2023 AWS, KT AICE와 함께하는 빅데이터·AI경진대회

제목

건전한 댓글문화 형성을 위한 사용자 맞춤형 AI 댓글 순화 서비스

(Session) 명 : KT AICE

(팀) 명 : 바글바글

(팀원) 조수환, 정우성, 김만서,김종한

상명대학교 SW중심대학사업단

인하대학교 SW중심대학사업단

신청일자: 2023. 05. 19

Team

학교 명 : 상명대학교(서울 캠퍼스)

	학과	학번	이름	연락처
팀장	융합전자공학과	201810892	조수환	01092051136 shan6517@naver.com
팀원1	융합전자공학과	201810890	정우성	01077919868 jwoos0705@naver.com
팀원2	휴먼지능정보공학과	201910777	김만서	01091337591 akstj123@naver.com
팀원3	휴먼지능정보공학과	201910792	김종한	01040873461 entere7761@gmail.com

주제

건전한 댓글문화 형성을 위한 사용자 맞춤형 AI 댓글 순화 서비스

1 목적

1. **주제를** 선정한 **배경**

날마다 새로운 컨텐츠가 생성되고, 인터넷 사용자들은 해당 컨텐츠에 다양한 댓글을 달며
서로의 의견을 표출하고 있다.

500 hours / min

6,000 tweets / sec

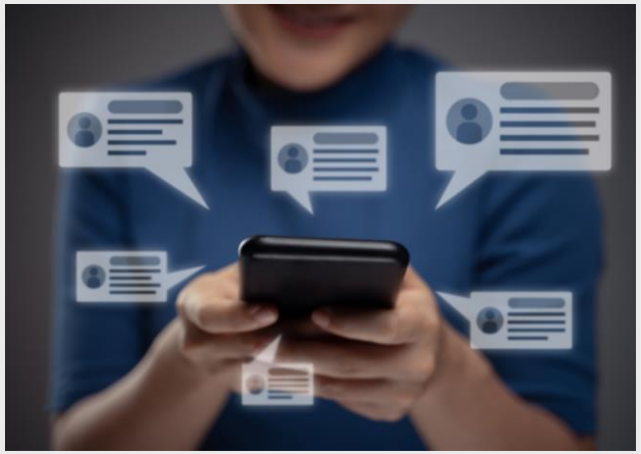
95,000 photos / sec

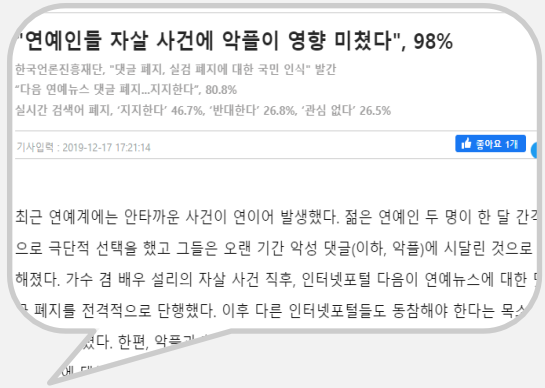
???

1 목적

표현의 자유를 지키는 것도 중요하지만, **악성댓글(악플)**로 인한 **사회적 문제**가 심각해지는
상황 속에서 많은 커뮤니티들은 **댓글검열**이라는 방식을 사용하여 이러한 문제들을 방지하고자 한다.

그러나 **기존의 단순한 검열방식은 사용자에게 만족감을 주지 못하고** 다른 문제들을
야기할 수 있기에 **생성형 AI**를 활용하여 댓글을 **순화**하는 방식을 연구하고자 한다.





2. 왜 생성형 AI (Chat GPT) 로 접근해야 하는가?

1. 문맥 이해 및 의미 해석

생성형 AI는 문맥을 이해하고 의미를 해석하는 능력이 뛰어나기 때문에, 댓글 순화 서비스에서는 문맥에 맞게 댓글을 분석하여 적절한 순화를 수행할 수 있다. 이는 단순히 특정 단어나 문구의 필터링을 넘어서서, **사용자 의도를 파악하고 적절한 수정을 제안할 수 있는 점에서 기존의 댓글 검열 서비스와 차별화 된다**.

2. 개인화 및 유연성

생성형 AI는 사용자와의 대화를 통해 학습하고 개인화될 수 있는 능력을 가지고 있다. 따라서 댓글 순화 서비스는 각 사용자의 취향과 요구에 맞게 작동하여, 더욱 효과적인 댓글 순화를 제공할 수 있다. 이는 기존의 댓글 검열 서비스에서는 찾아보기 힘든 장점으로, **사용자의 필요에 따라 유연하게 작동하는 개인화된 서비스를 제공할 수 있다.**

3. **해결**하고 싶은 **문제**

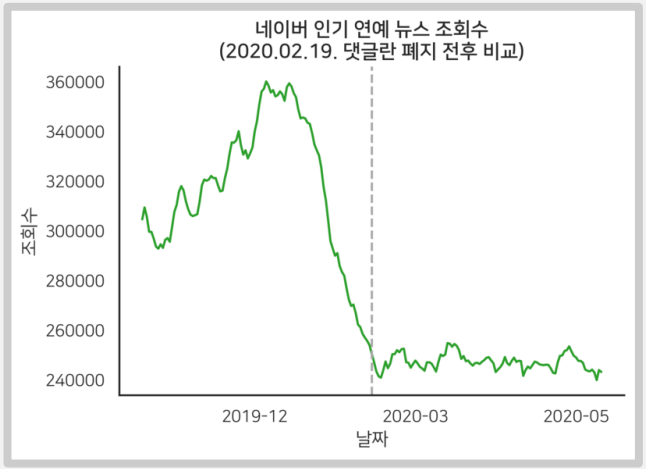
기존 댓글 검열 서비스들은 욕설과 비속어가 포함된 댓글은 삭제하나,

비꼬거나 공격적인 댓글의 검열은 불가능하다.

1 목적

또한 기존의 무차별적인 댓글 검열과 삭제는 문제 해결에 도움이 되지 않는다.

2020.02 네이버 연예 뉴스
댓글란 폐지 전후 비교



1 목적

마지막으로 인터넷 상에서 일어나는 '동조' 행위가 악플 증가에 한 몫 한다.

"기존 댓글들이 공격적이지 않을 때에는 비슷하게
공격적이지 않은 댓글을 달았지만, 기존 댓글이
공격적이면 언어 폭력이 현저히 증가하는 경향을 보였다.

사람들이 기존 댓글들의 톤에 큰 영향을 받아 우르르
악플을 달게 된다는 점을…"

따라서 댓글 순화를 서비스를 통해

- **욕설 및 비속어를 포함하여 공격적이거나 비꼬는 댓글**
- **인터넷 상의 동조 행위**

를 해결하기 위해 노력할 것이며 건전한 댓글 문화를 형성하는 것이 최종적인 목표이다.

[박진영의 사회심리학] 첫 댓글에 동조하는 악플의 심리학 1)

2 분석 방법

Target Users

댓글 컨텐츠를 이용하는 사용자

- 댓글 쓰는 이 측면
- 댓글 보는 이 측면

동일한 댓글이라도 사람마다, 상황마다 느끼는 불쾌함이 다를 수 있기 때문에

↓

댓글의 순화수준을 **단계**별/**대상**별로 사용자가 직접 지정할 수 있도록 선택권을 부여하여 맞춤 서비스를 제공

2 분석 방법

1. Survey & interview

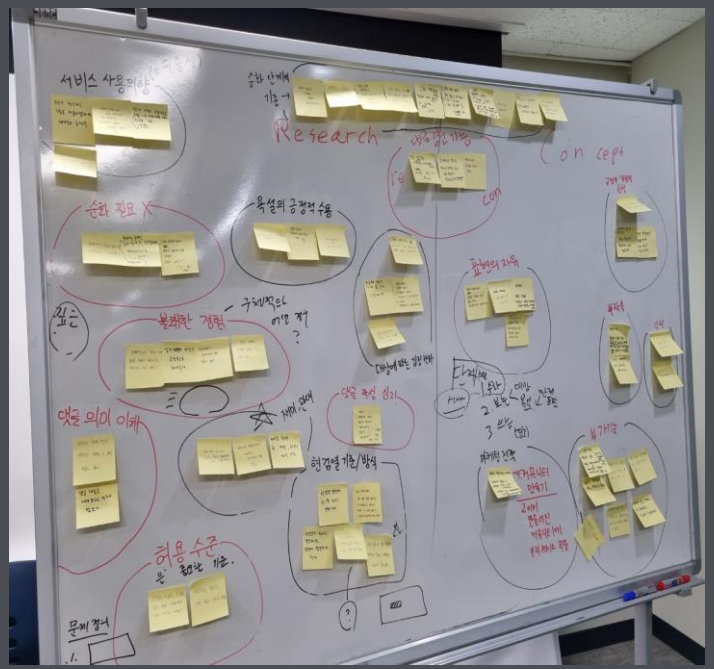
Survey
106

Age
20~29 : 88.7%
30~39 : 5.7%

Interview
11

구글 폼을 이용하여
댓글 관련 서비스에 대한
사용자의 멘탈 모델을 이해하기 위해
설문조사와 인터뷰를 진행

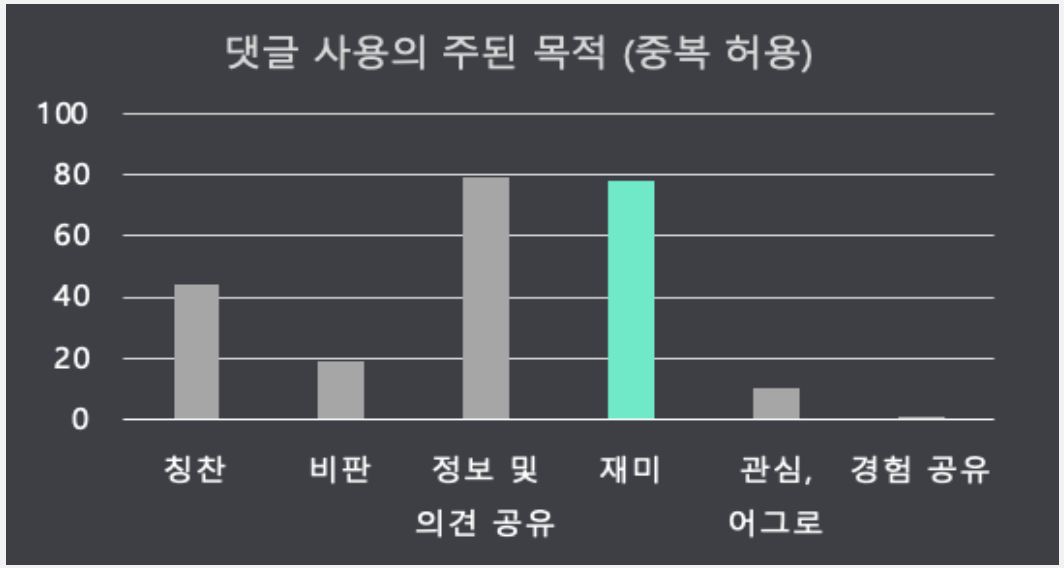
2. Affinity diagram



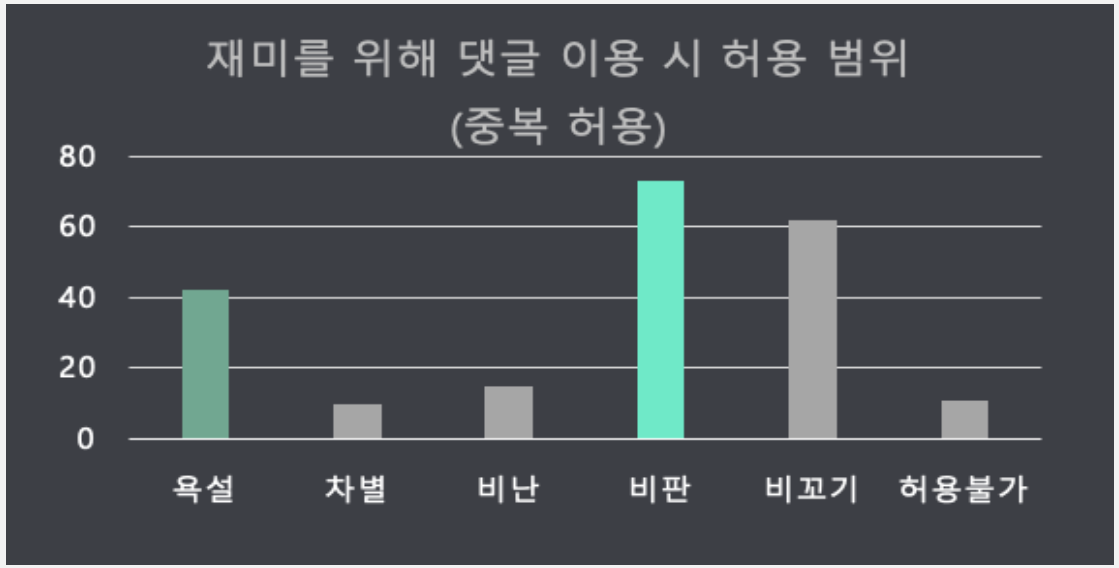
설문조사 결과를 바탕으로 affinity diagram 제작
유사성이 있는 아이디어를 그룹화
아이디어와 컨셉에 대하여 정리
서비스의 최종 기능정의

3. Survey Result

같은 댓글이라도 사용자에 따라 순화가 필요한 단계는 다르다



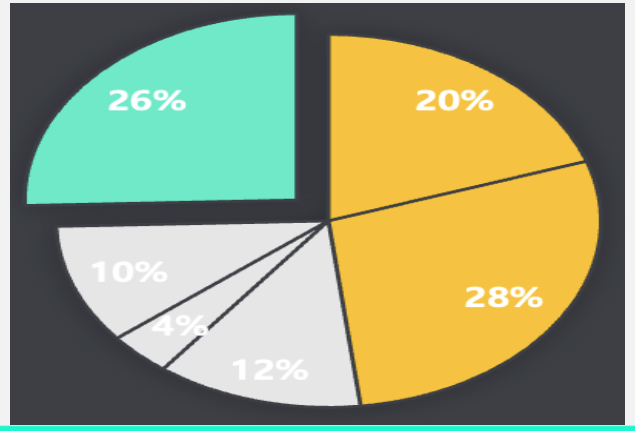
댓글을 사용하는 주된 목적을 조사한 결과 **재미**가 상당 부분 차지한다는 결과를 얻었다.



댓글 이용시 허용 범위에 대해 조사한 결과 **사람마다 그 기준이 다양**하다는 것을 알 수 있었다.

3. Survey Result

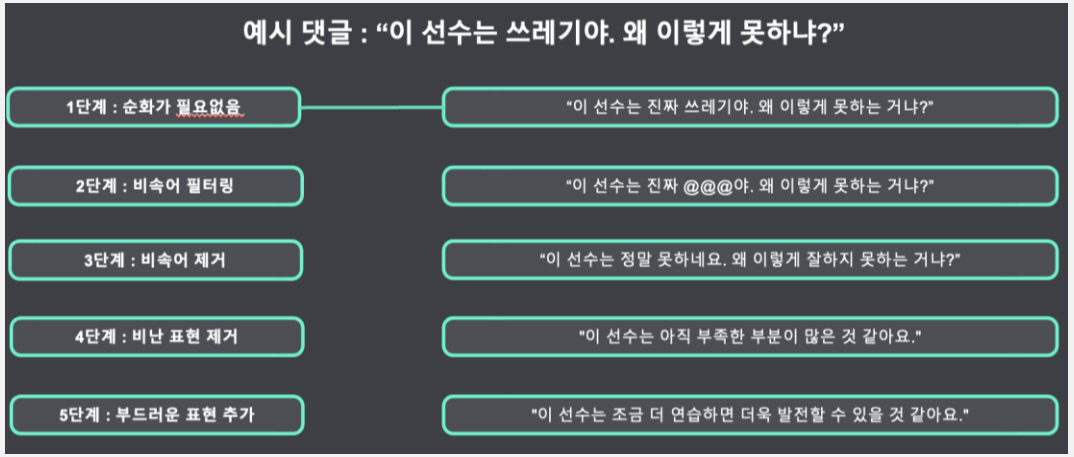
*같은 댓글이라도 사용자에 따라 순화가 필요한 **단계**는 다르다*



A 비난적 표현 순화 **48%**

B 부정적 의도 순화 **26%**

C 순화 필요 없음 **26%**



예시 댓글을 선정하여 임의의 단계로 나누어 순화를 진행했을 때 어떤 단계가 가장 적절한지를 조사해 본 결과 **'순화가 필요 없다' 는 답변이 예상보다 많이 나왔다**.

3. Survey Result

같은 댓글이라도 댓글의 대상에 따라 순화 단계는 달라진다

비난적 표현의 대상에 따른 순화 방식 변화를 조사한결과 대상이 제 3자인 경우 순화단계가 낮지만 **대상이 자신으로 바뀌는 순간 순화단계가 상승**하는 것을 볼 수 있다.

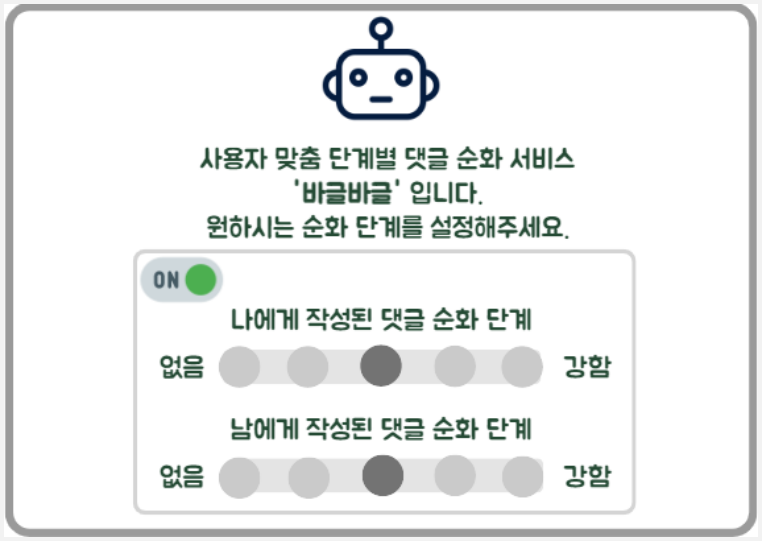
2 분석 방법

4. Core Insight

*같은 댓글이라도 사용자에 따라 순화가 필요한 **단계**는 다르다*

*같은 댓글이라도 댓글의 **대상**에 따라 순화 단계는 달라진다*

5. Conclusion



사용자별 맞춤형 댓글 순화 서비스의 필요성 확인

2 분석 방법

예시 댓글 원문 : “이 선수는 쓰레기야. 왜 이렇게 못하냐?”

사용자가 직접 단계별/ 대상별 댓글 순화단계를 설정하여 맞춤 순화가 가능

“이 선수는 진짜 쓰레기야. 왜 이렇게 못하는 거냐?”

“이 선수는 정말 못하네요. 왜 이렇게 잘하지 못하는 거냐?”

"이 선수는 아직 부족한 부분이 많은 것 같아요."

3 분석 방법

- **Data Definition & Model (Example)**

Description of Dataset Columns

Col A. Content

텍스트 데이터를 나타낸다. 이 데이터는 문장, 단락 또는 문서 등의 형태로 구성될 수 있다. 예를 들어, 이 속성은 고객 리뷰, 뉴스 기사, 소셜 미디어 게시물 등과 같은 텍스트 데이터를 포함한다.

Col B. Valence

(부정) 1-2-3-4-5 (긍정)으로 텍스트 데이터의 긍정적인 정도 또는 부정적인 정도를 나타낸다.

Pretrained Model

GPT-3.5 모델, 또는 다른 언어 생성 모델을 사용해 목적에 맞게 Fine Tuning 고려

3 분석 방법

- Data Acquisition

크롤링

인터넷 커뮤니티에 게시된 댓글의 내용을 얻기 위해 크롤링을 진행하여 데이터 확보

유튜브 커뮤니티, 뉴스기사 댓글, 블로그 댓글 등

공공 데이터

댓글 순화 기능의 성능을 향상시키기 위해 한국말의 의미와 여러 말뭉치 데이터를 확보

모두의 말뭉치 데이터셋, ETRI 감성 대화 데이터셋

데이터 제작

공공데이터로 확보할 수 없는 여러 욕설 말뭉치, 유행어, 줄임말 등을 직접 제작하여 데이터 확보

4 기대효과

댓글을 보는 사용자 측면

1. 악성 댓글에 의한 피해 최소화

2. 댓글의 순기능 극대화

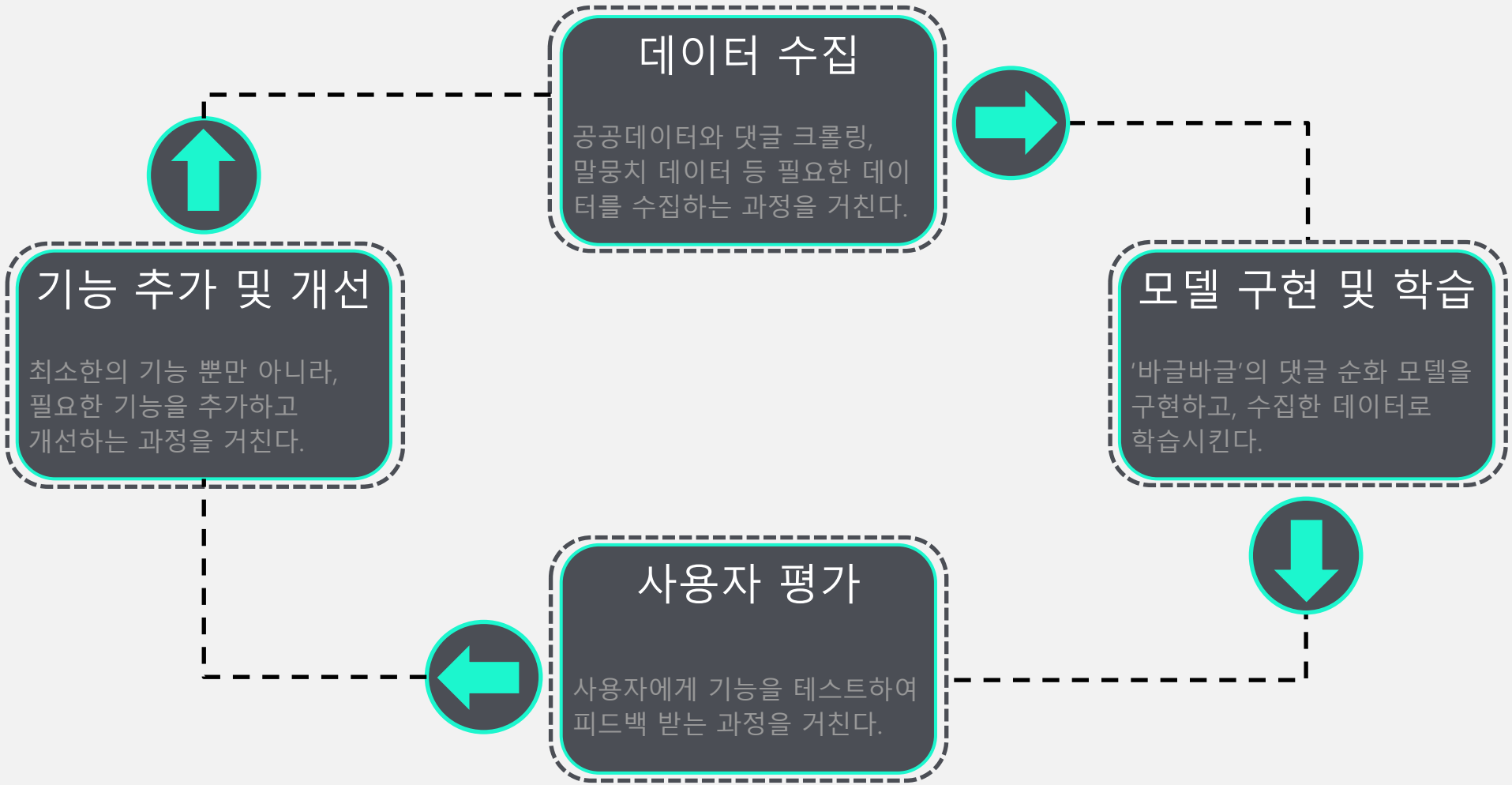
댓글을 쓰는 사용자 측면

1. 악성 댓글 동조효과를 방지하여
악성 댓글 사용률 감소

2. 무분별한 검열에 의해 발생한
사용자의 이탈을 방지하고
결과적으로 컨텐츠 이용률 재고

맞춤 순화서비스를 제공하여 **사용자의 자유**를 **존중**하면서도
적절한 **순화** 기능을 수행하여 **건전한 댓글문화 형성**에 기여

5 Plan


데이터 수집
공공데이터와 댓글 크롤링, 말뭉치 데이터 등 필요한 데이터를 수집하는 과정을 거친다.
모델 구현 및 학습
'바글바글'의 댓글 순화 모델을 구현하고, 수집한 데이터로 학습시킨다.
사용자 평가
사용자에게 기능을 테스트하여 피드백 받는 과정을 거친다.
기능 추가 및 개선
최소한의 기능 뿐만 아니라, 필요한 기능을 추가하고 개선하는 과정을 거친다.

6 기술 지원 요청 (기술 지원 및 SW중심대학사업단)