Better **W**ords for **K**ids Solution

아이들을 위한 미디어의 비윤리적, 혐오 표현 탐지 및 순화 Solution

고려대 경영대 CDTB DAB 경진대회
팀명 : Ai-Gen-cy
팀원 : 산업경영공학 이병주
경영 김태영/ 통계 박종혁 / 통계 하예은

목차

Part 1

서비스 개요

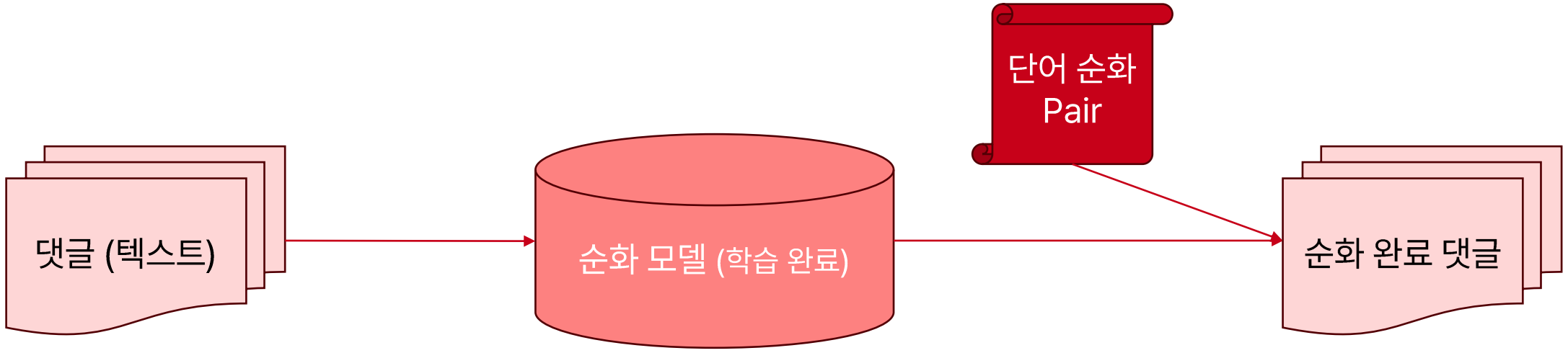
서비스 Flow Chart

문장 입력 → **순화 문장 생성** → **순화 단어 교체**

실제 존재하는 댓글이나
임의로 생성한 표현 입력

혐오 표현에 대한
순화된 문장 생성

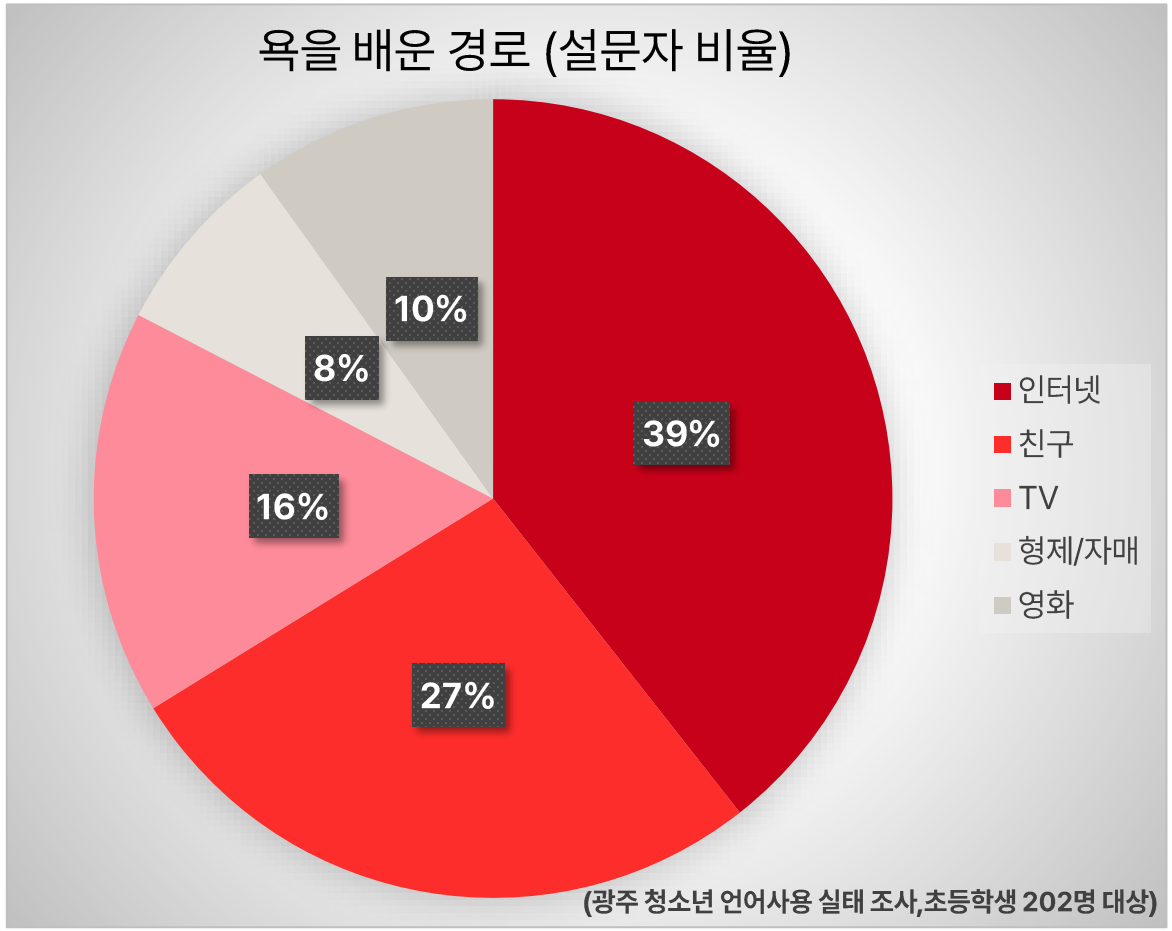
순화문장에서 바뀌지
않은 일부 단어 교체



Part 2

서비스 개발 배경 및 목적

이런 어린이 / 청소년의 언어 문제는 주로 인터넷과 미디어를 통해 비롯됨.
이런 언어문제는 학교 폭력 및 사회 문제로 이어지고 있음.



사회

'에버랜드 살인예고' 10대 적발..."게임서 욕 듣고 화나서"

진영기 기자 ☆

입력 2023.08.05 17:46 수정 2023.08.05 17:46 가가 오늘의

전국

"욕설, 비하에 성추행까지"...졸업생이 폭로한 '학생 인권 침해'

2022년 03월 15일 23시 18분 댓글

혜린이의 비극 그후

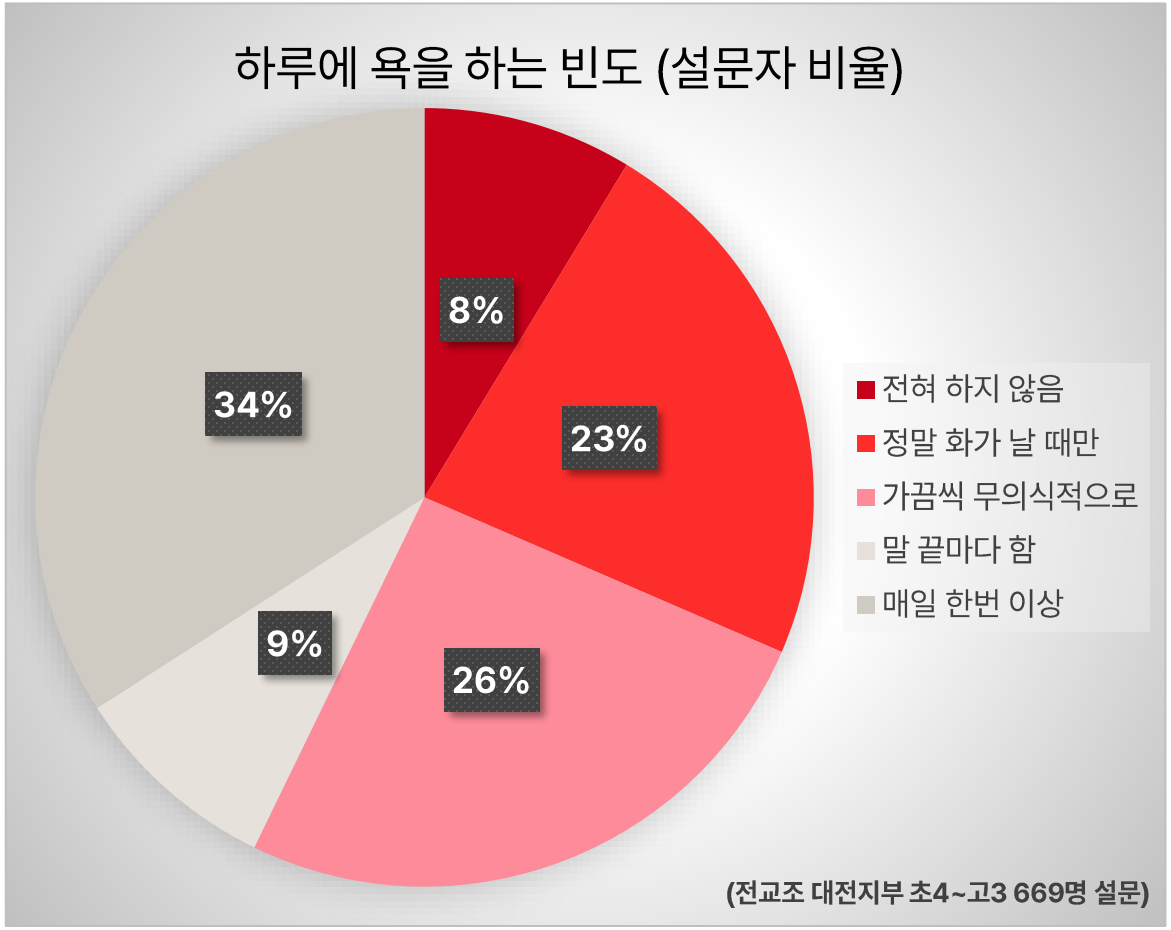
교실선 얌전하던 애도 온라인선 욕설... 채팅방은 출구 없는 감옥

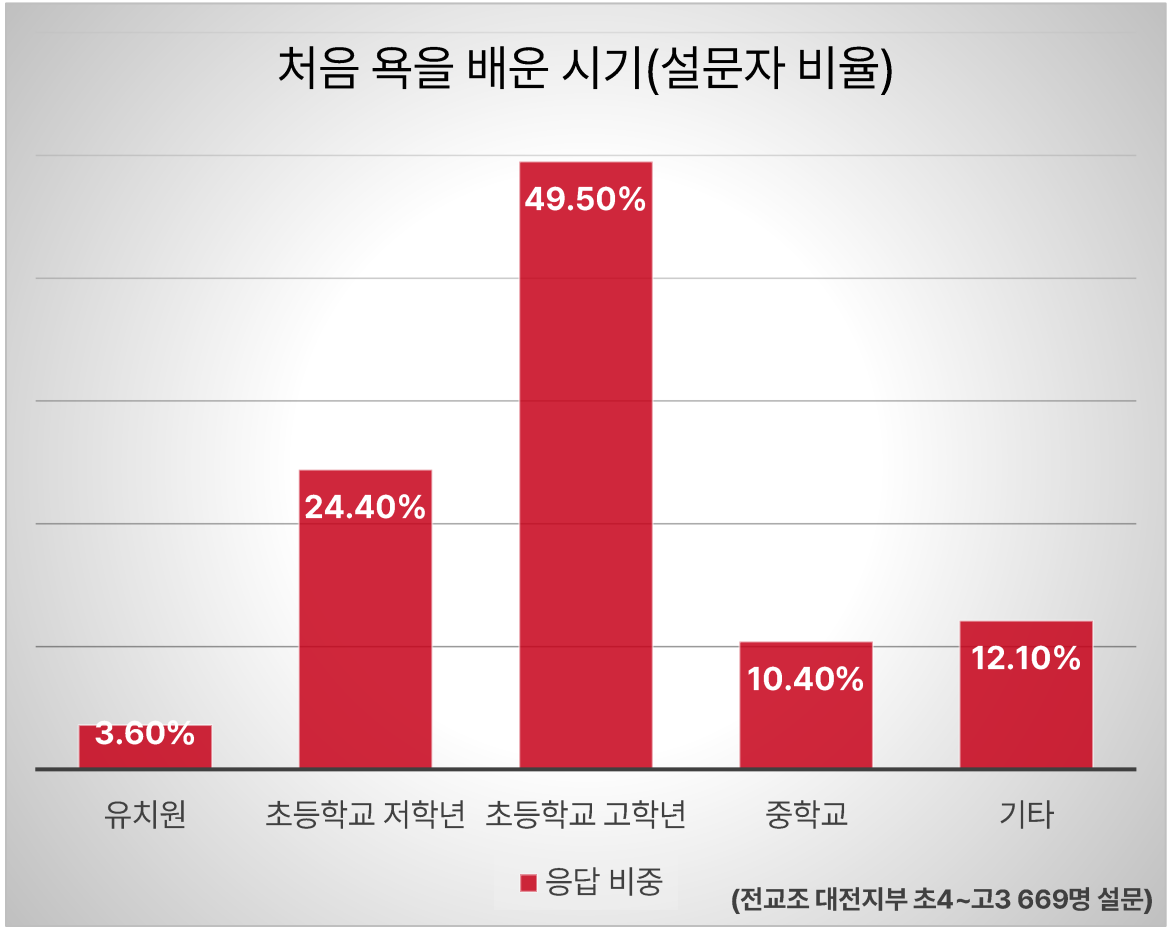
입력 2021.02.01 06:30 수정 2021.02.06 12:52 | 5면 17 2

Part 2

서비스 개발 배경 및 목적

어린이 / 청소년의 언어 문제 (비속어, 욕설)의 심각성이 대두되고 있음.





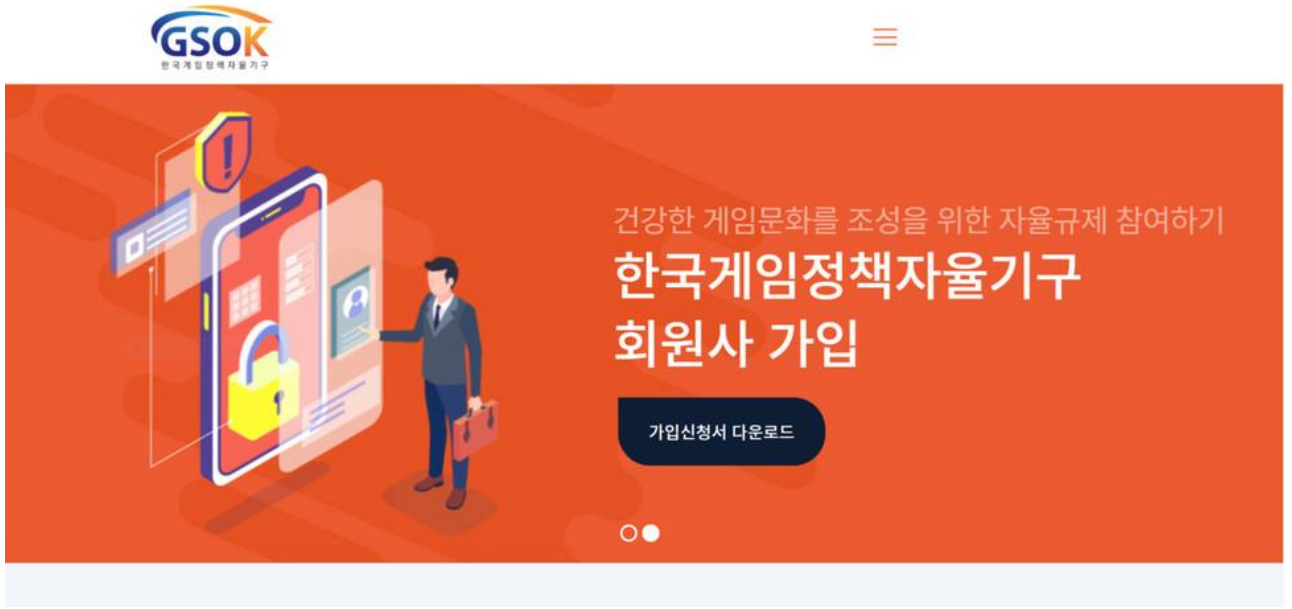
Part 2

서비스 개발 배경 및 목적

kakao NAVER

Smilegate®

혐오 표현을 탐지할 수 있는 플랫폼은 다수 존재.
그러나 탐지하여 제거 및 블라인드 처리할 뿐,
순화된 표현으로 바꾸는 기능의 부재



욕설 및 비속어 관련 공공 정책도 명확히 제시되어 있지 않음.
청소년이 이용가능한 채널에서는 욕설 채팅이 불가함만 명시하고
그 가이드라인은 구체적이지 않음.

서비스 개발 상세 – (1) 사용 데이터

UnSmile

Smilegate AI

Smilegate AI 제공

특정 사회적 집단에 대한 적대적 발언, 조롱, 희화화, 편견을 재생산하는 표현으로 정의된 혐오표현으로 구성

① 혐오표현 : 10,139 → 다중 레이블(8가지)
② 악플/욕설 : 3,929
③ clean : 4,674

문장	여성/가족	남성	성소수자	인종/국적	연령	지역	종교	기타 혐오	악플/욕설	clean	개인지칭
한남 답이 없노	0	1	0	0	0	0	0	0	0	0	0
옛말에 암탉이 울면 나라 망한다고 했다	1	0	0	0	0	0	0	0	0	0	0
꼭 키 작은 급식충이 이런 글 씀	0	0	0	0	1	0	0	0	0	0	0
뭐 어쩌라고 시발	0	0	0	0	0	0	0	0	1	0	0
혐오가 당연한 건 없습니다	0	0	0	0	0	0	0	0	0	1	0

서비스 개발 상세 – (2) 분류 모델

1. 과정 : 혐오표현인지, clean인지 이진분류 → 혐오 분류 중 어느 표현에 해당하는지 분류 (다중분류, # class = 9)
2. 사용 모델 : BertForSequenceClassification 사전학습 모델을 이전의 데이터로 학습시킴
3. 학습 세팅 : optimizer = AdamW, epoch = 3, scheduler = linear scheduler warmup
4. 분류 성능 : valid acc = 81%(이진분류), valid acc = 54%(다중분류)

```
1 for result in pipe("이래서 여자는 게임을 하면 안된다")[0]:
2 | print(result)
```

```
{'label': '여성/가족', 'score': 0.82530522346496 58}
{'label': '남성', 'score': 0.0397251546382904 05}
{'label': '성소수자', 'score': 0.0121443327516317 37}
{'label': '인종/국적', 'score': 0.0231818854808807 37}
{'label': '연령', 'score': 0.0103153083473443 98}
{'label': '지역', 'score': 0.0184548869729042 05}
{'label': '종교', 'score': 0.0112700257450342 18}
{'label': '기타 혐오', 'score': 0.0207340400665998 46}
{'label': '악플/욕설', 'score': 0.057331461459398 27}
{'label': 'clean', 'score': 0.1401053222003364 56}
```

```
1 for result in pipe("일하는 건 즐거워")[0]:
2 | print(result)
```

```
{'label': '여성/가족', 'score': 0.0146993724629282 95}
{'label': '남성', 'score': 0.011865005828440 19}
{'label': '성소수자', 'score': 0.0146441422402858 73}
{'label': '인종/국적', 'score': 0.0149406837299466 13}
{'label': '연령', 'score': 0.0099867479875683 78}
{'label': '지역', 'score': 0.0137908756732940 67}
{'label': '종교', 'score': 0.0134607506915926 93}
{'label': '기타 혐오', 'score': 0.0089365355670452 12}
{'label': '악플/욕설', 'score': 0.076820373535156 25}
{'label': 'clean', 'score': 0.9109398126602173}
```

서비스 개발 상세 – (3) 혐오 표현 순화

1. 혐오 표현 – 순화 표현 문장 쌍 구축

나이쳐먹고 피시방가는 놈들은 대가리에 똥만찬 놈들임	나이먹고 피시방가는 사람들은 머리에 똥만찬 사람들임
이 새끼 ㄹㅇ 좇됐네..징역 30년 맞겠네ㅋㅋㅋㅋㅋㅋㅋ	이 사람 ㄹㅇ 큰일났네..징역 30년 맞겠네ㅋㅋㅋㅋㅋㅋㅋ
그냥 쯧쯧하다 오지 왜 저지랄했누	그냥 쯧쯧하다 오지 왜 저난리쳤냐
개독새끼들 지들끼리 급 나누는 것도 웃김 모든 교회마다 우리교회가 진짜 교회다이지랄하는데 하는짓보면 하나같이 똑같음	개신교들 지들끼리 급 나누는 것도 웃김 모든 교회마다 우리교회가 진짜 교회다이러는데 하는짓보면 하나같이 똑같음
513 한남 내려치기도 아니고 그냥 한남의 잇는그대로를 말한 팩트일 쁜이노	남자 내려치기도 아니고 그냥 남자의 잇는그대로를 말한 팩트일 쁜이다
스페인어 쓰지 말라고 해도 쓰는 개ㅂㅅ 히스패닉들이 많음 미국인데 그거 빡쳐하는 영상있응ㅅ	스페인어 쓰지 말라고 해도 쓰는 멍청한 히스패닉들이 많음 미국인데 그거 화내는 영상있응ㅅ

- 14068개의 문장에서 순화 가능한 문장과 순화 불가능한 문장(단순 욕설)을 분류함
- 순화 가능한 문장 4638개에 대해 혐오 표현과 순화 표현 문장 쌍 데이터를 구축함
- Text Style Transfer(TST)를 위한 학습 데이터로 사용함
  * Test Style Transfer : 문장의 내용은 보존하고 스타일을 바꿔서 출력하는 태스크

서비스 설계 – (3) 혐오 표현 순화

2. 혐오 표현 – 순화 표현 단어 쌍 구축

빠개다	웃다
존나	정말
가시나	여자
와꾸	얼굴
좆본	일본
새끼	사람
년	사람
좆같	싫
스시남	일본남자
검둥이	흑인
짱깨	중국인
…	…
양키	미국인
맘충	자식만 중시하는 엄마
놈	사람
기레기	기자

- TST로 순화되지 않는 표현을 직접 대체하는 태스크 진행
- 혐오 단어에 대해 400개의 혐오-순화 단어 데이터 쌍 구축
- 단순 욕설인 경우 마스킹 처리 예) ㅅㅂ, ㅂㅅ, …

서비스 설계 – (4) 결과

결과 예시

서비스 설계 – (4) 결과

실제 디시인사이드 사이트의 댓글을 순화한 예시 화면

Part 4

서비스 기대 효과

1. 어린이, 청소년 보호

문제 상황
- 어린이와 청소년들이 **온라인의 혐오표현과 자극적인 언어**에 노출되어 있음

서비스 기대효과
- 서비스를 통해 혐오 표현이 필터링되면, 어린이와 청소년들의 **올바른 가치관 형성 가능**

2. 타 사이트로의 확장 가능성

문제 상황
- 여러 온라인 커뮤니티 내에서 의견 공유 시 혐오표현으로 인해 **의견공유에 피로를 느끼는 유저들 발생**

서비스 기대효과
- 타 사이트에 혐오표현 필터링 서비스 제공하여 **온라인 내 정보의 양과 질 개선**, 혐오표현 카테고리 분류를 통해 **자율적으로 순화 여부 조정**

Part 5

개선사항

완벽하지 못한 순화

1. 순화 후에도 부정적인 문맥을 전부 없애진 못함
2. 혐오표현인데 순화하지 못하는 경우 & 혐오표현이 아닌데 변환 되는 경우 발생
3. 표현의 자유 침해 가능성

개선사항

1. 게임 내 채팅, sns 게시물 등 더 다양한 유형의 텍스트 학습 통해 모델 성능 개선 기대
2. 댓글 작성자 연령대, 성별 등 추가 정보 활용하여 추가적인 필터링 제공 가능

감사합니다