클린봇이 악성댓글을 감지합니다. 설정

blah****
2020.09.12. 03:22
[redacted] 들어오라 하세요
답글 작성 4 2

kise****
2020.09.19. 11:55
악플 신고하면 포인트 좀 주고 그것을 기반으로 빅데이터 구성하면 될것을 아이티 강국 수준 이런 알고리즘 하나도 생성도 못하냐
답글 작성 0 0

hell****
2020.09.13. 08:26
악플은 상대방에게 보이지 않는 상처를 남기는 폭력입니다. 다른 상대방을 폄하하고 욕하는 것은 자신의 의견을 주장하는 것과는 다른 것입니다. 사이버 공간에서도 인격은 존중되어야 합니다. 이제 악플은 그만!
답글 작성 0 0

pinc****
2020.09.12. 10:51
작성자에 의해 삭제된 댓글입니다.

ybwo****
2020.09.13. 01:57
스포츠 기사에 댓글다는게 사라져서 한편으로는 다행이네요. 댓글 보고 응원받고 힘내시는 선수분들도 있는데 100개중 1개라도 악플이면 그걸 보는 선수 분들도 힘들겠죠. 댓글안달고 해서 악플이 줄어든거 같아 좋네요/
답글 작성 1 2

jaus****
2020.09.12. 15:34
클린봇이 부적절한 표현을 감지한 댓글입니다.

meta****
2020.09.12. 03:11
[redacted]는 기사 자체가 악플 아닌가??
답글 작성 1 4

5. 출시 후 반응

https://www.chosun.com/site/data/html_dir/2020/06/25/2020062500399.html?utm_source=naver&utm_medium=original&utm_campaign=news

경제 >

[Tech & BIZ] 문맥으로 비하 발언 감지, 욕설 없는 악플도 찾아내는 AI

네이버 'AI 클린봇 2.0' 도입

최인준 기자

입력 2020.06.25 04:09

네이버가 개발한 '인공지능

최근 네이버와 ㄷ
댓글(악플)이 줄
뉴스 댓글 제재
뉴스 댓글 서비스
이력을 공개하면

朝鮮日報

[아무튼, 주말] 하루에 악플만 수십만 개… 댓글 청 쁘다

B11면 1단 | 기사입력 2020.09.12. 오전 3:09 최종수정 2020.09.13. 오전 7:36 기사원문 스크랩 본문

13 8

요약

클린봇에 비친 한국 사회

일러스트=안병현

'클린봇이 부적절한 표현을 감지한 댓글입니다.'

https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=102&oid=023&aid=0003561288

클린봇이 악성댓글을 감지합니다. 설정

blah****
2020.09.12. 03:22
[redacted] 들어오라 하세요
답글 작성 4 2

kise****
2020.09.19. 11:55
악플 신고하면 포인트 좀 주고 그것을 기반으로 빅데이터 구성하면 될것을 아이티 강국 수준 이런 알고리즘 하나도 생성도 못하냐
답글 작성 0 0

hell****
2020.09.13. 08:26
악플은 상대방에게 보이지 않는 상처를 남기는 폭력입니다. 다른 상대방을 폄하하고 욕하는 것은 자신의 의견을 주장하는 것과는 다른 것입니다. 사이버 공간에서도 인격은 존중되어야 합니다. 이제 악플은 그만!
답글 작성 0 0

pinc****
2020.09.12. 10:51
작성자에 의해 삭제된 댓글입니다.

ybwo****
2020.09.13. 01:57
스포츠 기사에 댓글다는게 사라져서 한편으로는 다행이네요. 댓글 보고 응원받고 힘내시는 선수분들도 있는데 100개중 1개라도 악플이면 그걸 보는 선수 분들도 힘들겠죠. 댓글안달고 해서 악플이 줄어든거 같아 좋네요/
답글 작성

jaus****
2020.09.12. 15:34
클린봇이 부적절한 표현을 감지한

jaus****
2020.09.12. 15:34
[redacted] 경영진 임직원 기자색히들은 부모형제 처가 자식에 손주색히들까지 전부 잡아다 고문하고 처죽여 후쿠시마 원전 앞바다에 수장시킬 색히들
답글 작성 0 1

meta****
2020.09.12. 03:11
[redacted]는 기사 자체가 악플 아닌가??
답글 작성 1 4

5. 출시 후 반응

한국일보

쌍욕 바꾸고 지우고... AI, 한 달 1억건 걸러도 진화하는 악플

16면 TOP | 기사입력 2020.10.08. 오전 4:31 기사원문 스크랩 본문듣기 • 설정

5 11 요약봇 가

> 사람이 직접 댓글 보면서 삭제 불가능
> 댓글 빅데이터 분석해 악플 분류하는 AI 개발
> "해커와 백신의 관계, 건전한 인터넷 문화 확산돼야"

https://www.hankookilbo.com/News/Read/A2020092812570002916?did=NA

5. 출시 후 반응

뉴스 1-10 / 1,466건

한국일보

쌍욕 바꾸고 지우고... AI, 한 달 1억건 걸러도 진화하는 악플

16면 TOP | 기사입력 2020.10.08. 오전 4:31 기사원문 스크랩 본문듣기 • 설정

5 11 요약봇 가

사람이 직접 댓글 보면서 삭제 불가능
댓글 빅데이터 분석해 악플 분류하는 AI 개발
"해커와 백신의 관계, 건전한 인터넷 문화 확산돼야"

https://www.hankookilbo.com/News/Read/A2020092812570002916?did=NA

뉴스검색 가이드

검색결과 자동고침 시작

네이버뉴스

기사에 붙은 댓글을 읽
숨겨주는 인공지능(AI)

순차적으로 도입했다. A
블라인드 처리했다. 올

다시 열 수도...

월 간의 개발 끝에 선보
가했다. 단어가 아닌 문

네이버, AI 클린봇 욕설 댓글까지 탐지
기자협회보 PICK | 2020.07.21. | 네이버뉴스
네이버는 "댓글 이력 공개, 특정인이 작성한 댓글 차단, 클린봇 업그레이드가 악성 댓글 노출을... AI 클린봇을 업그레이드했다. 네이버는 이와 함께 6월의 댓글 수는 연초 대비 0.7% 감소했지만, 작성자 수는 8...
└ 네이버 'AI 클린봇'으로 악플 잡았다 서울경제 | 2020.07.21. | 네이버뉴스
└ 네이버, 댓글 개편 후 악플 절반 이상… 한국경제 | 2020.07.21. | 네이버뉴스
└ 네이버, AI 클린봇 효과 톡톡… 악성… 브릿지경제 | 2020.07.21.
└ 네이버 'AI 클린봇' 댓글 무례한 표현… EBN | 2020.07.21.
관련뉴스 8건 전체보기 >

[대한민국청소년기자단] '악성댓글과의 전쟁' 이제 악플도 AI 알고리즘이 선별...
굿모닝충청 | 3일 전
인스타그램이 사람이 최종 판단하여 정확도를 올리는 데 집중했다면, 네이버는 '클린봇'이라는 악플 제거 AI를 도입하였다. 기존 악플 제거 AI들이 단어를 보고 파악하는 경우가 많았지만, '클린봇'은 문맥을 읽고...

'18어게인' 이기우, 순정남→악역 반전 "욕 먹어도 즐겁다"
뉴스1 | 6시간 전 | 네이버뉴스
이기우는 "일권이 나쁜놈 맞다"며 "지인이 포털사이트 클린봇 활동 열심히 한다던데 앞으로 더 열심히 해야할 것 같다"고 유쾌하게 답했다. 또 "욕을 먹어도 즐겁다. 일권은 학생을 비롯해 친구 등 주변의 모든 인물을...
└ '18어게인' 이기우, 두 얼굴의 반전 … 스포츠조선 | 7시간 전 | 네이버뉴스

5. 출시 후 반응

뉴스 1-10 / 1,466건

한국일보

쌍욕 바꾸고 지우고... AI, 한 달 1억건 걸러도 진화하는 악플

16면 TOP | 기사입력 2020.10.08. 오전 4:31 | 기사원문 | 스크랩 | 본문듣기 • 설정

5 | 11 | 요약봇 | 가

> 사람이 직접 댓글 보면서 삭제 불가능
> 댓글 빅데이터 분석해 악플 분류하는 AI 개발
> "해커와 백신의 관계, 건전한 인터넷 문화 확산돼야"

https://www.hankookilbo.com/News/Read/A2020092812570002916?did=NA

네이버, AI 클린봇 욕설 댓글까지 탐지
기자협회보 PICK | 2020.07.21. | 네이버뉴스
네이버는 "댓글 이력 공개, 특정인이 작성한 댓글 차단, 클린봇 업그
출을... AI 클린봇을 업그레이드했다. 네이버는 이와 함께 6월의 댓글
감소했지만, 작성자 수는 8...
- 네이버 'AI 클린봇'으로 악플 잡았다 서울경제 | 2020.07.21. | 네
- 네이버, 댓글 개편 후 악플 절반 이상… 한국경제 | 2020.07.21.
- 네이버, AI 클린봇 효과 톡톡… 악성… 브릿지경제 | 2020.07.21
- 네이버 'AI 클린봇' 댓글 무례한 표현… EBN | 2020.07.21.

관련뉴스 8건 전체보기 ›

[대한민국청소년기자단] '악성댓글과의 전쟁' 이제 악플도 AI
굿모닝충청 | 3일 전
인스타그램이 사람이 최종 판단하여 정확도를 올리는 데 집중했다면
라는 악플 제거 AI를 도입하였다. 기존 악플 제거 AI들이 단어를 보.
지만, '클린봇'은 문맥을 읽고...

'18어게인' 이기우, 순정남→악역 반전 "욕 먹어도 즐겁다"
뉴스1 | 6시간 전 | 네이버뉴스
이기우는 "일권이 나쁜놈 맞다"며 "지인이 포털사이트 클린봇 활동 열심히 한다던데 앞으로
더 열심히 해야할 것 같다"고 유쾌하게 답했다. 또 "욕을 먹어도 즐겁다. 일권은 학생을 비롯
해 친구 등 주변의 모든 인물을...
- '18어게인' 이기우, 두 얼굴의 반전 … 스포츠조선 | 7시간 전 | 네이버뉴스

클린봇이 악성댓글을 감지합니다. 설정

sain****
2020.10.08. 06:25
욕이 절로 나오는 세상이긴 하다만 그래도 예쁜말 고운말 쓰는 것을 일상화 해 보자구요 ^^
답글 2 | 3 | 0

nowt****
2020.10.08. 06:28
AI 를 패스하는 AI를 개발해서 악플을 다는 세상이 올듯.
답글 작성 | 1 | 0

mkta****
2020.10.08. 05:18
ⓘ 클린봇이 부적절한 표현을 감지한 댓글입니다.

hell****
2020.10.11. 13:47
자신의 의견을 주장하는 것이나 다른 사람의 의견에 반대의 견을 내는 것은 악플과는 다른 차원의 문제입니다. 악플은 다른 사람의 마음에 지울 수 없는 상처를 냅니다. 댓글은 온라인 인격입니다. 악플은 이제 그만!
답글 작성 | 0 | 0

idki****
2020.10.09. 04:57
거시기가 거시기해도 거시기하니 AI 할비가와서 영어는 걸러도 한국어는 못 걸른다는...
답글 작성 | 0 | 0

mypi****
2020.10.08. 14:19
정치적인 의도를 가지고 입력값 지정하여 정부여당 비판 못하게 막은거 모를거 같더냐??
답글 작성 | 0 | 0

lieb****
2020.10.08. 10:47
눈가리고 아웅이지
답글 작성 | 0 | 0

5. 출시 후 반응

뉴스 1-10 / 1,466건

한국일보

쌍욕 바꾸고 지우고... AI, 한 달 1억건 걸러도 진화하는 악플

16면 TOP | 기사입력 2020.10.08. 오전 4:31 | 기사원문 | 스크랩 | 본문듣기 • 설정

5 | 11 | 요약봇 | 가

> 사람이 직접 댓글 보면서 삭제 불가능
> 댓글 빅데이터 분석해 악플 분류하는 AI 개발
> "해커와 백신의 관계, 건전한 인터넷 문화 확산돼야"

https://www.hankookilbo.com/News/Read/A2020092812570002916?did=NA

네이버, AI 클린봇 욕설 댓글까지 탐지
기자협회보 PICK | 2020.07.21. | 네이버뉴스
네이버는 "댓글 이력 공개, 특정인이 작성한 댓글 차단, 클린봇 업그
출을... AI 클린봇을 업그레이드했다. 네이버는 이와 함께 6월의 댓글
감소했지만, 작성자 수는 8...
- 네이버 'AI 클린봇'으로 악플 잡았다 서울경제 | 2020.07.21. | 네
- 네이버, 댓글 개편 후 악플 절반 이상… 한국경제 | 2020.07.21.
- 네이버, AI 클린봇 효과 톡톡… 악성… 브릿지경제 | 2020.07.21
- 네이버 'AI 클린봇' 댓글 무례한 표현… EBN | 2020.07.21.

관련뉴스 8건 전체보기 >

[대한민국청소년기자단] '악성댓글과의 전쟁' 이제 악플도 AI 클
굿모닝충청 | 3일 전
인스타그램이 사람이 최종 판단하여 정확도를 올리는 데 집중했다면
라는 악플 제거 AI를 도입하였다. 기존 악플 제거 AI들이 단어를 보.
지만, '클린봇'은 문맥을 읽고...

'18어게인' 이기우, 순정남→악역 반전 "욕 먹어도 즐겁다"
뉴스1 | 6시간 전 | 네이버뉴스
이기우는 "일권이 나쁜놈 맞다"며 "지인이 포털사이트 클린봇 활동 열심히 한다던데 앞으로
더 열심히 해야할 것 같다"고 유쾌하게 답했다. 또 "욕을 먹어도 즐겁다. 일권은 학생을 비롯
해 친구 등 주변의 모든 인물을...
- '18어게인' 이기우, 두 얼굴의 반전 … 스포츠조선 | 7시간 전 | 네이버뉴스

클린봇이 악성댓글을 감지합니다. 설정

sain****
2020.10.08. 06:25
욕이 절로 나오는 세상이긴 하다만 그래도 예쁜말 고운말 쓰는 것을 일상화 해 보자구요 ^^
답글 2 | 3 | 0

nowt****
2020.10.08. 06:28
AI 를 패스하는 AI를 개발해서 악플을 다는 세상이 올듯.
답글 작성 | 1 | 0

mkta****
2020.10.08. 05:18
클린봇이 부적절한 표현을 감지한 댓글입[redacted]개44ㅅ @ ㅐ ㄲ ㅣ!
답글 작성

mkta****
2020.10.08. 05:18

hell****
2020.10.11. 13:47
자신의 의견을 주장하는 것이나 다른 사람의 의견에 반대의 견을 내는 것은 악플과는 다른 차원의 문제입니다. 악플은 다른 사람의 마음에 지울 수 없는 상처를 냅니다. 댓글은 온라인 인격입니다. 악플은 이제 그만!
답글 작성 | 0 | 0

idki****
2020.10.09. 04:57
거시기가 거시기해도 거시기하니 AI 할비가와서 영어는 걸러도 한국어는 못 걸른다는...
답글 작성 | 0 | 0

mypi****
2020.10.08. 14:19
정치적인 의도를 가지고 입력값 지정하여 정부여당 비판 못하게 막은거 모를거 같더냐??
답글 작성 | 0 | 0

lieb****
2020.10.08. 10:47
눈가리고 아웅이지
답글 작성 | 0 | 0

6. Media Tech는 오늘

6.1 Cleanbot Red and Orange ver.

DEVIEW 2020

- Red ver.

  저 감수성 모델로,

  댓글 그 자체만으로 **명백히** 성희롱 또는 음란한 의도로

  인식할 수 있는 댓글을 분류합니다.

- Orange ver.

  고 감수성 모델로

  본문에 따라서, 혹은 개인에 따라서 **성희롱 또는 음란한** 의도로

  작성된 댓글이라고 **인식 될 수도** 있는 댓글을 분류합니다.

댓글	Orange	Red	Clean V2
요염하게 생겨서[redacted]싶어요 ㅋㅋ	0.999	1.000	0.050
[redacted]안하면 안돼냐	0.994	0.999	0.079
[redacted]물?	1.000	1.000	0.037
골프는 가슴으로 치는거다	0.995	0.589	0.041
세계에서 가장 섹시한 골퍼	0.965	0.186	0.054
미드가 너무 좋은데	0.670	0.013	0.039
ㅓㅜㅑ 오늘은 이거다	0.997	0.008	0.034
언니 일주일에 [redacted]?	0.911	0.268	0.273

6.2 Cleanbot Phobia and Aggression ver.

악플 세분화

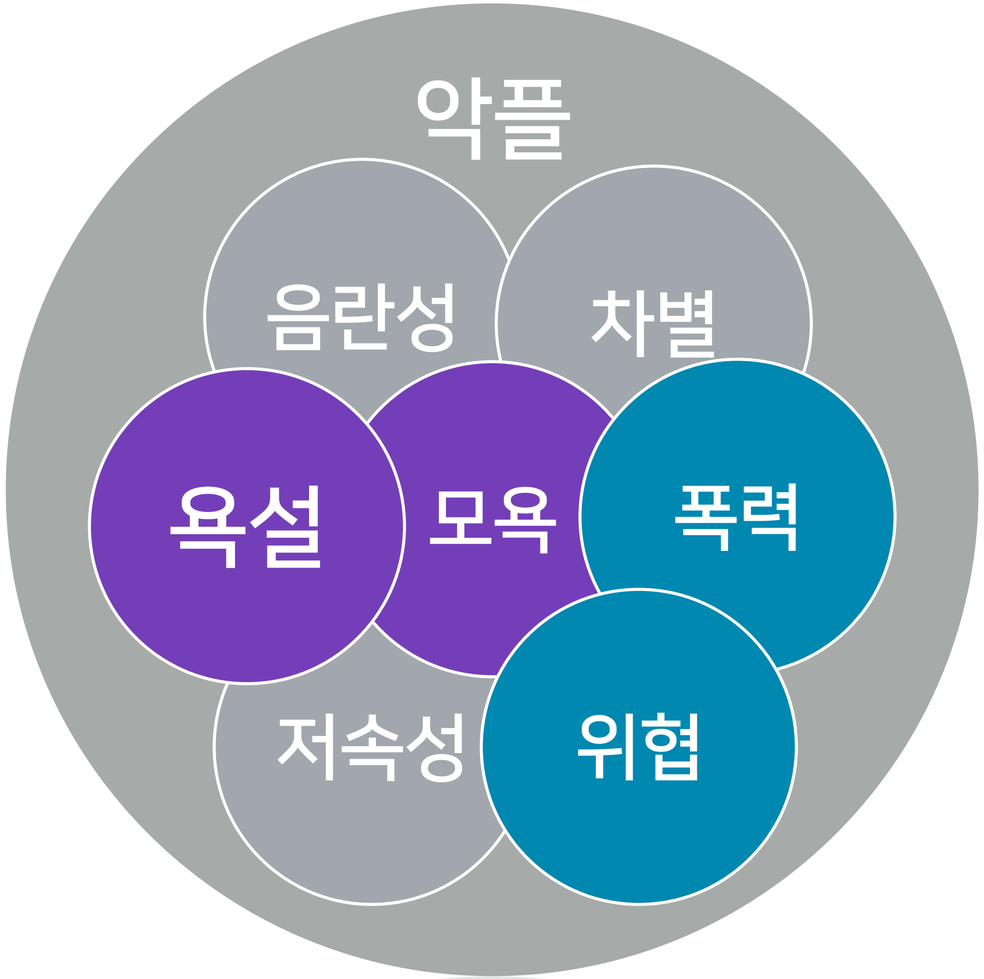
- Phobia

  혐오(비하) 표현을 찾아내는 모델입니다. 단순 욕설이 아니라 대상에 대한 비하 또는 혐오 표현을 찾아내는 모델입니다.

- Aggression

  공격성에는 준위가 있다고 판단하였으나 정의한 각 class 간의 수치적 정도가 linear 하지 않다고 판단하여, multi class classification 문제로 만들어 풀었습니다.

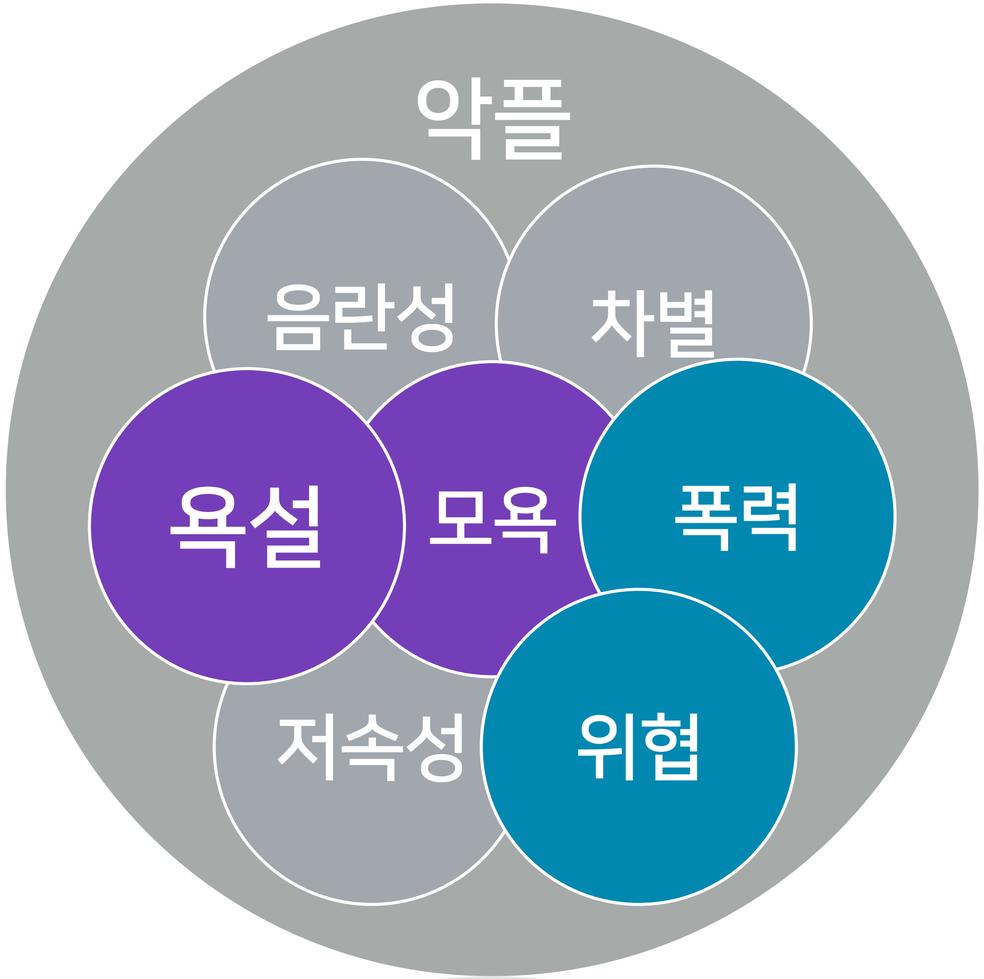
  작성자의 정신적, 물리적 가해의도 여부가 있는 경우와, 의도가 없으나 피격이 발생할 여지가 있는 경우를 판별합니다.



6.2 Cleanbot Phobia and Aggression

DEVIEW 2020

- Phobia or Aggression
- 이색기 시즌초반에 헛스윙 안한다고 물고빠는 기사 개 많던데 3할초반ㅋㅋㅋㅋㅋㅋㅋ
- 아직도 전범풀을 빠네 키~야 개.돼지들 널렸다
- 손해보는 게 호구
- 핫코너 경쟁같은소리하네 딱봐도 민수 낙점인데특히 빳다돌릴때 과감하게 돌리는게 인상적이다 펀치력도 잇고
- 광안대교 앞 마린시티만 비싸고 나머진 평당 1000도 안하잖어 ㅋㅋㅋㅋㅋㅋㅋㅋ엌 그나마 엘시티 하나 들어와서 ㅈㅇ라도 해보지만 현실은 부실공사에 건축법 위반 ㅋㅋ루삥뽕
- 밖에 모여서 파티하는거 지적하는 사람들이 아무도 없네 코로나가 종식된것도 아니고 마스크 안쓴사람들도 있던데 집단감염되겠네
- 시러ㅜ븅



6.3 Open API 준비

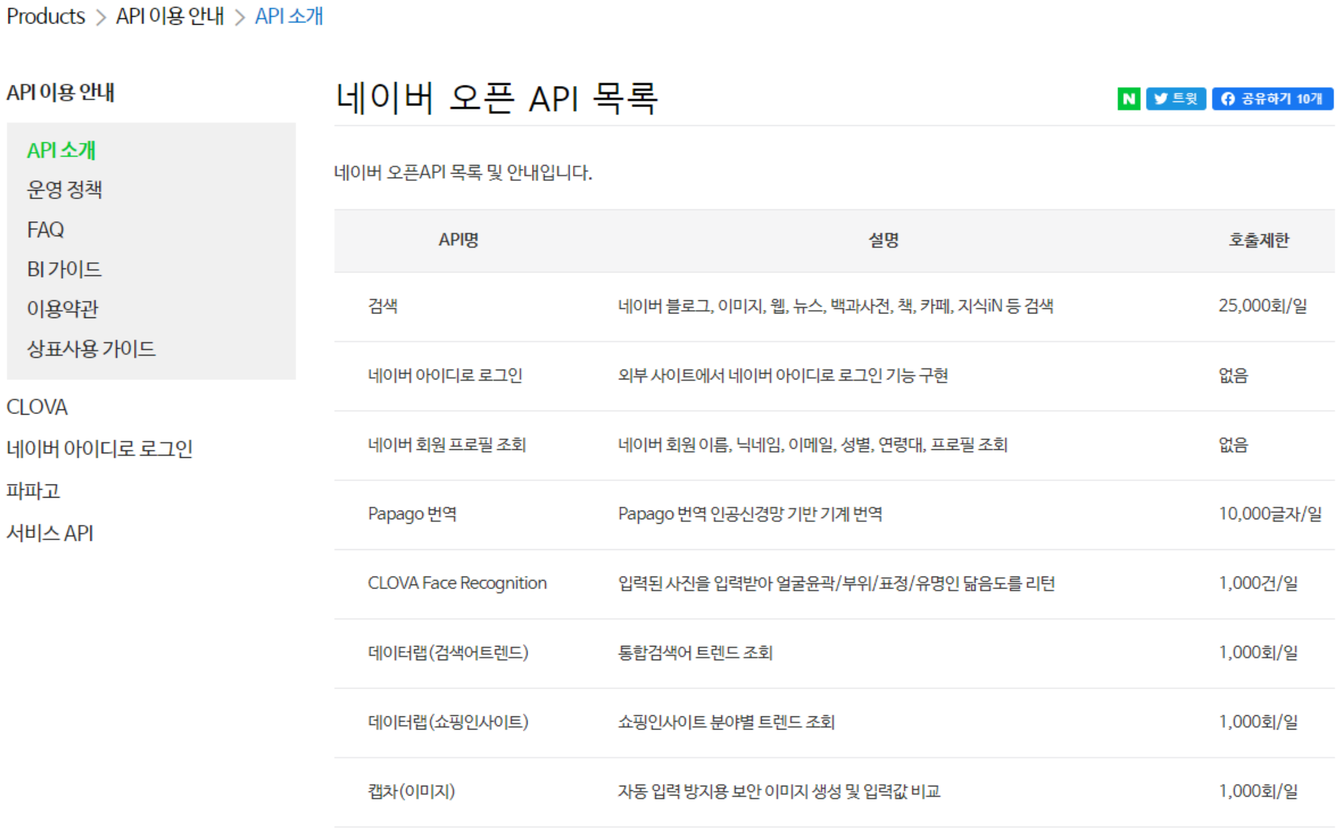
DEVIEW 2020

클린봇을 네이버 개발자 센터에 open api (https://developers.naver.com/products/intro/plan/)로 공개할 준비를 하고 있습니다.

앞서 밝힌 것처럼 저비용 모델에 강한 집착을 했던 이유는 여기에 있습니다.

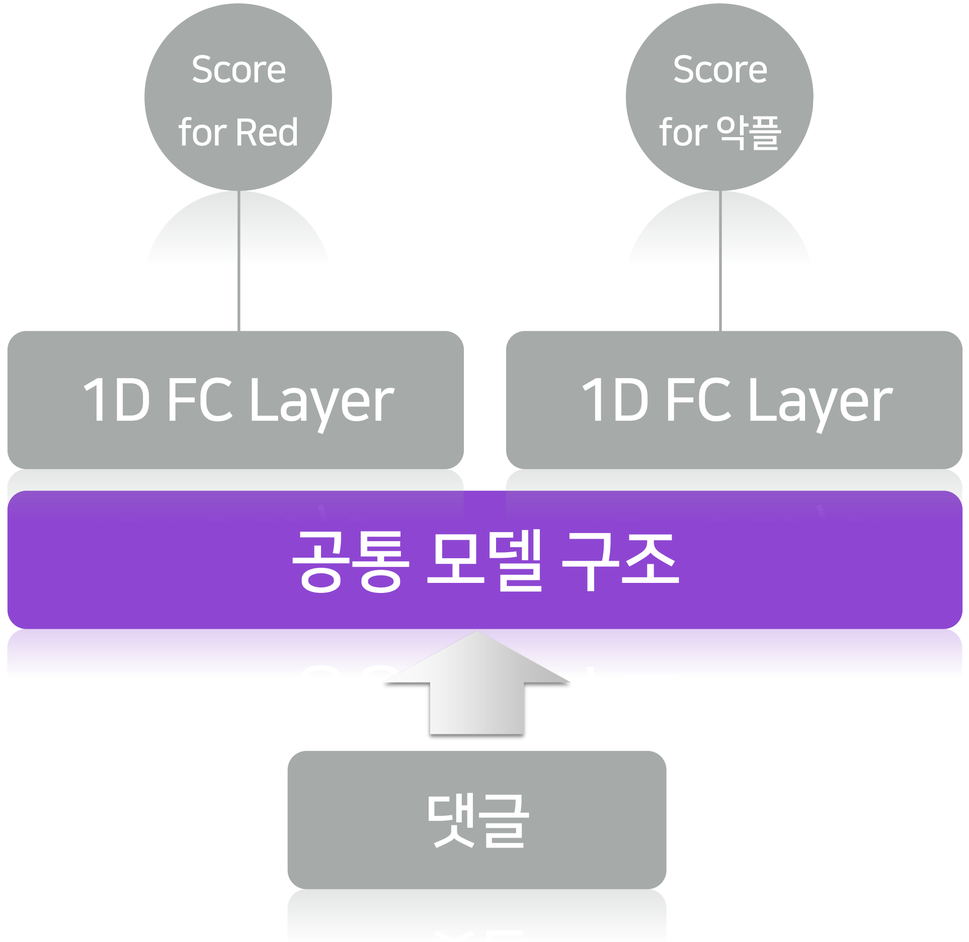
모델의 **운용 비용과 정확도 성능** 간의 적절한 합의점을 찾고, inference serve를 최적화 하여, open api로 공개하였을 때 이용자와, 네이버가 부담해야 하는 비용을 최소화 하고자 하였습니다.

현재 더 **많은 호출을 안정적으로 감당** 할 수 있도록 inference 모듈을 개선하고 있습니다.



7. ETC

7 Multi-task Learning



7 Multi-task Learning

red – 악플 multi-task learning

전처리	multi task	model	f1	pr	re
O	O	cnn_rnn	71	95	57
O	X	cnn_rnn	73	85	65
X	O	cnn_rnn	77	86	71
X	X	cnn_rnn	82	82	82
O	X	persona	67	90	53
X	X	persona	81	87	76

Q & A

Thank You