

LIMEBOT

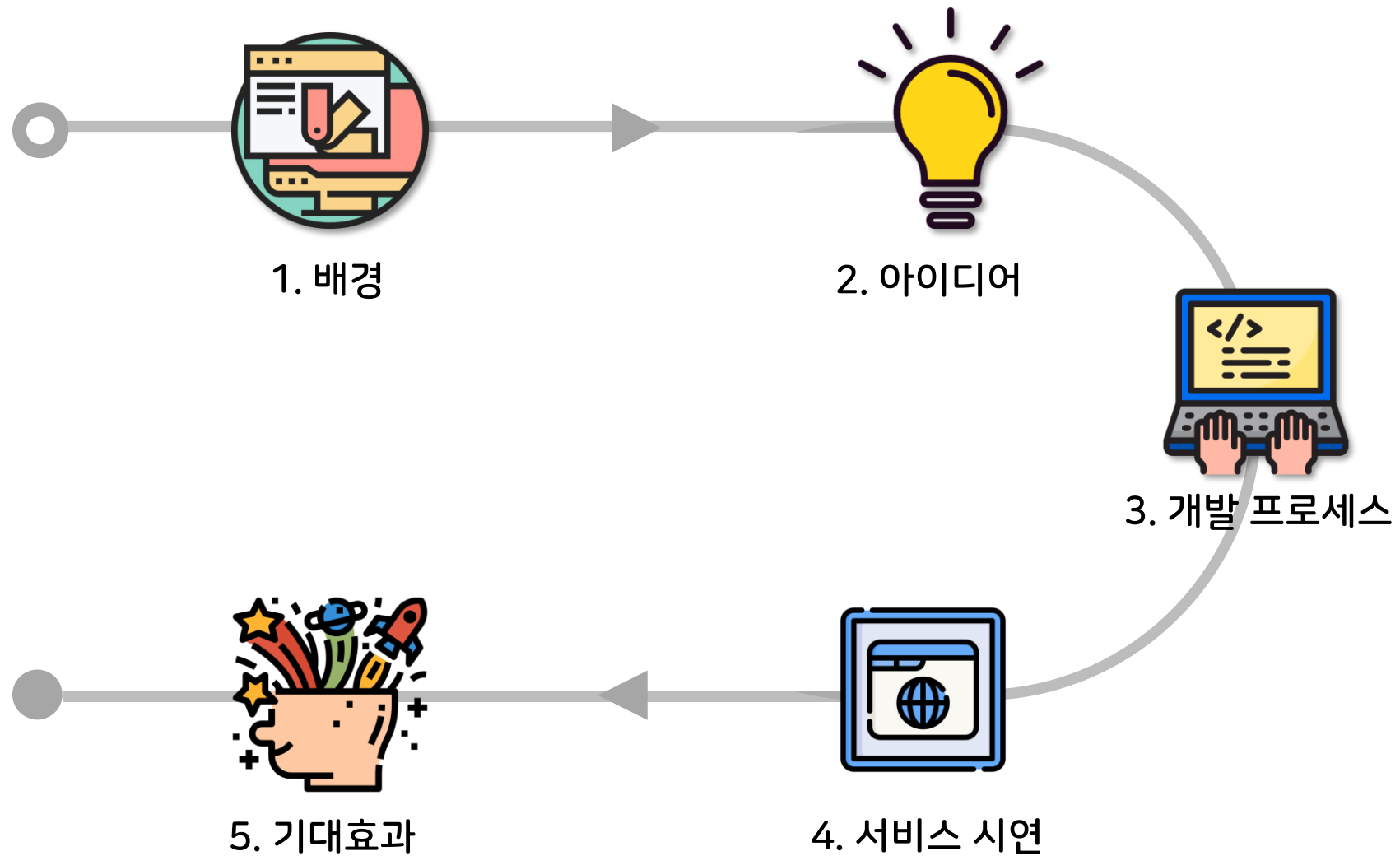
뉴스 악성댓글 필터링 서비스

나의LIME오렌지나무



이은호, 이원재, 임수영, 천지우

CONTENTS





이 발표에는 **비속어와 욕설**이 포함되어 있습니다.

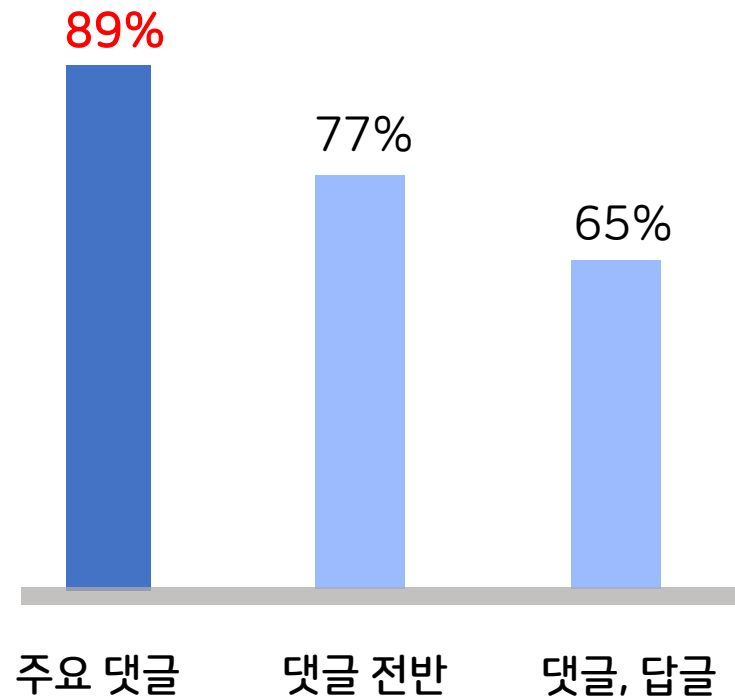
실제 뉴스 기사의 악성 댓글을 활용해 만든 자료로
불쾌감을 일으킬 수 있는 내용이 포함되어 있습니다.

1. 배경

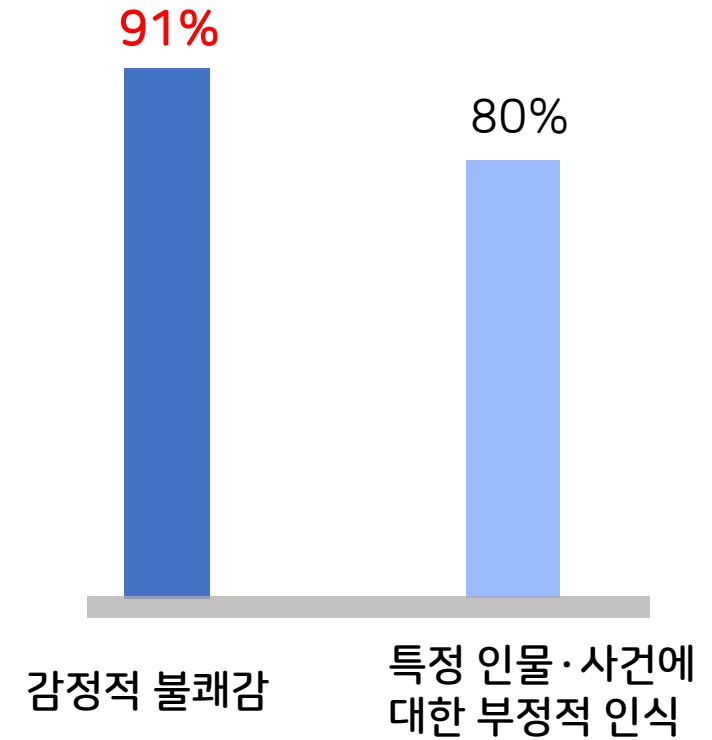
출처 : 한국리서치

댓글을 읽는 뉴스 소비 문화

온라인 뉴스 소비 시 "댓글"을 함께 읽는다



댓글을 읽으며 "불쾌감"을 느낀 경험이 있다



1. 배경

포털 뉴스 아웃링크 논의



빨라지는 '아웃링크' 시계... 언론사 수익 양극화로 이어지나

인수위의 포털 뉴스 개편·민주당의 망법 개정안이 언론계에 미칠 파장

포털 뉴스에 대한 아웃링크 도입 논의

언론사 홈페이지의 뉴스 긍정적 소비경험 부족
(많은 광고 배너, 댓글 어뷰징, 낮은 속도)



포털사이트의 댓글 서비스에 대한 AI모델 개발,
악성 댓글 운영 정책 개편 등 소비경험 향상을 위한 노력



댓글을 선별하여 게재하는 뉴욕타임즈
악성 댓글을 걸러내는 알고리즘 개발 및 인력에 대한 투자

1. 배경

악성 댓글 관리 현황 ①

1) 포털 뉴스(네이버)

AI를 활용한 악성 댓글 검출



씨발

스브

2) 언론사 뉴스 홈(동아일보)

금칙어와 정확히 일치하는 댓글만 검출



1. 배경

악성 댓글 관리 현황 ②

예쁜 쓰레기 아닌가


예쁜 쓰레기 아닌가

10 / 300

등록



예쁜 쓰레기



**상처 주는 표현이
포함되어 있지 않나요?**

클린봇 탐지결과 부적절한 표현이
감지됩니다.
반복 등록시 이용이 제한될 수 있습니다.

10 / 300

취소

등록



사용자가 의아해할 수 있는
클린봇의 악플 제재 방식

악성 댓글을 정확히 분류하면서도
이유를 설명할 수 있는 모델을 만들 수는 없을까?



00일보 웹사이트 운영 관리자

매일 작성되는 많은 댓글을
손쉽게 관리하기 위한 모델

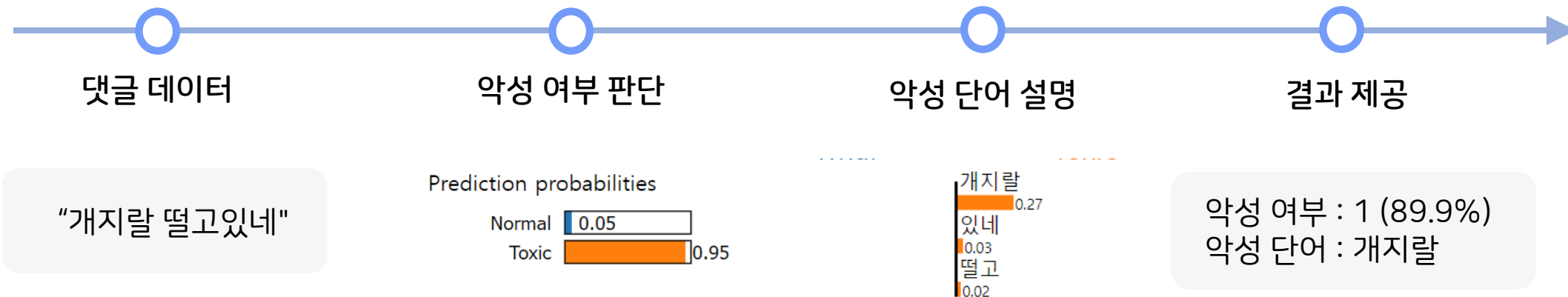
뉴스의 특성을 반영하여
언론사에 적용가능한 모델

악성 댓글 판단 기준이 명확하여
쉽게 유지/보수 가능한 모델

2. 아이디어 내용



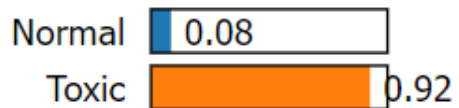
악성 댓글 이유를 제공하는 투명한 클린봇
뉴스 카테고리에 맞춤형된 클린봇



설명가능한 AI 모델

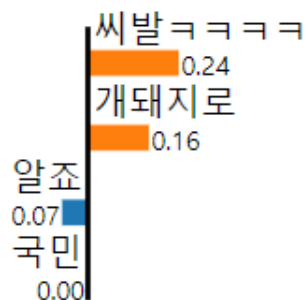
Input : 씨발ㅋㅋㅋㅋ 국민 개돼지로 알죠?

Prediction probabilities



Normal

Toxic



Text with highlighted words
씨발ㅋㅋㅋㅋ 국민 개돼지로 알죠?

설명가능한 AI 모델(XAI)의 일종인 LIME 모델

- ① 모델이 댓글을 악성으로 판단할 확률과 근거를 모두 설명함
- ② 모델이 판단 근거를 제공하기 때문에 모델의 유지/보수에도 유용

서비스 주요 차별점 ①

2. 아이디어 내용

뉴스 카테고리별 기준 산정

성범죄 기사 中 일부 댓글



밤늦게 남자랑 단둘이 술먹는년이.제정신 이냐



판새 부터싸다 잡아 죽이자



더치페이 안 하는 창녀여서 죽인듯ㅋㅋㅋㅋㅋ

사회 기사 中 일부 댓글



도대체 나이 쳐먹고 왜 그렇게 사냐????



지금은 모든 범죄가 예전같지 않아요.국개들 제발 일좀하자.



얼굴을 알아야 잡든가~신고 하지요~얼른 공개해 주세요~

범죄 관련 기사에 상대적으로 높은 욕설 및 혐오 표현



카테고리 특성에 맞게 악성 판별 기준치를 조절하여
2차 가해, 무분별한 비난 방지

서비스 주요 차별점 ②

2. 아이디어 내용

언론사 정책에 따른 다양한 결과물 제공


기사 ID	댓글	작성자
32794125	지랄을하네 주둥이찢어야됨	wkd****
...



경고 메시지 팝업



악성 댓글이 탐지되었습니다

 **경고**
욕설, 폭력, 혐오, 차별에
해당하는 단어가
포함되지 않았나요?

악성 단어 하이라이트



악성 댓글 판독 결과입니다

댓글	악성확률
지 랄 을하네 주둥이 찢 어야됨	0.89
...	...

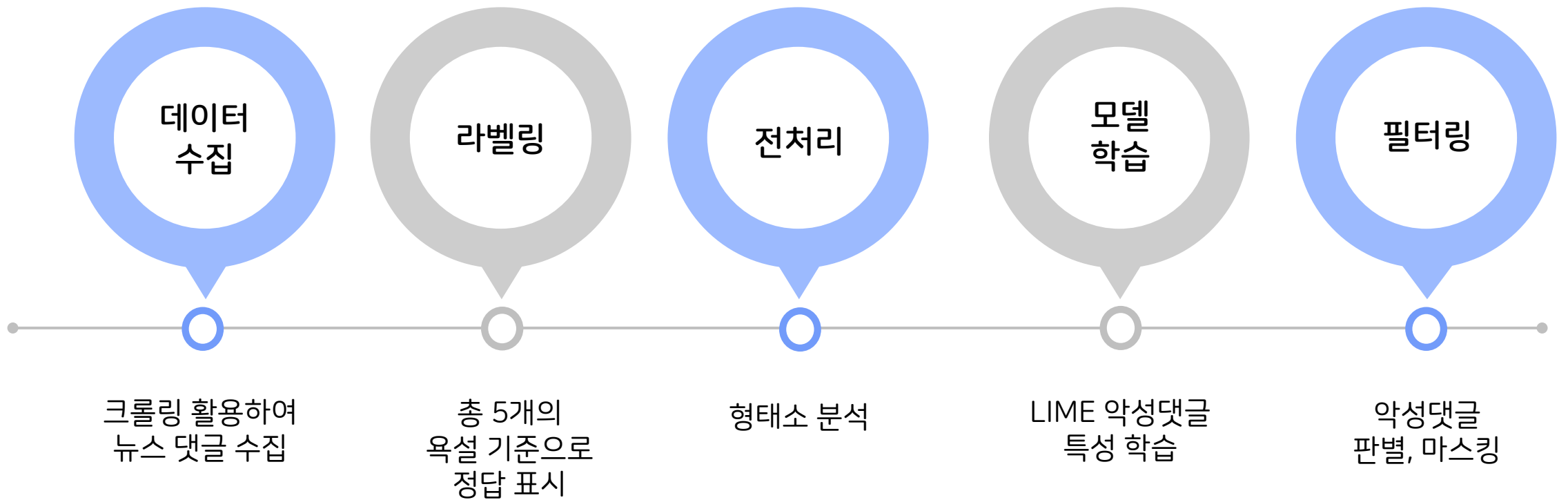
악성 댓글 마스크



악성 댓글을 마스크합니다

지랄을하네 주둥이찢어야됨
↓
을하네 주둥이**

3. 개발 프로세스



1) 데이터 수집

3. 개발 프로세스

크롤링

포털사이트

정치, 경제, 사회 뉴스 댓글 11,974개

오픈 데이터셋

Korean-malicious-comments-data¹ 10,000개

Curse-detection-data² 5,825개



총 27,799개의 한국어 뉴스 댓글 데이터셋

1. <https://github.com/ZIZUN/korean-malicious-comments-dataset>

2. <https://github.com/2runo/Curse-detection-data>

2) 라벨링

3. 개발 프로세스

분류		예시
욕설		병신, 존나
폭력위협/범죄조장		찢어 죽인다
외설		걸레, 창녀
혐오	성 혐오	똥꼬충, 한남, 한녀
	세대 혐오	틀딱, 찜민
	인종/지역 혐오	짱깨, 쪽빠리, 상도남
	장애 혐오	정신병자
	정치 성향 혐오	찢재명, 윤돼지
	직업 혐오	기레기, 검색

0 : 기준에 해당하지 않는 일반 댓글
1 : 기준에 해당하는 악성 댓글

3) 전처리

3. 개발 프로세스

BEFORE

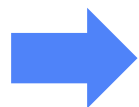
"짱개새끼들고국으로돌아가라"



빅카인즈
형태소 분석 API

AFTER

문장	짱개새끼들고국으로돌아가라						
형태소	짱개	새끼	들	고국	으로	돌아가	라
태그명	명사	명사	접미사	명사	조사	동사	어미



뉴스 댓글 특성상 띄어쓰기가 제대로 지켜지지 않기 때문에,
띄어쓰기가 아닌 **형태소 단위**로 구분하여 모델 학습

4) 모델학습

3. 개발 프로세스

① 데이터셋

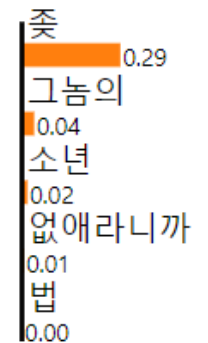
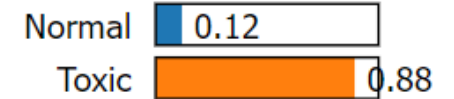
no	comment	label
1	그놈의 축법 = 좇법 없애라니까	1
2	앞날에 축복이 가득하기를...	0
3
4

② LIME : 악성댓글 학습



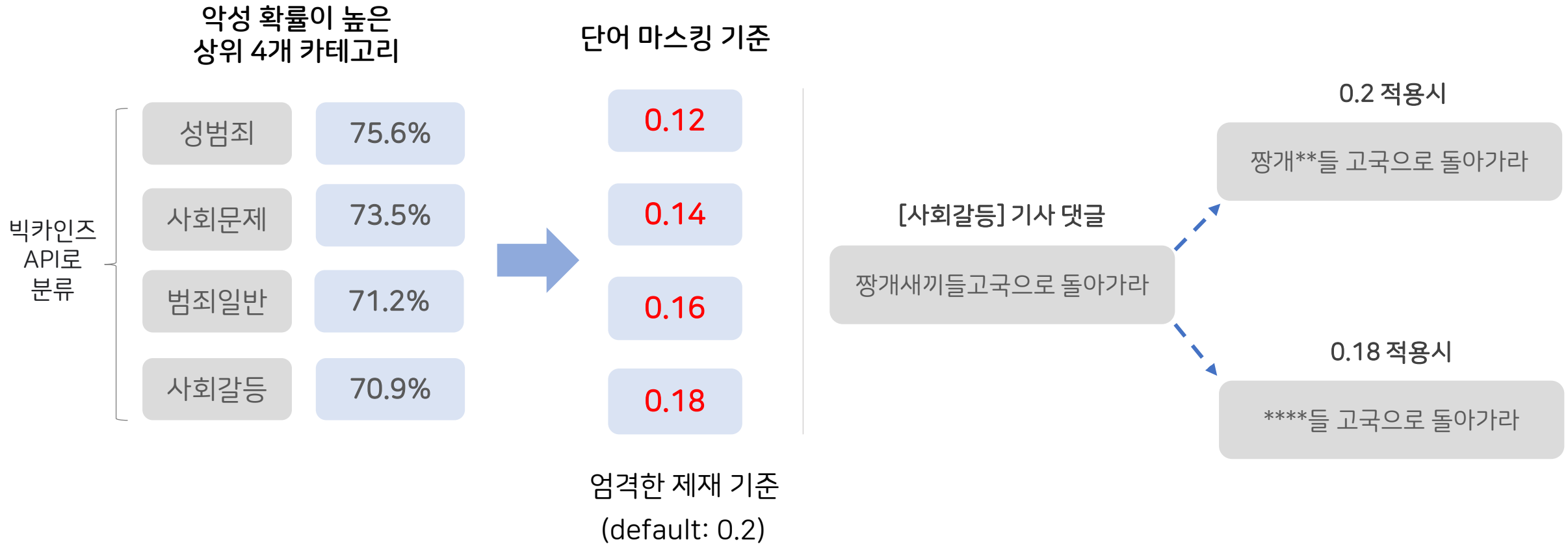
③ 댓글 설명

Prediction probabilities



분류 정확도 : 93%

5) 결과물 출력




※ 특정 카테고리에 속하지 않으면 본래 기준인 0.2를 적용

4. 서비스 시연

LIMEBOT_악성댓글 필터링 서비스

https://limebot.imweb.me/index

LIMEBOT

HOME악성댓글 필터링 서비스활용 예시라임봇이란?

☺️ 저희의 서비스는 다음과 같습니다 ☺️

악성 확률 판독

27,799개의 데이터셋으로 학습된 AI가 뉴스 댓글의 악성 확률을 판독합니다. 악성 정도가 높은 순서대로 정렬하여 결과를 제공합니다. 관리자는 LIME_BOT을 활용해 뉴스 홈페이지의 댓글 커뮤니티를 관리하고, 부적절한 댓글을 우선적으로 제재 할 수 있습니다.
* MORE를 눌러 더 많은 정보를 확인해보세요.

MORE

악성 단어 하이라이트

설명가능한 AI(XAI) LIME 모델을 활용하여 뉴스 댓글의 악성 단어를 찾아냅니다. 악성 정도가 높은 단어에 하이라이트로 표시해 결과를 제공합니다. LIME_BOT을 활용해 악플러에게 제재 이유를 설명할 수 있습니다. 표현의 자유를 침해하지 않으면서도 건전한 댓글 문화를 조성할 수 있습니다.
* MORE를 눌러 더 많은 정보를 확인해보세요.

MORE

자동 마스킹


기사 카테고리 별로 가중치를 조절하여 AI가 불편한 단어를 자동으로 마스킹 합니다. 성범죄, 사회 갈등 등 일반 개인에게 2차 가해가 우려되는 기사의 댓글은 더욱 엄격하게 마스킹 합니다. LIME_BOT을 활용해 부적절한 댓글 노출을 사전에 차단할 수 있습니다.
* MORE를 눌러 더 많은 정보를 확인해보세요.

MORE

4. 서비스 시연

LIMEBOT_악성댓글 필터링 서비스

https://limebot.imweb.me/index



HOME악성댓글 필터링 서비스활용 예시라임봇이란?

😊 악성 댓글 필터링 시연 😊

뉴스 댓글 파일을 업로드 하세요.

news_comments_sample.xlsx (10 Kb) X

원하는 서비스를 선택하세요.

악성 확률 출력

서비스 : 악성 확률 출력

뉴스 카테고리를 설정하세요.

☐ 성범죄

☐ 일반범죄

☐ 사회갈등

☐ 사회문제

☒ 그 외


카테고리 : 그 외

결과보기

4. 서비스 시연

LIMEBOT_악성댓글 필터링 서비스

https://limebot.imweb.me/archive

HOME악성댓글 필터링 서비스활용 예시라임봇이란?

👹 댓글 악성 확률 결과입니다 👹

* 악성 확률이 높은 상위 5개 결과를 확인하실 수 있습니다

* 전체 댓글 결과는 "다운로드"를 통해 엑셀로 확인하실 수 있습니다

선택 옵션

- file_name : "news_comments_sample.xlsx (10Kb)"
- service_menu : "악성 확률 출력"
- category : "기타"

결과


기사ID	댓글	작성자	작성일	악성 확률
98546281	쓰부 랄 년	rainy**	2022.10.05	96.47%
14365269	윤재앙 또 설치고있네 닥치고짜저있어라제발	96day**	2022.10.05	94.30%
54920004	짱개 개새끼들 그냥 알아서 뒤졌으면 ~~	march3**	2022.10.06	93.01%
52445677	씨발ㅋㅋㅋㅋ 국민 개돼지로 알죠?	dla****	2022.09.30	91.64%
15937802	엥간히 해라 미친 GSGG들아...	rkd***	2022.10.01	89.94%
...

4. 서비스 시연

LIMEBOT_악성댓글 필터링 서비스

+

← → ↻ https://limebot.imweb.me/archive



HOME 악성댓글 필터링 서비스 활용 예시 라임봇이란?

결과

기사ID	댓글	작성자	작성일	악성 확률
98546281	쓰부 랄 년	rainy**	2022.10.05	96.47%
14365269	온재앙 또 설치고있네 덕치고짜져잇어라제발	96day**	2022.10.05	94.30%
54920004	짱개 개새끼들 그냥 알아서 뒤졌으면 ~~	march3**	2022.10.06	93.01%
52445677	씨발ㅋㅋㅋㅋ 국민 개돼지로 알죠?	dla****	2022.09.30	91.64%
15937802	헐간히 해라 미친 GSG들아...	rkd***	2022.10.01	89.94%
***	***	***	***	***

다운로드

돌아가기

4. 서비스 시연

악성 확률 판독

기사ID	댓글	작성자	작성일	악성확률
98546281	쓰부 랄 년	rainy**	2022.10.05	0.9647
14365269	윤재양 또 설치고있네 닥치고짜져있어라제발	96day**	2022.10.05	0.9430
54920004	짱개 개새끼들 그냥 알아서 뒤졌으면 ~~	march3**	2022.10.06	0.9301

악성 단어 하이라이트

기사ID	댓글	작성자	작성일	하이라이트
15937802	엔간히 해라	rainy**	2022.10.05	쓰부 랄 년
79900021	ㅂㅅ 육갑 떨고	96day**	2022.10.05	윤재양 또 설치고있네 닥치고짜져있어라제발
32794125	미련싸가지로	march3**	2022.10.06	짱개 개새끼들 그냥 알아서 뒤졌으면 ~~
34998866	ㅂㅅ 국짐 클	dla*****	2022.09.30	씨발ㅋㅋㅋㅋ 국민 개돼지로 알죠?
45502794	ㅋㅋㅋㅋㅋ	rkd***	2022.10.01	엔간히 해라 미친 GSGG들아...

악성 댓글 마스크

기사ID	댓글	작성자	작성일	마스크
98546281	쓰부 랄 년	rainy**	2022.10.05	*** **
14365269	윤재양 또 설치고있네 닥치고짜져있어라제발	96day**	2022.10.05	*** 또 설치고있네 ***짜져있어라제발
54920004	짱개 개새끼들 그냥 알아서 뒤졌으면 ~~	march3**	2022.10.06	** ***** 그냥 알아서 ***** ~~
52445677	씨발ㅋㅋㅋㅋ 국민 개돼지로 알죠?	dla*****	2022.09.30	**ㅋㅋㅋㅋ 국민 ***** 알죠?
15937802	엔간히 해라 미친 GSGG들아...	rkd***	2022.10.01	엔간히 해라 미친 *****들아...
79900021	ㅂㅅ 육갑 떨고있네. 똤.저러	what**	2022.10.07	** 육갑 떨고있네. *** 고마
32794125	미련싸가지로생겼네주둥이를찢어야지 저런건	dnjs****	2022.10.07	미련***로생겼네주둥이를**** 저런건
34998866	ㅂㅅ 국짐 클래스	light***	2022.10.02	ㅂㅅ ** 클래스
45502794	ㅋㅋㅋㅋㅋ 그럼그렇지..	kim09**	2022.10.05	ㅋㅋㅋㅋㅋ 그럼그렇지..
23461357	인간적으로 너무한 거 아닌가	koe*****	2022.09.27	인간적으로 너무한 거 아닌가
87877236	우리나라엔 해당사항없는듯?	lms488**	2022.09.25	우리나라엔 해당사항없는듯?
39400876	유류세 어디까지 오르는거임 ㄹㅇ	luv2***	2022.09.27	유류세 어디까지 오르는거임 ㄹㅇ

5. 기대효과

언론사

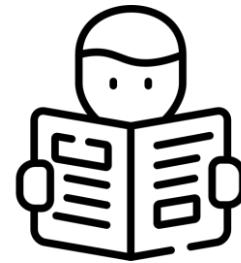


악성 댓글로 인한 피해 방지



LIMEBOT

이용자



표현의 자유 보장

5. 기대효과

언론사



1. 댓글 관리의 효율성

- 악성댓글 모니터링에 필요한 자원감소
- 포털 뉴스 아웃링크에 대비한 자체 댓글 관리 정책 마련

2. 뉴스 플랫폼 경쟁력 강화

- 건강한 토론이 오가는 댓글 커뮤니티 형성
- 자체 콘텐츠(폴 서비스, 커뮤니티)로 확대 가능

이용자



1. 불쾌감 감소

- 무분별한 욕설, 비속어로부터 노출 방지

2. 피해자 보호

- 사건, 사고 기사 피해자의 신상 노출, 2차 가해 방지

5. 기대효과

건전한 토론 문화 제공 → 언론이 가져야할 또 하나의 의무

언론의 댓글공간은 "건강하고 전문성이 묻어있는 토론 커뮤니티, 포럼"이 되어야 한다 - 뉴욕 타임즈

감사합니다

참고 레퍼런스

- [1] 이동환, [기획] 뉴스기사 댓글에 대한 인식, 한국리서치 정기조사 여론 속의 여론, 2022.02.08, <https://hrcopinon.co.kr/archives/20815>
- [2] 김달아, 빨라지는 '아웃링크' 시계... 언론사 수익 양극화로 이어지나, 한국기자협회, 2022.05.03, http://m.journalist.or.kr/m/m_article.html?no=51498
- [3] 금준경, 아웃링크가 대안? 언론이 네이버보다 잘할 수 있나, 미디어오늘, 2018.04.30, <http://www.mediatoday.co.kr/news/articleView.html?idxno=142462>
- [4] 한국어 욕설 감지 데이터셋, <https://github.com/2runo/Curse-detection-data>
- [5] 한국어 악성 댓글 데이터셋, <https://github.com/ZIZUN/korean-malicious-comments-dataset>
- [6] 박진원, 나영윤, 박규병(2021). 비윤리적 한국어 발언 검출을 위한 새 데이터 세트, ACK 2021 학술발표대회 논문집 (28권 2호)
- [7] 이원석, 이현상(2022). 딥러닝 기술을 활용한 차별 및 혐오 표현 탐지 : 어텐션 기반 다중 채널 CNN 모델링. 한국정보통신학회논문지, 24. 1595~1603.
- [8] 조용래(넥슨코리아 인텔리전스랩스 어뷰징탐지팀), 딥러닝으로 욕설 탐지하기, 2018.07.03, http://ndc.vod.nexoncdn.co.kr/NDC2018/slides/NDC2018_0033/index.html