

혐오 표현 교육용 챗봇

경희대학교 응용수학과
권회준

목차

table of contents

- 1 자기 소개
- 2 데이터 소개
- 3 챗봇 소개
- 4 결론

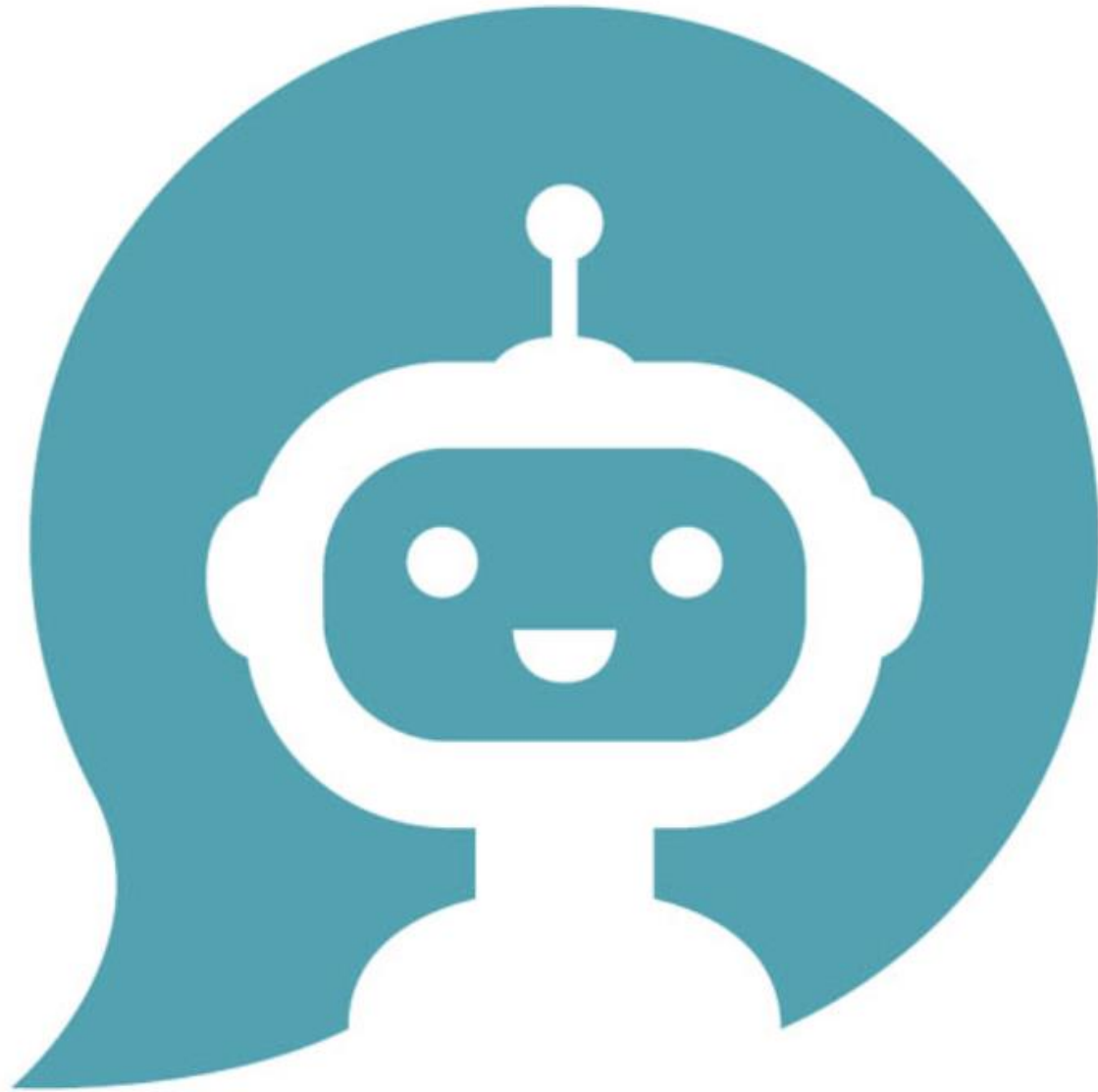
1

자기소개









2

데이터 소개

2018 – 2020 사이에 수집된 한국어 뉴스 기사 댓글

각 데이터는 'text'와 'label'로 구성되어 있음

'text'는 댓글

'label'은 어떤 (정치/연령..) 혐오 표현인지 / 혐오 표현이 아닌지

```
{ 'text' : '3040 대깨문 핵심지지 아닌가 지네들이 찍은사람이 퍼는정책이니 감사히 받겠지', 'label' : [2, 4] }
```

데이터 소개

	label	document	preprocessed_text
1	2,4	"자한당틀딱들.. 악플질 고만해라."	" 자 한 당 틀 딱 들 .. 악플 질 고만해라 ."
2	8	정치적으로 편향된 평론한은 분은 별로...	정치 적 으로 편향 된 평론 한 은 분 은 별로 ...
3	2	적당히좀 쳐먹지.그랬냐??? 안그래도 문재인 때문에 나라 엉망진창인데...	적당히 좀 쳐 먹지 . 그랬냐 ??? 안 그래도 문재인 때문 에 나라 엉망 진창 인...
4	4	"안서는 아재들 풀발기 ㅋㅋㅋ"	" 안 서는 아 재 들 풀 발기 ㅋㅋㅋ "
5	8	우와 ㅋㅋ 능력자	우와 ㅋㅋ 능력자

	label	document	preprocessed_text
1	8	서로만이 이렇게 있을때 보기 좋다♥	서로 만이 이렇게 있을 때 보기 좋다 ♥
2	8	승호 키스신에 누나는 읊니다..흑흑..너무 찼하자너..무방비상태였는데 충격먹음..손...	승호 키스신 에 누나 는 읊니 다 .. 흑 흑 .. 너무 찼하자너 .. 무방비 상태...
3	1,3	"미친돼지새끼 육갑떠네 ——"	" 미친 돼지 새끼 육갑 떠네 —— "
4	3	"전광훈이 교수 됐네 ㅋㅋㅋ 애라이 새까 니가 목사면 박쥐가 독수리다 임마"	" 전광훈 이 교수 됐네 ㅋㅋㅋ 애 라이 새까 니 가 목사 면 박쥐 가 독수리 다 ...
5	8	인덕션 꼭 누를때 기절하는줄ㅋㅋㅋ 아재개그통한다	인덕션 꼭 누를 때 기절 하는줄 ㅋㅋㅋ 아 재개 그 통한 다

데이터 소개

트레이닝 샘플 갯수: 78977

검증 샘플 갯수: 8776

테스트 샘플 갯수: 21939

레이블: 80

레이블 분포:

label

8 42909

3 6359

2 4967

0 4083

1 3930

...

1,3,7 1

0,2,3,7 1

1,2,3,5 1

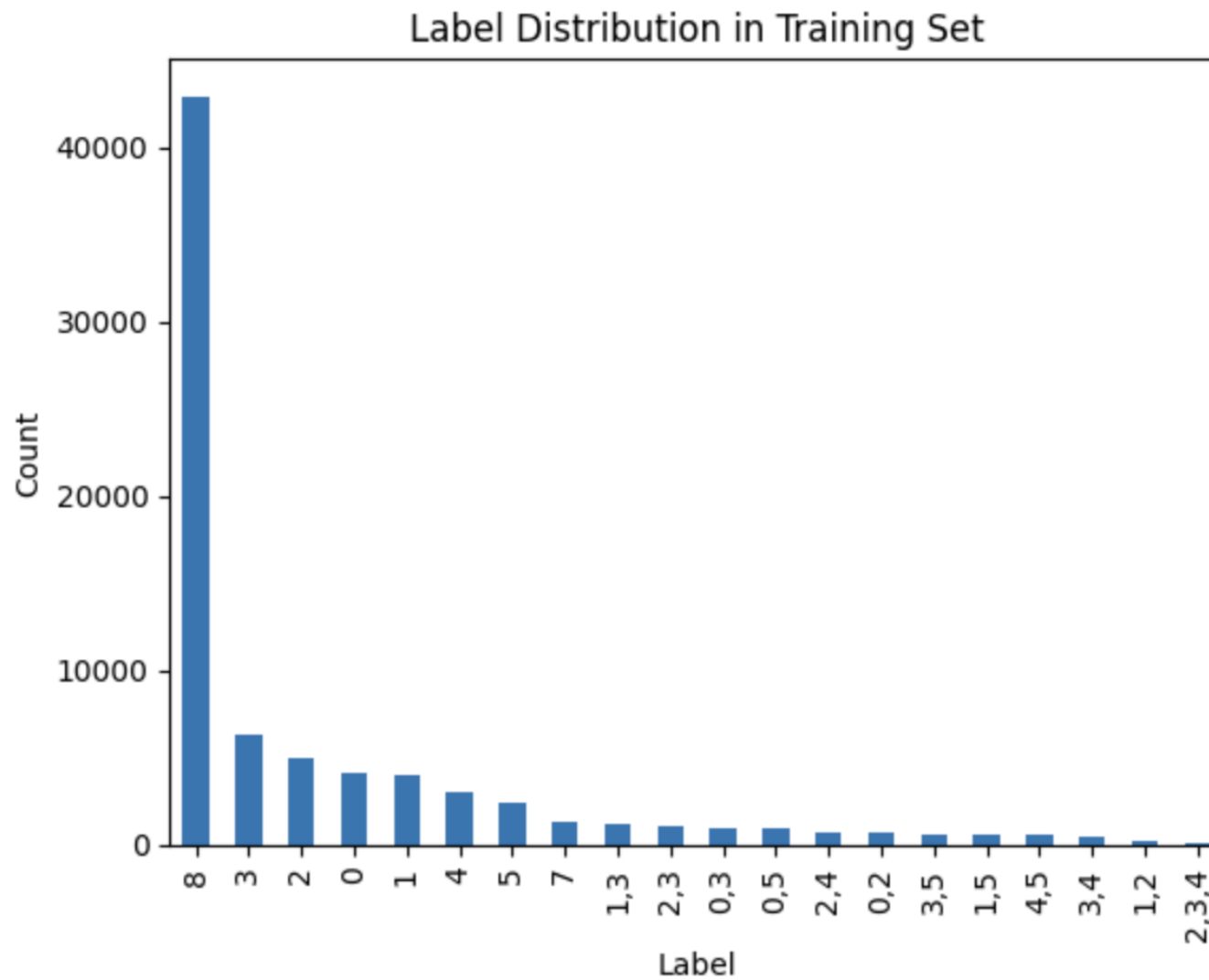
0,2,4,5 1

0,1,6 1

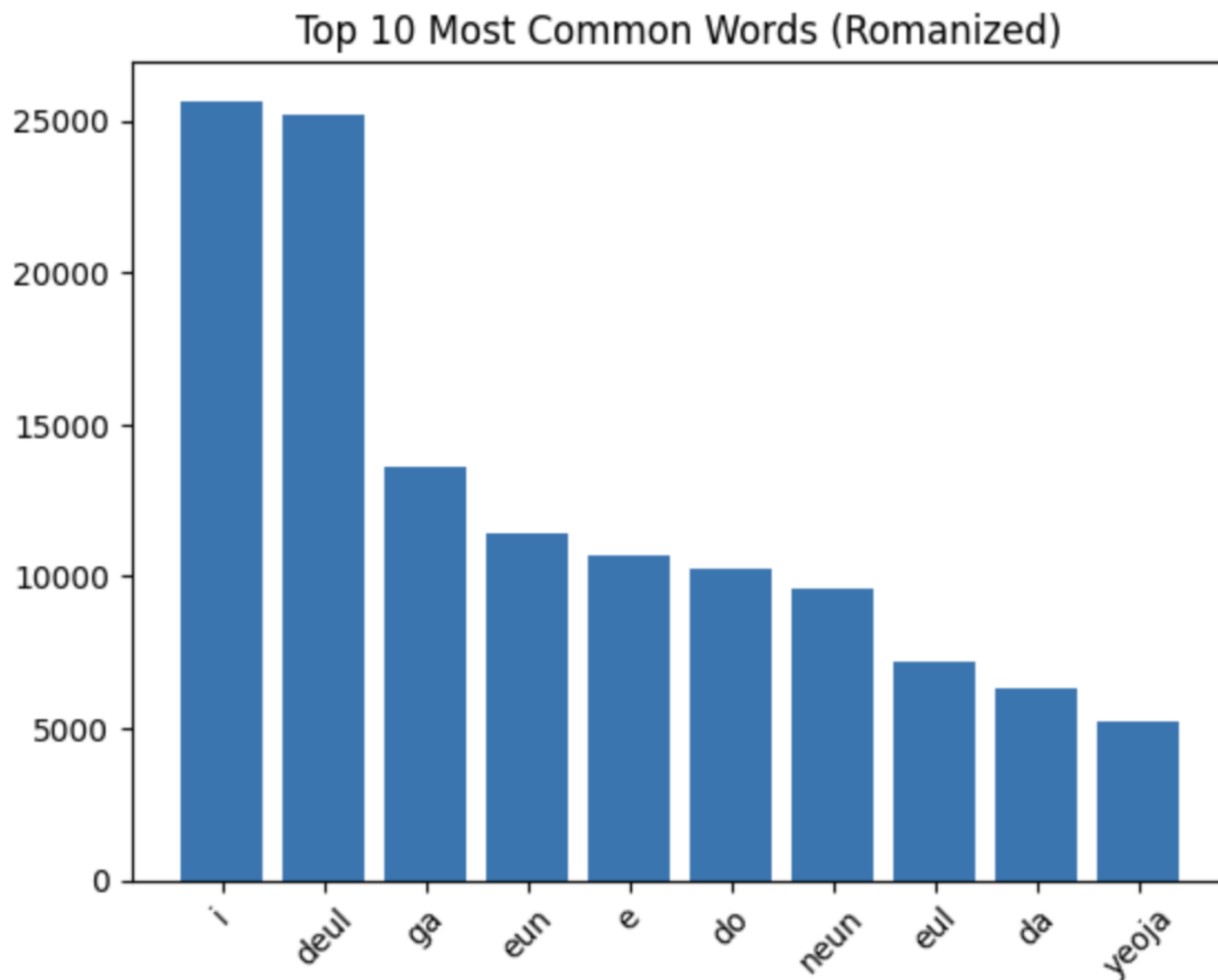
Name: count, Length: 80, dtype: int64

평균 문장 길이: 14.420109652177215

데이터 소개



데이터 소개



데이터 소개

	Word	Frequency	
180	여자	5228	←
3	으로	3585	
213	진짜	3355	
67	남자	3345	←
49	사람	3240	
63	하는	2930	
225	에서	2889	
570	얼굴	2383	←
263	이다	2084	
1331	너무	1960	
158	새끼	1945	
132	돼지	1935	←
66	한국	1864	
551	한테	1799	
182	그냥	1736	
118	댓글	1719	
16	인데	1676	
508	이나	1635	
130	같은	1488	
13	나라	1476	
1299	생각	1465	
1016	나이	1447	
101	중국	1436	
419	재앙	1428	←
362	이런	1408	
11	문재인	1398	←
194	까지	1315	
117	기사	1279	
536	전라도	1274	←

혐오 욕설/정치성향 차별 레이블들이 많이 분포함

+

재앙/문재인 높은 분포로 나옴



문정권에 대한 비판적인 여론이 많았다

여자/남자는 어디 레이블에 있는 것인가?

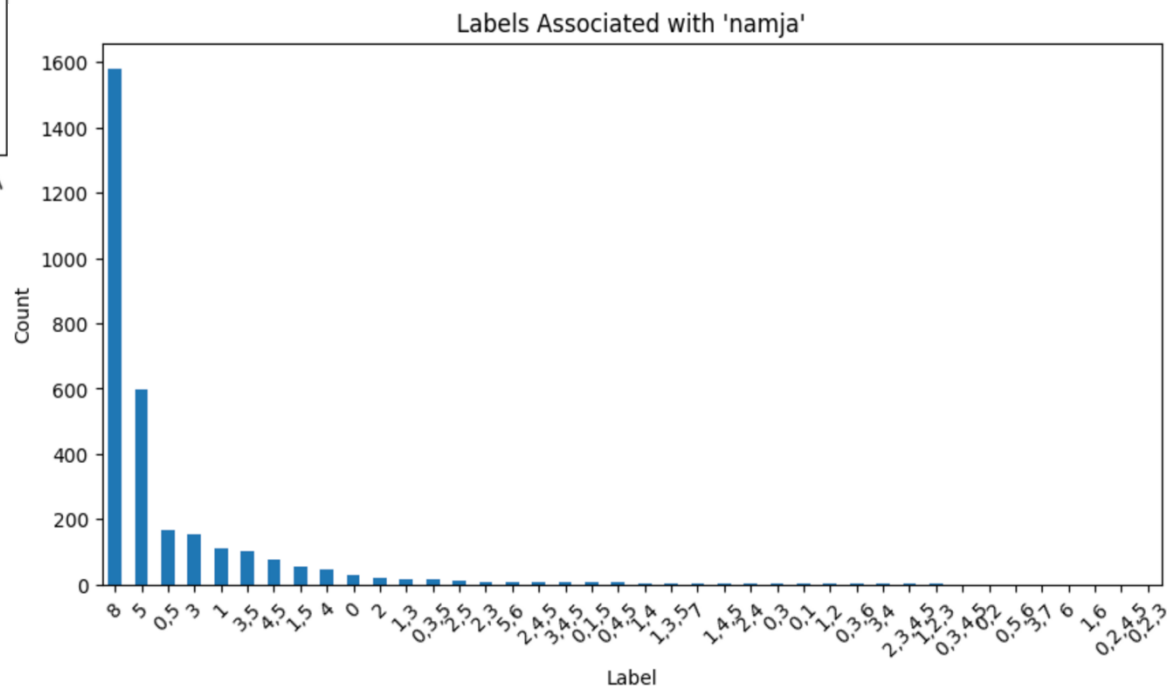


성차별 레이블속에 많이 존재하나?



	Word	Frequency
31	여자	1219
44	남자	700
127	..	291
253	ㅋㅋ	210
91	...	201
92	댓글	151
102	페미	135
392	ㅋㅋㅋ	131
220	하는	127
88	으로	122
158	맘충	120
...

나머지 4천개는 어디에..?



MISLABELING

데이터 소개

11	공통점 은 한국 남자 ?
36	근데 더러운건 지 디 인데 왜 여자 욕 만 함 ㅋㅋ
61	어린 여자 만나야 하는 건가 ..
64	남자 도 성형 이 많다니 .. 성형 공화국 답 다
92	화장 안 한 얼굴 이 훨씬 이쁜것 같음 . 실제 로 보니 몸매 도 날씬한것 같 드 ...
100	송지효 남친 있지 알 ㅎ 음 ? 그 말죽거리 남자 매니저 옛 던애
121	여자 얼굴값 하게 생긴네 .. 색 기도 많아 보이 고 ... 남자 가 불쌍타 .. ㅈㅈ
139	남자 만 불리해
142	남자 가 봐도 ㄹㅇ 잘생겼는데 수염 은 깎자 ... 리틀 소지섭 별명 이 있지만 수...
176	자식 을 낳았으면 어떨 게 든 책임 을 져야지 .. 자식 입장 에선 똑같은 자식 인...
194	박서준 은 좋은데 윤식당 이 윤식당 으로써 빛 이나고 힐링 이 되어야 하는데 박서준...
213	여자 가봐도 반하겠는데 ㅋ 이동욱 이랑 어울린다 웬지 결혼 까지도 갈수있다는 느낌 이 뵈
217	이야 ~ 역시 남자 는 튼 튼 ! 이다 !
237	여자 는 처음 부터 끝 까지 A 로 해주고 남자 는 바로 성 까고 이름 도 까네 ㅋㅋㅋ
252	성비 봐라 ㅋㅋㅋ 여자 들 다 욕 하네
282	시후 . 조심해라 . 여자 들 아
291	역시 여자 는 어려야
312	남자 들 은 다 똑같은 .. 진리



레이블 8을 자세히 살펴보았다



데이터 소개

646 " 와 ... 저 얼굴 이 일본 에션 아이돌 이라 하는구나 ㅋㅋㅋ "



914 대통령 자녀 들 도 얼굴 공개 되는 세상 인데 · 엄마 는 죽어라고 사진 올리는 게...

924 " 엉덩이 는 튼실한데 ㅋㅋ 근데 얼굴 은 왜 ? 가린 거 ??? 뉘집 도련님 인지...

984 유부녀 얼굴 이 궁금하다 . 이쁘겠지 ?

1094 여태 까지 중 에 가장 놀란 듯 한 얼굴 이었음 . 나 모현 하고 싶은대로 하라 고...

1100 조현재 교수 랑 홍경민 석은 얼굴 같어 ~ 잘생겼다

1329 얼굴 은 ? 90년 대같으면 티비 에 나오지도 못 할 얼굴



분명 레이블링이 잘 못된 댓글들이 존재한다



모델로 레이블링을 다시 해보면 어떨까?

데이터 소개

시사점

단어들의 분포를 살펴보면서

2018–2020 시대상황을 반영하고 있다

문제점

레이블의 분포를 살펴보면서

레이블링이 잘 못된 댓글들이 존재한다

데이터 소개

혐오 표현 아님으로 레이블링된 문장들을 학습된 모델로 분류 시도



혐오 표현으로 예측된 문장 1: 공통점은 한국남자?
혐오 표현으로 예측된 문장 2: 50어 애나서 언제 길러요?
혐오 표현으로 예측된 문장 3: "이여인 관상자체가.."
혐오 표현으로 예측된 문장 4: 남자도 성형이 많다니..성형공화국답다
혐오 표현으로 예측된 문장 5: 누구신지. . .존재감 없는
혐오 표현으로 예측된 문장 6: "섬나라로 가라..."
혐오 표현으로 예측된 문장 7: "느그나라 북한? 전라도?"
혐오 표현으로 예측된 문장 8: 이혼이 무슨 자랑이라고.. 옛날같으면 평생고개를 못들고 다니는 죄인인데.
혐오 표현으로 예측된 문장 9: 여자 얼굴값하게 생긴네..색기도 많아 보이고... 남자가 불쌍타..ㅈㅈ
혐오 표현으로 예측된 문장 10: 니미 백년전꺼도 꺼내봐라



MISLABELING

3

챗봇 소개

교육

인천경기

교실 멍드는 혐오발언, '학교 문화' 때문

“극도의 모욕감이나 멸시감, 폭력을 당할지도 모른다는 위협까지”

이민선 (doule10)

등록 | 2019.01.07 17:30

수정 | 2019.01.07 17:30

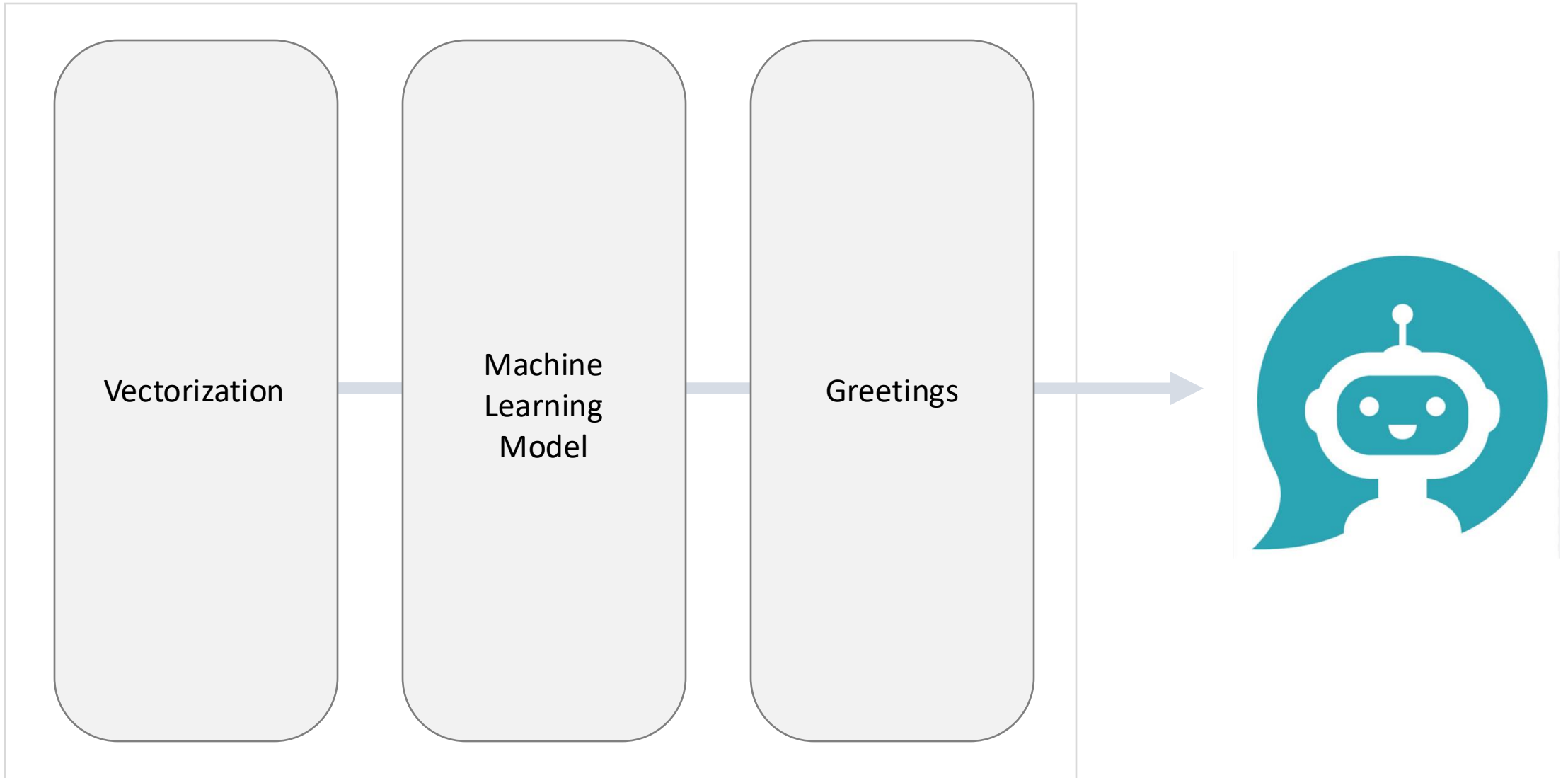


▲ 경기도교육청 전경 © 경기도 교육청

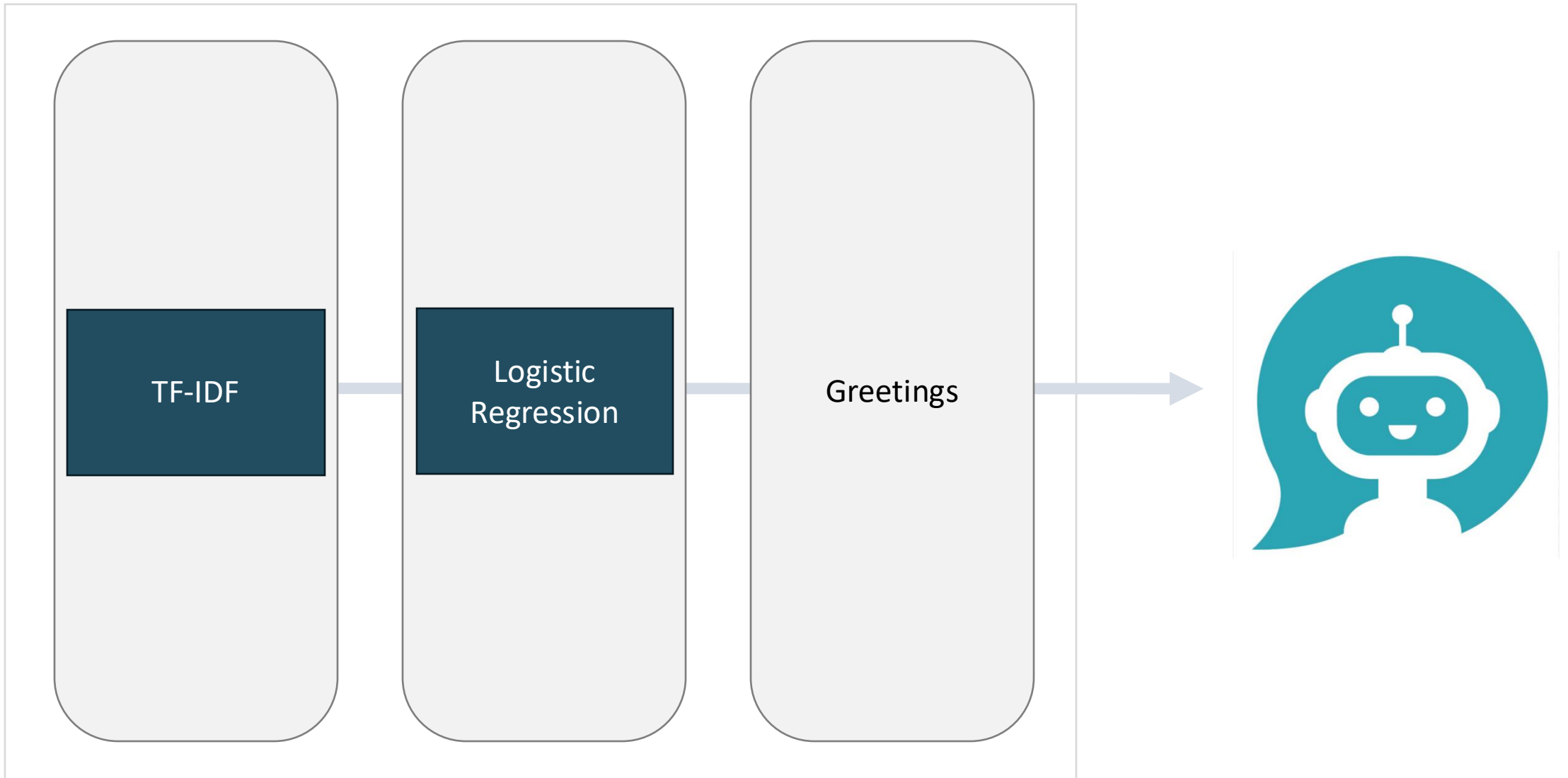


'네가 우리 반 폭탄이야, 너 생긴 게 정말 거지같아!'

챗봇 소개



챗봇 소개



WHY TF-IDF



협오표현은 많은 경우 명확한 단어나 표현으로 드러남 (한남,좌빨)

+

사전 학습된 모델은 댓글에 나오는 신조어나 변형된 문법에 약할 수 있다

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

tf_{ij} = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

텍스트 내에서 단어의 중요도 → 수치화

```
Shape of TF-IDF training data: (78977, 5000)
Shape of TF-IDF validation data: (8776, 5000)
Shape of TF-IDF test data: (21939, 5000)
<class 'numpy.ndarray'>
array(['말아먹고', '말투', '맘대로', '맘충', '맙시'], dtype=object)
```


WHY LOGISTIC REGRESSION



이진 분류

+

F1-score가 괜찮게 나왔다

챗봇 소개

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.70	0.75	3889
1	0.78	0.87	0.83	4887
accuracy			0.80	8776
macro avg	0.80	0.79	0.79	8776
weighted avg	0.80	0.80	0.79	8776

Confusion Matrix:

```
[[2719 1170]
 [ 625 4262]]
```

검증 데이터셋으로 무엇 확인?



모델이 훈련 데이터셋에 오버피팅 되어있지는 않는가?

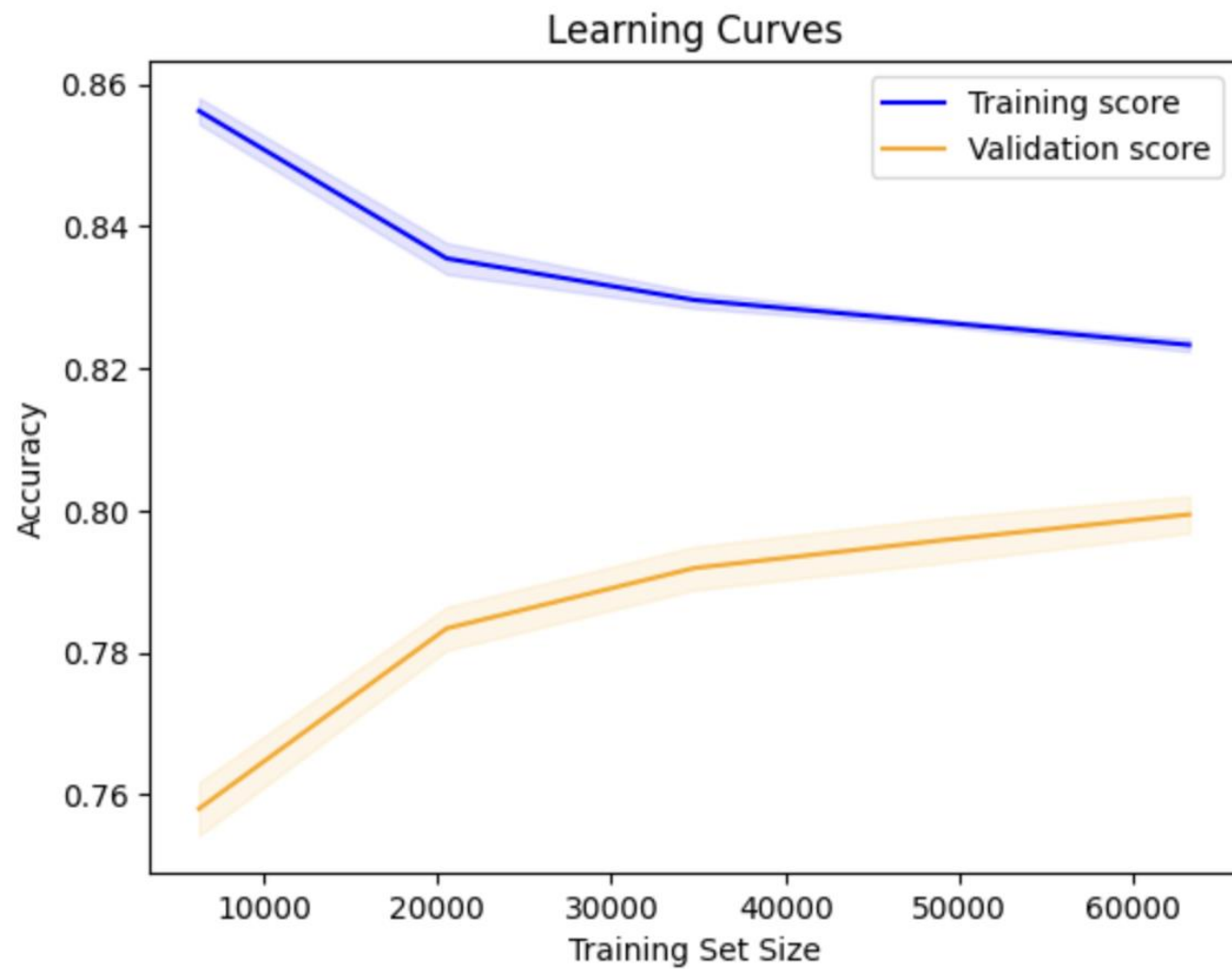
+

recall 값을 높여보는 것은 어떨까?

모델이 훈련 데이터셋에 오버피팅 되어있지는 않는가?

+

recall 값을 높여보는것은 어떨까?



모델이 훈련 데이터셋에 오버피팅 되어있지는 않는가?

+

recall 값을 높여보는것은 어떨까?

2719	TP	FN	1170
625	FP	TN	4262

챗봇 소개

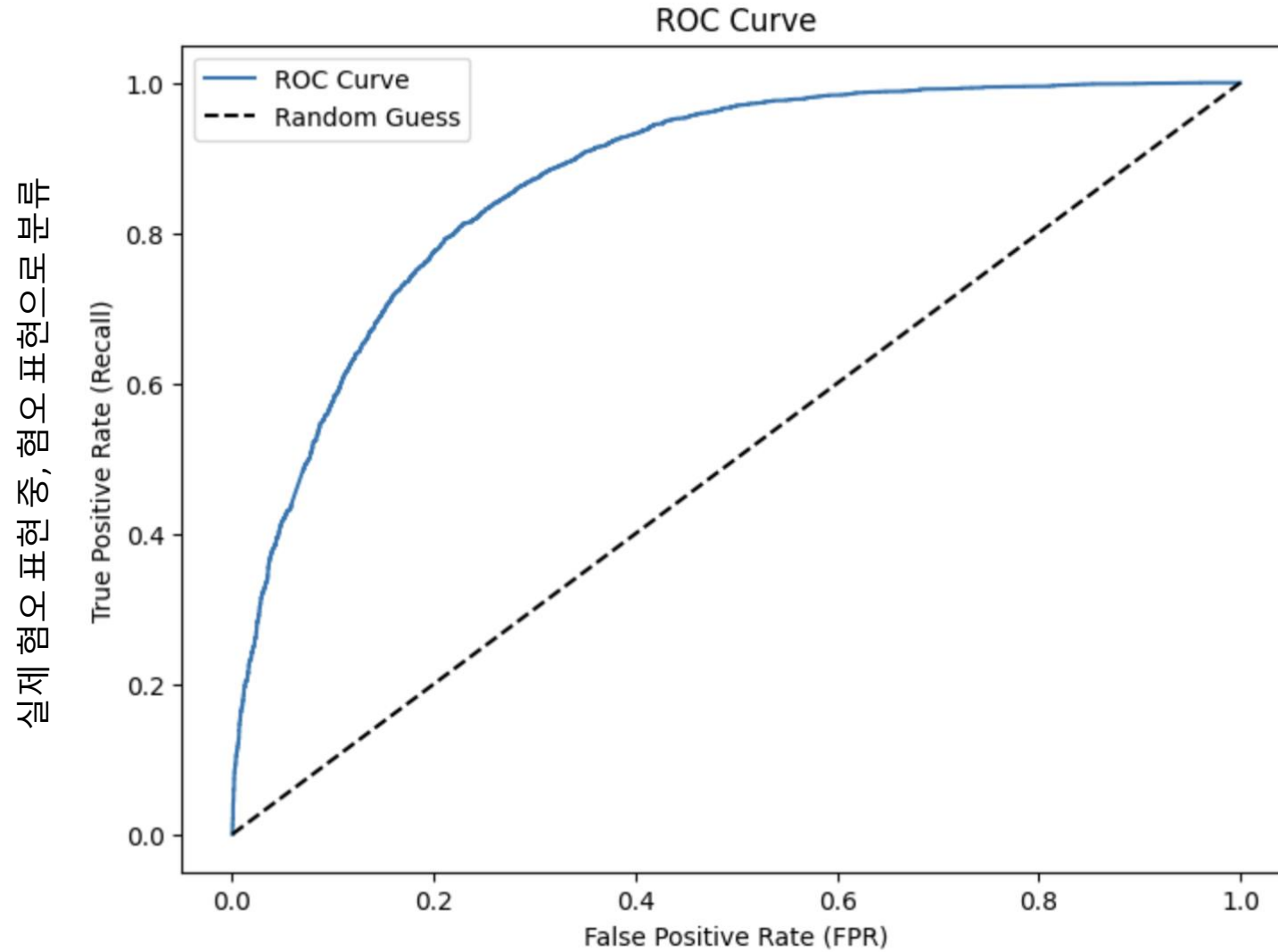
혐오 표현 아닌걸, 혐오 표현 아닌걸로 판단	혐오 표현인데, 아닌것으로 판단
<div>2719</div> <div>TP</div>	<div>1170</div> <div>FN</div>
<div>625</div> <div>FP</div>	<div>4262</div> <div>TN</div>
혐오 표현 아닌데, 맞는것으로 판단	혐오 표현인걸, 혐오표현으로 판단

혐오 표현을 30%나 놓치는 문제가 있다



$$Recall = \frac{TP_{(2719)}}{TP_{(2719)} + FN_{(1170)}}$$

혐오 표현에 민감한 모델 !



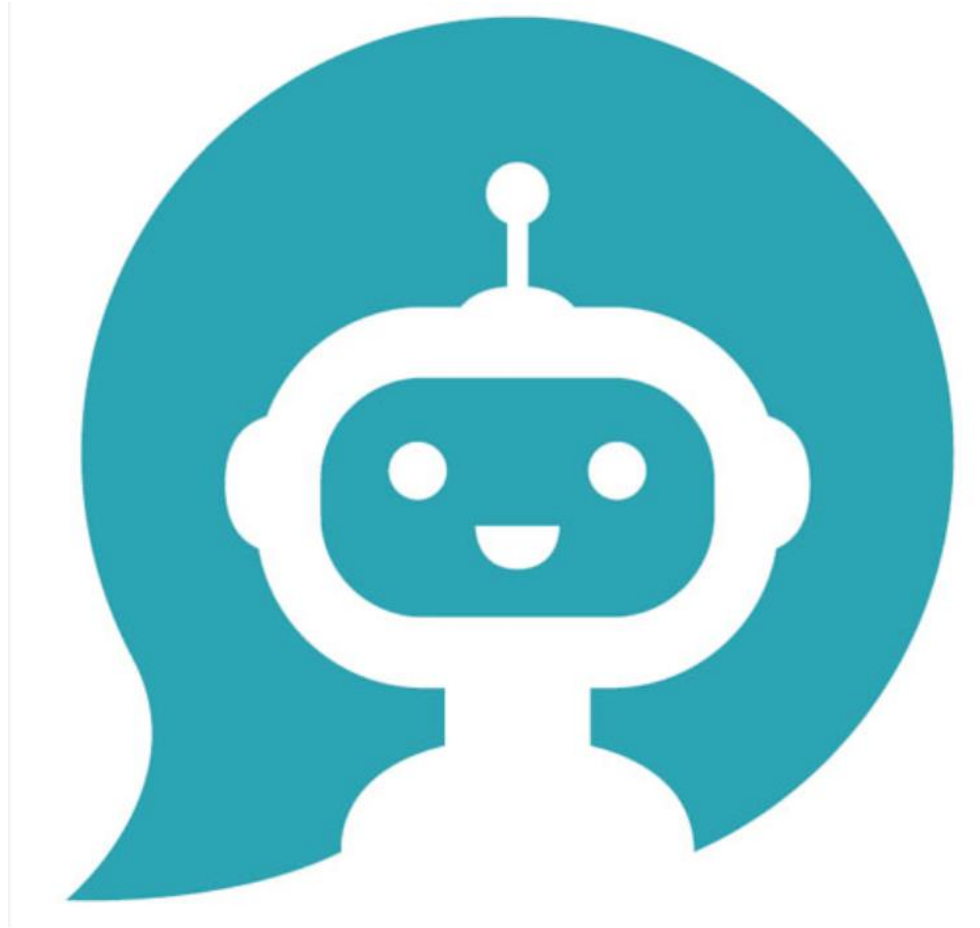
비혐오 표현을, 혐오 표현으로 잘못 분류

Classification Report with Optimal Threshold:

	precision	recall	f1-score	support
0	0.77	0.77	0.77	3889
1	0.82	0.81	0.81	4887
accuracy			0.79	8776
macro avg	0.79	0.79	0.79	8776
weighted avg	0.79	0.79	0.79	8776

Confusion Matrix with Optimal Threshold:

```
[[3000  889]
 [ 918 3969]]
```



챗봇 소개

혐오 표현 챗봇 준비가 완료되었습니다! 입력을 해주세요!

학생: 네이년 댓글들이 개거품 무는거 보이 잘되고 있는듯.

챗봇: 해당 문장은 사용하시기에 충분합니다!



학생: 욕 쳐하면서 방송하는 새깡이를 공중파에 나오게하네...

챗봇: 해당 문장은 사용하시기에 충분합니다!



학생: 한남들 너무 싫다

챗봇: 해당 문장은 혐오 표현을 담고 있습니다 ㅏㅏ

학생: 국민팔이 개색 똥고물 송편이나 쳐먹어라 괴뢰들아

챗봇: 해당 문장은 혐오 표현을 담고 있습니다 ㅏㅏ

학생: 전라도 + 공무원 + 빨갱이 = 대국민 사기꾼

챗봇: 해당 문장은 혐오 표현을 담고 있습니다 ㅏㅏ

학생: 안녕히계세요

다음번에 봐요!

최근 기사에는 어떻게 반응할까?



혐오 표현 챗봇 준비가 완료되었습니다! 입력을 해주세요!

학생: 저것들 1찍들인가???

챗봇: 해당 문장은 사용하지기에 충분합니다!



학생: 니들이 환경오염 같으다.

챗봇: 해당 문장은 사용하지기에 충분합니다!



학생: 설령 건희는 개인 문제이고 재인이는 국가의 안위와 관련있는것이고 재명이는 비리 문제인것 같은데 ~~ 더듬어 개들은 판단도 못하는 개들의 집단들~~

챗봇: 해당 문장은 혐오 표현을 담고 있습니다 ㅏㅏ

학생: 무슨 특검 ? 이죄명 대장동등 특검이나 해라 또 내로남불

챗봇: 해당 문장은 사용하지기에 충분합니다!



학생: 줄리 집구석은 사기꾼 집단 허위 위조 협박 공갈 투기 등 범죄 종합 선물세트

챗봇: 해당 문장은 혐오 표현을 담고 있습니다 ㅏㅏ

학생: 문제양이 간첩인 확실한 증거가 나왔다

챗봇: 해당 문장은 혐오 표현을 담고 있습니다 ㅏㅏ

학생: 이 폐미기자 또 선동하네. 어디 어린애들이 방시혁을 지우고 싶다고 하나. 그런애를 본적이 없구만. 방시혁 지우라고 선동을 하고 있구만 방법까지 가르쳐주면서.

챗봇: 해당 문장은 혐오 표현을 담고 있습니다 ㅏㅏ

모델의 한계?



문맥을 이해하진 못 함

&

학습되지 않은 단어들에 취약하다

모델의 한계?



문맥을 이해하진 못 함

&

학습되지 않은 단어들에 취약하다

챗봇 소개

원래 목적은 학생들 아니었나?



학생들이 사용하는 어휘는 다르다



신조어	뜻
월거지/전거지	월세/전세 거주자를 거지에 빗대 비하하는 말
엘사	LH 아파트(임대 아파트)에 사는 사람
휴거	휴먼시아 거주자를 거지에 빗대 비하하는 말
이백충/삼백충	부모의 월 소득 200/300만 원 이하인 사람을 벌레에 빗대 낮잡아보는 말
-수저	출생 시 부모의 부유 정도를 수저 계급으로 나눈 말. 최근엔 금수저 위에 다이아몬드수저, 흙수저 밑에 똥수저가 생김
조물주 위에 건물주/장래 희망 건물주	우주를 만들고 다스리는 조물주보다 건물을 세워 소유하고 있는 건물주가 낫다는 말/돈을 얼마 못 버는 평범한 직업보다 건물주를 장래 희망으로 생각하는 요즘 학생들의 말

▲ 온라인에서 논란이 된 초등학생들이 쓰고 있는 은어들

챗봇 소개

"이백충" : "해당 문장은 혐오 표현을 담고 있습니다 ㅜㅜ",
 "삼백충" : "해당 문장은 혐오 표현을 담고 있습니다 ㅜㅜ",
 "폭탄이야" : "해당 문장은 혐오 표현을 담고 있습니다 ㅜㅜ",
 "엠이" : "(엄마)를 써보는건 어떨까요?",
 "담탱이" : "(담임)을 써보는건 어떨까요?",
 "한남" : "(남자)를 써보는건 어떨까요?",
 "한녀" : "(여자)를 써보는건 어떨까요?",



 혐오 표현 챗봇 준비가 완료되었습니다! 입력을 해주세요!

학생: 담탱이 완전 싫지 않냐?

챗봇: (담임)을 써보는건 어떨까요?

 학생: 너네집 이백충이야?

챗봇: 해당 문장은 혐오 표현을 담고 있습니다 ㅜㅜ

 학생: 톨딱들 왓케 많냐

챗봇: (어르신)을 써보는건 어떨까요?

단어사전 추가

대체 어휘 제공

학습되지 않은 단어 추가

4

결론

결론 |

mislabeled

recall

어휘사전 추가

Thank You !

질의응답