캡스톤디자인(2) 최종보고서

프렌들리 키보드

Friendly Keyboard

Team members

노현진 20183784

이보림 20190277

정현규 20186984

차례

1. 프로젝트 기획
    1) 팀원 소개
    2) 프로젝트 소개
    3) 개발 동기
    4) 기존 사례 분석
2. 프로젝트 개발
    1) 사용한 기술
    2) 안드로이드 앱 개발 (+앱 스크린샷 포함)
        단계별 교정, 미션 등등 & 각 view별 기능 소개
    3) 비속어 및 혐오표현 동작 로직
3. 시행착오 및 프로젝트 스케줄
    1) 팀 개발 일정

1. 프로젝트 기획

1) 팀원 소개

GitHub https://github.com/CAU-CAPSTONE-2-Friendly-Keyboard

2) 프로젝트 소개

프렌들리 키보드는 사용자의 비속어 및 혐오표현 사용을 줄이고자 언어 습관 교정 어플리케이션이다.

사용자가 스마트폰 내의 어떠한 서비스에서 텍스트를 입력하면, 해당 텍스트 내의 비속어와 혐오표현 사용 유무를 찾아낸다. 이를 바탕으로 단계별 교정을 통하여 자연스러운 교정과 여러 미션을 통하여 사용자가 즐겁게 교정하도록 유도하고자 하였다.

3) 개발 동기

메신저, SNS 를 이용하다 보면 종종 무의식적으로 혹은 습관적으로 비속어나 혐오표현을 사용하고 후회한 경험이 있다. 또한 자주 노출된 혐오표현은 혐오표현이라고 인지하지 못하고 사용했던 경험이 있다. 이와 같은 경우가 사회적으로 얼마나 발생하고 있는지 알아보았다.

① 혐오표현

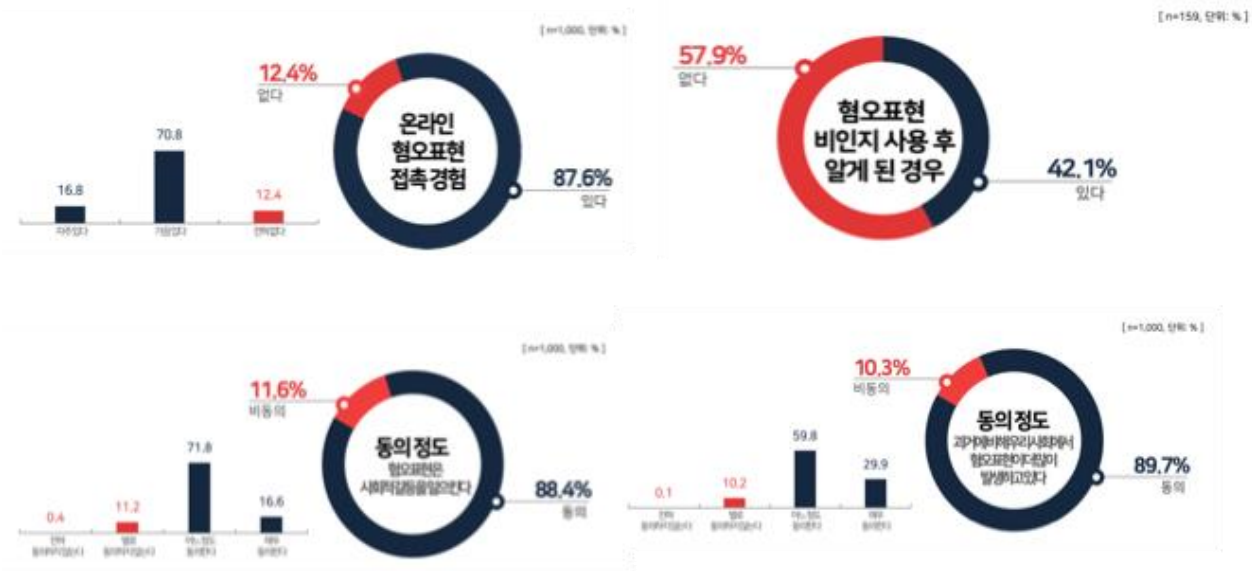


Figure 1 2021 서울시 청년 인권인식 및 혐오표현 실태조사

'2021 년도 서울시 청년 인권인식 및 혐오표현 실태조사'에 따르면, 응답자 중 무려 90%에 가까운 사람들이 온라인에서 혐오표현을 접한다고 응답하였다. 이를 통해 온라인에서 빈번하게 혐오표현이 사용되고 있음을 알 수 있다. 온라인에서의 혐오표현 사용뿐만 아니라 빈번한 접촉 또한 혐오표현 비인지 사용으로 이어지며 실제 응답자 중 42%가 혐오표현 비인지 사용 후 알게 된 경우가 있다고 응답하였다.

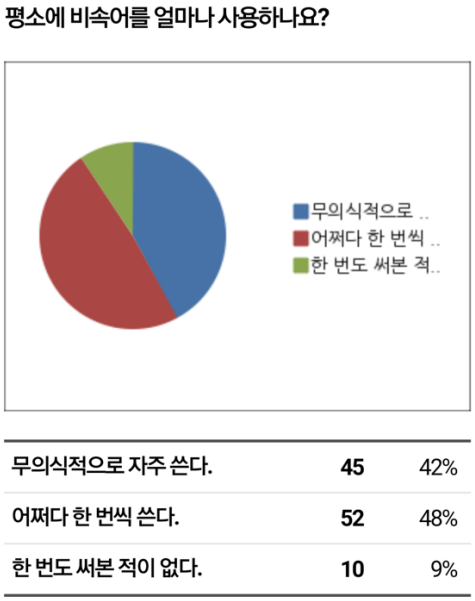
혐오표현 사용이 심각한 문제가 될 수 있는 이유는 개인으로 끝나는 것이 아닌 사회적 문제로 발전할 수 있기 때문이다. 약 90%의 대다수 응답자들이 혐오표현은 사회적 갈등을 조장할 위험이 있다고 답하였을 뿐만 아니라 과거에 비해 우리사회에서 혐오표현이 더 많이 발생하고 있다고 답변하였다.

이와 같은 결과를 바탕으로 혐오표현을 줄이기 위한 노력이 필요한데, 적지 않은 사람들이 혐오표현을 인지하지 못하는 상황에서 사용하기에

단순히 혐오표현을 줄여야 한다고 알려주는 것은 효과적으로 혐오표현 사용을 줄이지 못한다.

그래서 스마트폰 키보드에서 혐오표현 사용을 알리고 무슨 혐오표현에 해당하는 것인지를 제공함으로써 해결할 수 있을 거라 생각하여 프렌들리 키보드를 개발하게 되었다.

② 비속어



무의식적으로 자주 쓴다.	45	42%
어쩌다 한 번씩 쓴다.	52	48%
한 번도 써본 적이 없다.	10	9%

Figure 2 2020 국립국어원 언어의식 조사

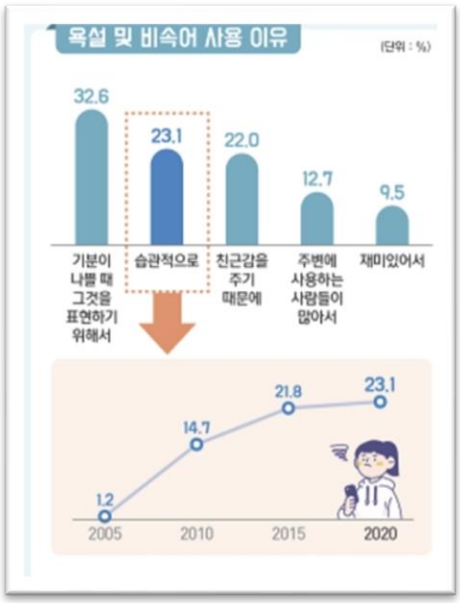


Figure 3 2019 대한민국 청소년 기자단

'2019 대한민국 청소년 기자단'에서 조사한 결과에 따르면 90%의 청소년들이 비속어를 사용한다고 응답하였다. 이 중 42%가 무의식적으로 자주 사용한다고 한다. 이러한 비속어 사용은 해마다 점차 늘어나는 추세이다. 2005 년도 불과 1.2%였던 것이 2020 년도에는 23.1%로 대폭 상승하였다.

습관적인 무분별한 비속어 사용을 줄이기 위해서는 비속어 사용시 알림을 보내 무의식에서 의식으로 바꿔주는 것이 필요하다. 어플리케이션으로 사용자에게 얼마만큼의 비속어 사용을 하고 있는 지와 비속어 텍스트 입력 시 알림, 대체어 제공으로 해결하고자 프렌들리 키보드를 개발하게 되었다.

4) 기존 사례 분석

① 바른말 키패드

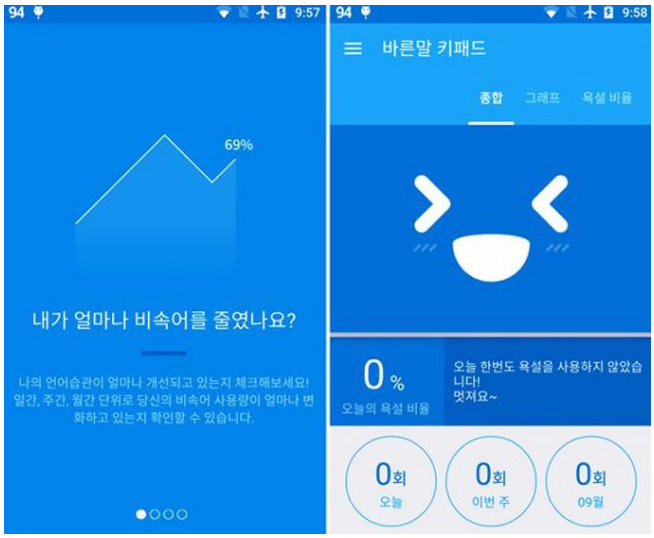


Figure 4 바른말 키패드

위의 앱은 바른말 키패드로 비속어 사용 횟수를 기록하거나 알람으로 경고, 비속어 사용 시 이모티콘으로 대체해주는 기능을 갖고 있다. 현재 해당 앱은 서비스 종료된 상태이다.

② 네이버 스마트보드

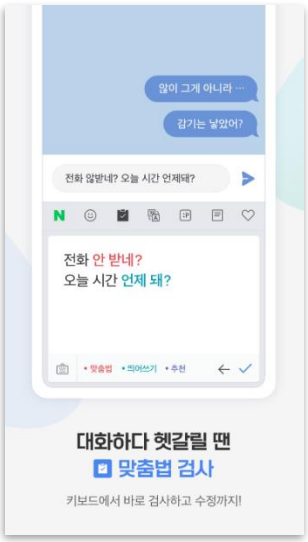


Figure 5 네이버 스마트보드

위의 네이버 스마트보드 앱은 채팅을 할 때, 키보드에서 맞춤법 검사를 할 수 있는 기능을 제공하여 잘못된 맞춤법으로 채팅하는 습관을 개선시켜준다.

③ 차별점

- 바른말 키패드와의 차별점

  본 프로젝트('프렌들리 키보드')는 바른말 키패드가 제공해주는 비속어 알람과 대체 기능뿐만 아니라 혐오표현 기능도 함께 제공한다. 바른말 키패드의 문제점은 비속어 알람 및 대체 기능이 사용자의 불편함을 계속 유발한다는 점이라고 판단하였다.

  프렌들리 키보드에서는 챗 GPT 와의 대화, 업앤다운 게임, 자판이 뒤섞인 랜덤 키보드 등 다양한 교정 방법과 미션으로 재미를 주어 문제점을 개선하였다.

- 네이버 스마트보드와의 차별점

  네이버 스마트보드와 같은 키보드로 문제점을 개선시켜주는 점이 같지만 프렌들리 키보드는 맞춤법이 아닌 비속어와 혐오표현 사용을 개선시켜준다. 그리고 프렌들리 키보드에서는 사용 통계 또한 확인할 수 있다.

2. 프로젝트 개발

1) 사용한 기술

① Android Studio

안드로이드 개발 환경에서 가장 널리 쓰이는 Kotlin 언어를 사용하여 안드로이드 OS 의 기반의 서비스로 제작.

터치 스크린을 통해 입력을 수행하기 위한 가상 키보드 SoftKeyboard 제작. InputMethodManager 클래스를 통해 본 custom keyboard 를 사용자 키보드로 등록 및 사용.

② Kotlin Coroutine

비속어 제거를 위한 마스킹 기능이 적용된 후 문자 전송이 이루어지도록 비동기 방식의 텍스트 마스킹 기능을 동기적으로 수행.

③ Retrofit2

서버와 통신을 위해 사용한 라이브러리

④ **ChatGPT**

교정 기능 해제를 위한 미션 수행 과정에서 ChatGPT 의 feedback 활용.

⑤ **Flask**

Python 의 라이브러리 중 하나로 API 서버 구축에 활용.

⑥ **MariaDB**

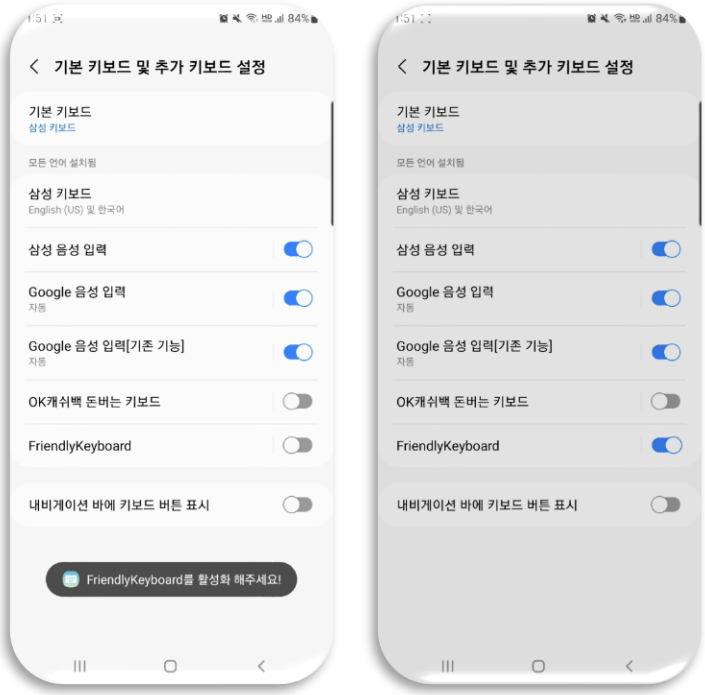
계정당 혐오표현 전체 사용 횟수, 날짜별 혐오표현 종류별로 사용 횟수, 자신이 입력한 채팅방 내용을 저장하는데 사용한 DB.

⑦ **Raspberry Pi**

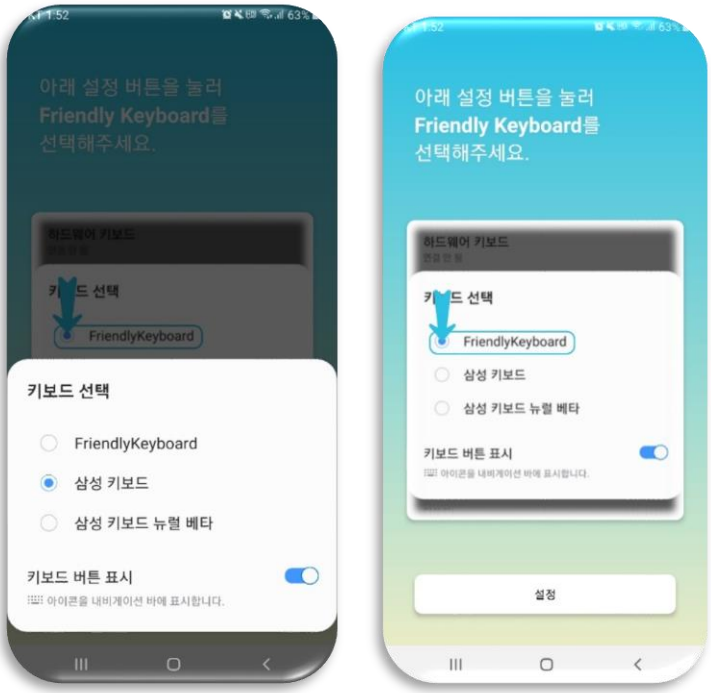
Python 의 Flask 라이브러리를 사용하여 구축한 서버를 배포하는데 사용.

2) 안드로이드 앱 개발

① 키보드 선택



앱 사용을 위해 본 서비스의 SoftKeyboard 를 기본 키보드로 설정하여야만 실행 가능.

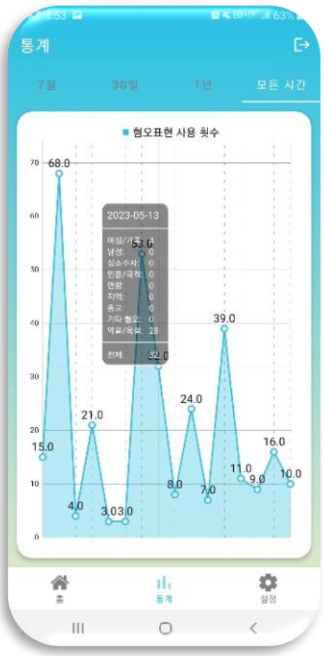


앱에서 로그인을 한 직후 현재 사용하고 있는 키보드가 Friendly Keyboard가 아닌 경우 키보드 선택 설정창을 통해 사용하는 키보드를 Friendly Keyboard로 변경합니다.

② 홈 화면

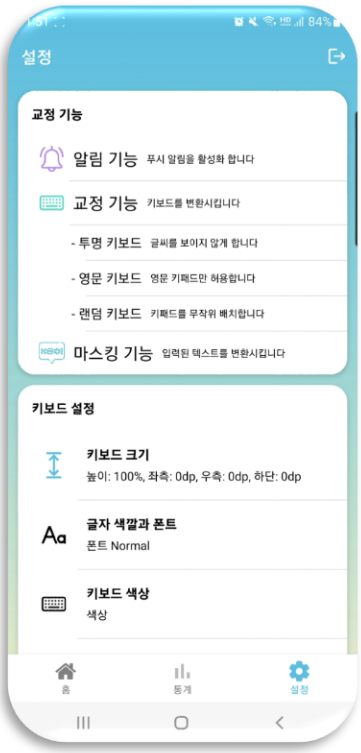
홈 화면에서는 현재 적용 중인 교정 기능과 자신이 지금까지 사용한 혐오표현 전체 사용 횟수를 표시합니다.

③ 통계 화면



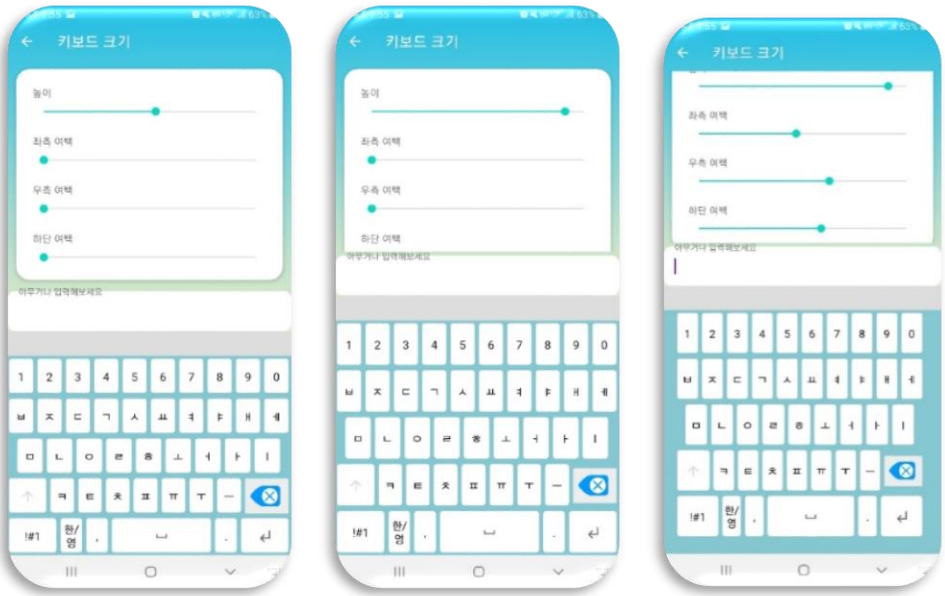
통계 화면에서는 자신이 날짜별로 사용한 혐오표현 횟수를 종류별로 분류하여 보여줍니다. 또한 통계를 볼 때 7일, 30일, 1년, 또는 모든 시간으로 기간을 설정하여 데이터를 볼 수 있습니다.

④ 설정 화면



설정 화면에서는 키보드의 크기나 색상 등을 수정하여 자신만의 키보드를 커스터마이징할 수 있습니다.

■ 키보드 크기



위의 사진과 같이 키보드의 높이를 조절하거나 좌우, 하단 여백을 수정할 수 있습니다.

■ 키보드 글자

위의 사진과 같이 키보드 글자의 색깔을 수정하거나 볼드체 여부를 설정할 수 있습니다.

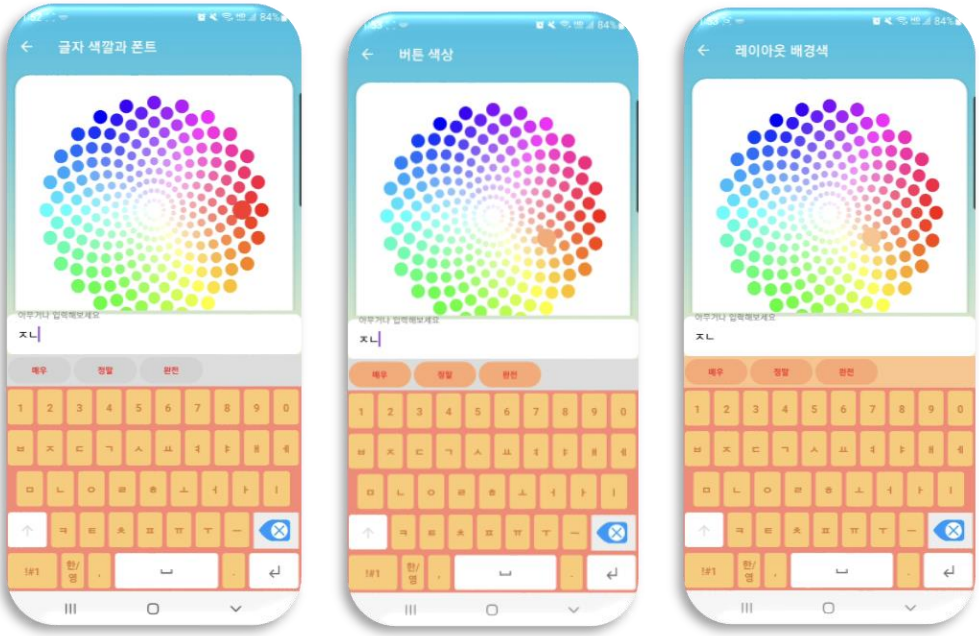
■ 키보드 색상

위의 사진과 같이 키보드 자판의 색상을 수정할 수 있습니다.

■ 키보드 배경

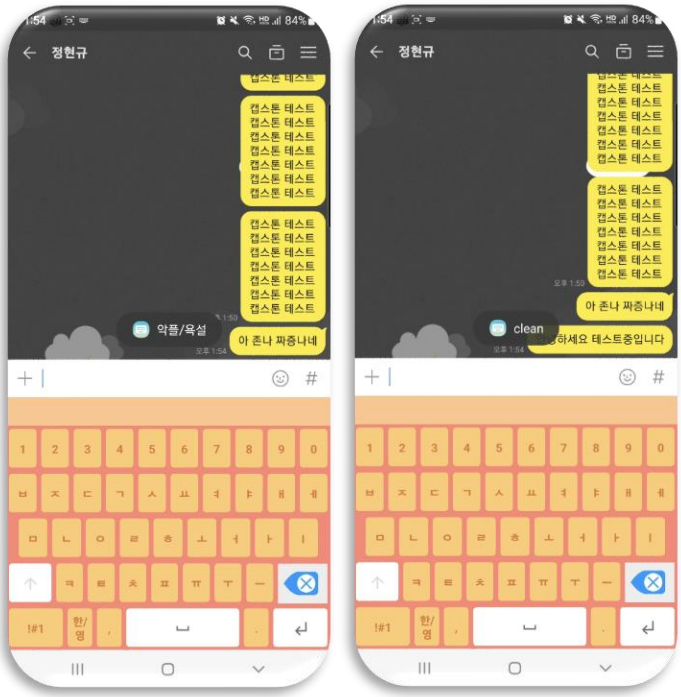
위의 사진과 같이 키보드의 배경색을 수정할 수 있습니다.

■ 대체어



colorPicker library를 이용해 비속어 입력 시 제공되는 대체어의 UI RGB값을 직접 수정할 수 있도록 합니다.

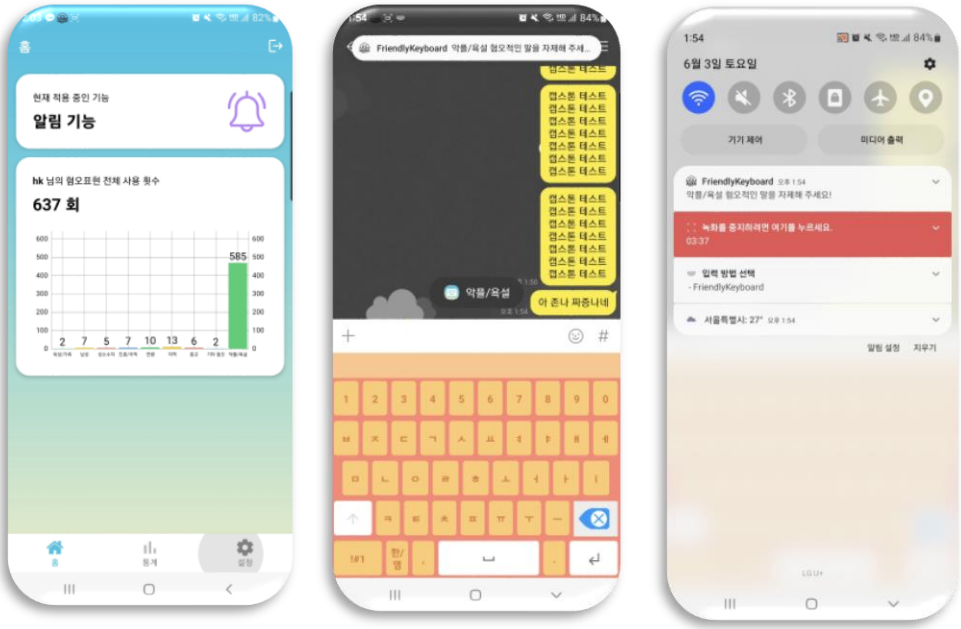
⑤ 텍스트 내의 혐오 표현 유무 판별



메신저 어플에서 본 키보드를 사용하여 텍스트를 전송을 할 경우 먼저 서버로 전송된 텍스트의 혐오 표현 유무가 AI 모델에 의해 판별됩니다. 이후 결과 값을 토대로 유저의 혐오 표현 사용 횟수 및 단계별 교정 기능 실행을 결정합니다.

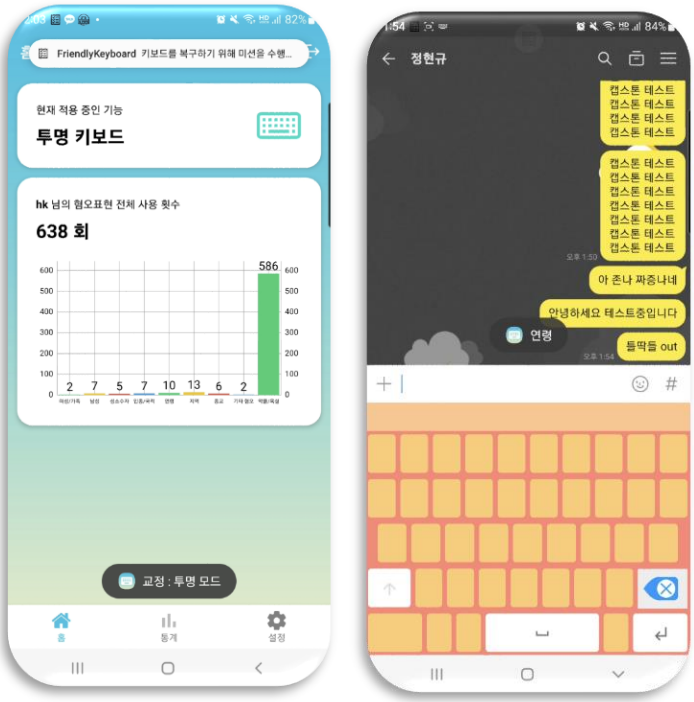
⑥ 교정 기능

■ 알림 기능



알림 기능은 단계별 교정 1단계로서 전송한 메시지의 혐오 표현 포함 유무를 판단해 존재 시 세부 종류를 알림으로 알려줍니다. NotificationCompat 클래스를 활용하여 위 단계에서 푸시 알림이 항상 팝업되도록 하였습니다.

■ 투명 키보드

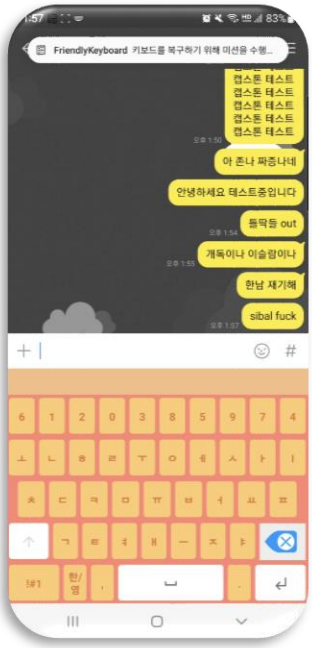


투명 키보드는 단계별 교정 2단계인 키보드 변환의 첫 기능으로서 키보드 글자를 동적으로 바탕색과 일치시켜 글자의 위치를 보이지 않게 합니다.

■ 영문 키보드

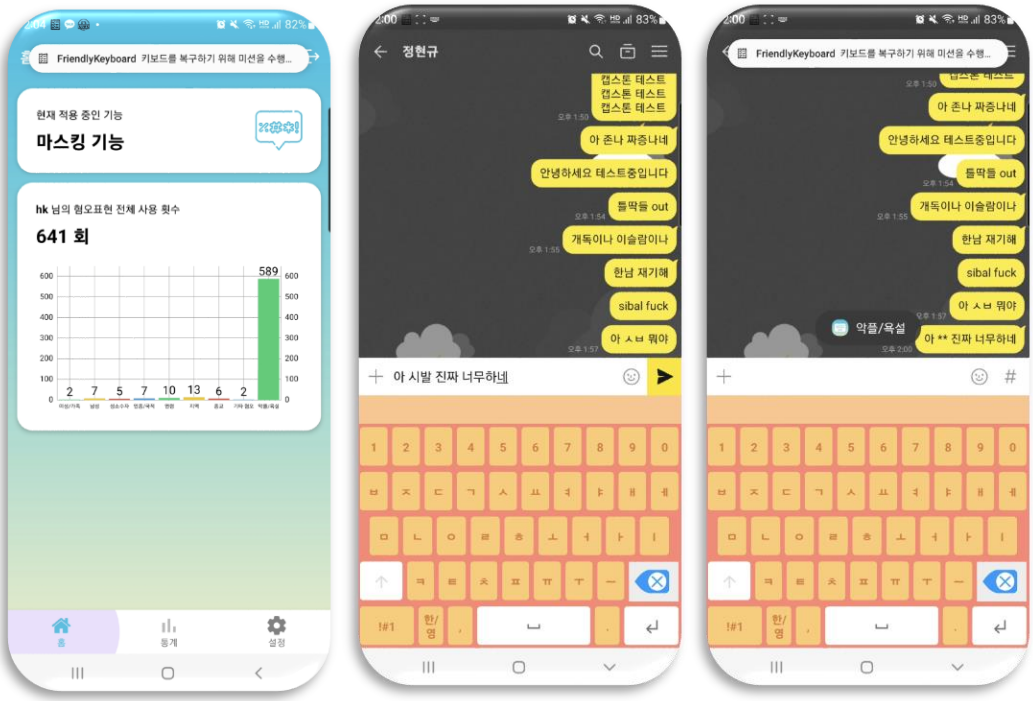
영문 키보드는 2단계의 두 번째 기능으로서 한글 및 특수문자 키보드로의 변환을 금지시켜 오직 영문 키보드의 사용만을 허용합니다.

■ 랜덤 키보드



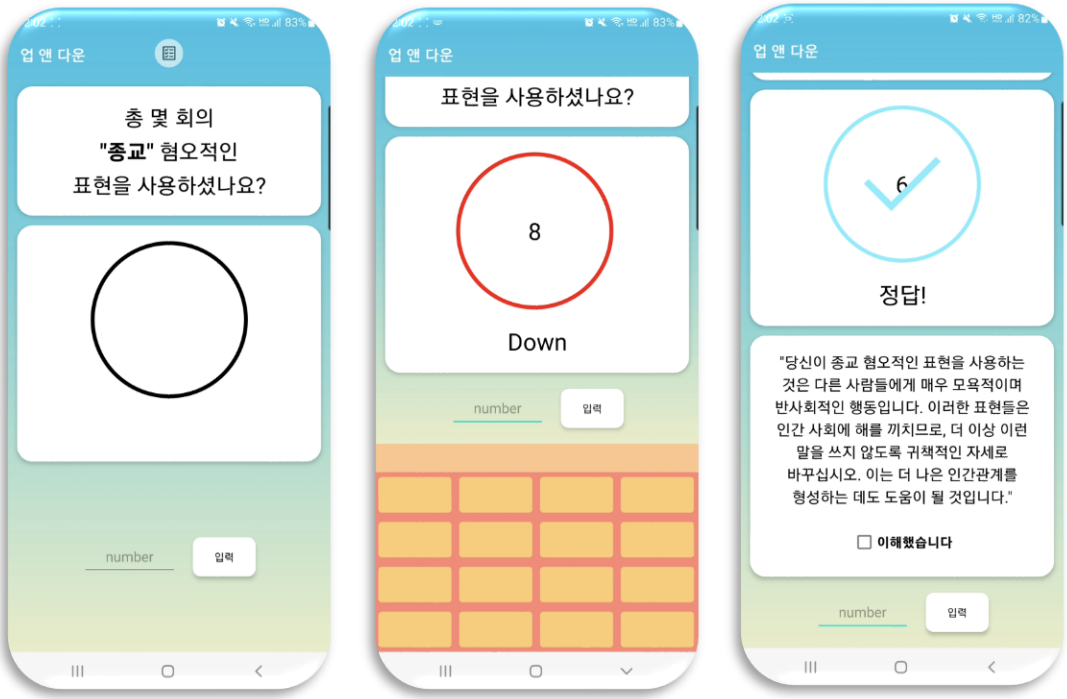
랜덤 키보드는 2단계의 세 번째 기능으로서 해당 기능 활성화 중에는 채팅 전송 시 매번 키보드 내 글자를 무작위로 배치합니다.

■ 마스킹 기능



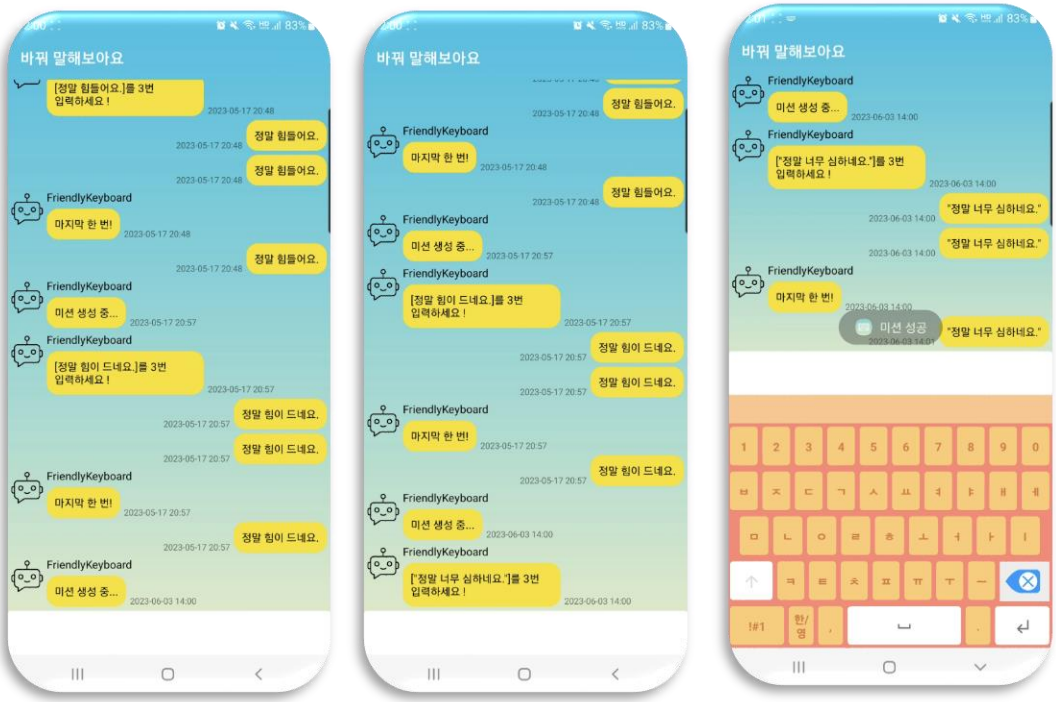
단계별 교정의 마지막 단계인 마스킹 기능은 메시지 전송이 완료되기 직전에 수행해야 하는 작업으로서 문장 내 비속어만을 대상으로 마스킹을 실행합니다. 여타 교정 기능들과 다르게 텍스트가 마스킹 완료 후 전송되어야 하기 때문에 서버에서 비동기적으로 이루어지던 작업을 coroutine의 suspend 및 withContext를 활용해 동기적으로 바꾸어 실행합니다.

⑦ 업 앤 다운 기능



본 키보드 서비스는 유저가 활성화된 단계별 교정 기능을 해제하고 자신이 사용했던 혐오 표현에 대해 되돌아볼 수 있도록 총 두 가지의 미션 기능을 제공합니다. 첫 번째는 업 앤 다운 게임으로서 유저가 사용했던 특정 종류의 혐오 표현 횟수를 맞춘다면 사전에 입력해둔 prompt로 API를 사용해 ChatGPT가 해당 언행에 대한 feedback을 주고 이를 읽은 유저는 미션을 클리어합니다.

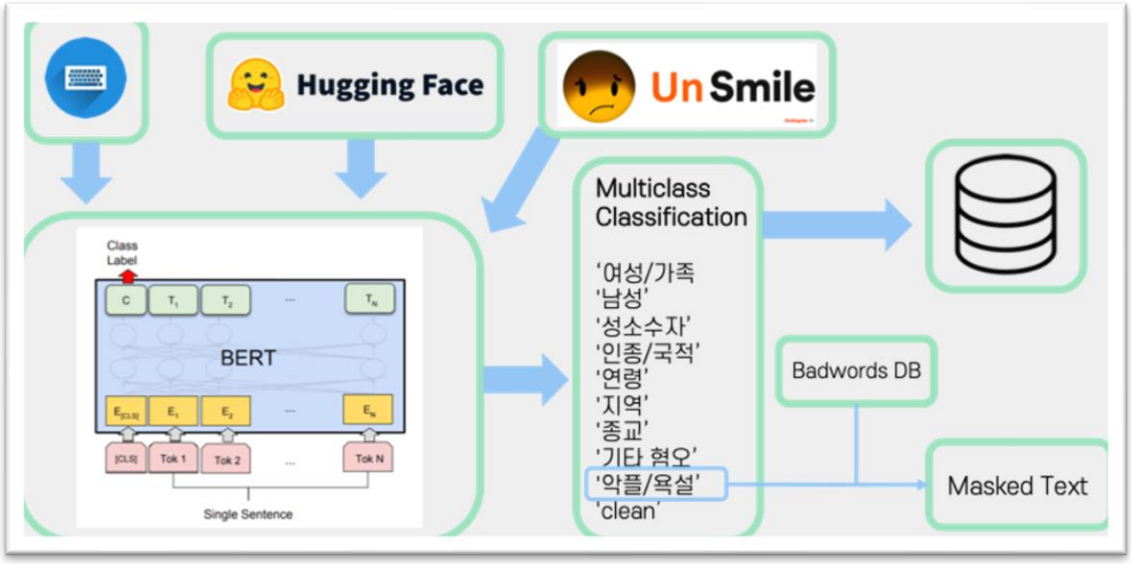
⑧ 채팅방 기능



두 번째는 유저가 사용했던 비속어 및 혐오 표현을 토대로 시스템으로부터 순화된 표현의 문장을 받아와 유저가 이를 특정 횟수만큼 입력해야 클리어 가능한 미션입니다. 마찬가지로 ChatGPT API를 활용하였습니다.

3) 비속어 및 혐오표현 동작 로직

- **대체어 이외의 설계도**



① 텍스트 분류

사용자가 입력하는 텍스트에 비속어, 혐오표현이 있는 지 확인하기 위하여 텍스트로 Multiclass classification 을 진행한다.

Pretrained Model 로는 hugging face 에서 한국어로 학습된 bert model(hugging face 의 beomi/kcbert-base)을 사용한다. Bert 모델은 트랜스포머를 이용하여 전체적인 문장을 모두 반영하여 문장 내용을 잘 이해하기 때문에 선택하였으며, 서버가 크지 않기 때문에 large 보다는 base 모델을 선택했다.

데이터셋은 **Smilegate AI** 의 **'한국어 혐오표현 Unsmile 데이터셋'**을 사용한다. 해당 데이터셋은 혐오표현을 '특정 사회적 (소수자) 집단에 대한 적대적 발언, 조롱, 희화화, 편견을 재생산하는 표현'으로 정의하고 전문가 집단을 통하여 10 가지[여성/가족, 남성, 성소수자, 인종/국적, 연령, 지역,

종교, 기타혐오, 악플/욕설, Clean]의 레이블로 데이터를 분류하였다. 각 레이블에 대한 설명은 아래와 같다.

- 여성/가족 : 여성성 및 여성의 성역할에 대한 통념을 고착시키는 발언, 여성 차별을 희화화하는 발언, 페미니즘·여성가족부 전반에 대한 악플 등을 포함, 비혼주의자, 미혼모, 동성 부부 등 전통적이지 않은 형식의 가족에 대한 혐오 발언 역시 본 카테고리에 포함
- 남성 : 집단으로서의 남성 일반을 비하, 조롱, 희화화하는 발언들
- 성소수자 : 성소수자(레즈비언, 게이, 바이섹슈얼, 트랜스젠더 등)를 배척하는 발언, 이성애 이외의 섹슈얼리티를 부정적으로 묘사하거나 성소수자를 희화화하는 표현들을 포함
- 인종/국적 : 특정 인종(흑인, 아시안 등)과 국적(일본인, 아프가니스탄인, 베트남인 등)에 대한 욕설, 고정관념, 조롱 포함, 종교·인종·국가에 대해 암묵적으로 함께 지칭하는 소재 (e.g. 무슬림, 난민)의 발언들 포함
- 연령 : 특정 세대나 연령을 비하하는 은어의 사용 및 혐오 표현
- 지역 : 특정 지역에 대한 은어 및 혐오 표현
- 종교 : 특정 종교에 대한 혐오 및 종교인 집단에 대한 비난
- 기타혐오 : 위에서 정의한 카테고리 이외의 집단을 대상으로 하는 혐오 표현(e.g. 장애인, 정부, 기자, 경찰, 차별금지법 반대 등)
- 악플/욕설 : 타인 혹은 외모에 대한 비하/욕설이 포함되어 있거나, 불쾌감을 주거나, 악플과 음란성 문장을 분류
- Clean : 혐오표현, 욕설, 불쾌감, 음란성 내용을 포함하고 있지 않은 일반 문장

모델의 결과로 레이블이 나오면, 해당 내용을 사용자에게 알린다. 그리고 이 내용을 저장하여 사용자의 사용기록을 업데이트한다.

② 악플/욕설 마스킹

단계별 교정 중 3 단계에 속하는 문장 마스킹은 사용자가 입력한 문장을 전송할 때 비속어를 ***로 자동 변환하는 기능이다. 3 단계일 때, 사용자가 입력한 텍스트가 '악플/욕설'로 분류가 된다면 비속어만 저장되어있는 Badwords DB 에서 포함되는 비속어를 찾아 마스킹한다.

원래 계획된 방식은 비속어 위치를 AI 모델로 찾는 방법이었는데, LIME 알고리즘으로 구현하였지만, 최대한 시간단축을 시도한 결과 1 분 미만으로는 해결되지 않아 사용에 적합하지 않다고 판단하였다. 그래서 테스트시간이 짧은 DB 내에서 찾아 매칭하는 위와 같은 방법으로 해결하였다.

③ 비속어 대체어

키보드 텍스트 입력 시, 비속어 입력 시 곧바로 대체어를 제안하는 기능이다. Key : Value 형식으로 비속어 : 대체어 사전을 직접 구축하여 매칭 후 대체어를 키보드 자판 바로 위에 띄워 클릭 시 대체된다. 대체어 사전은 챗 GPT API 를 이용하여 'ooo 를 대체할 단어 추천해줘'의 답변들을 참고하여 구축하였다.

3. 프로젝트 스케줄

Project Schedule

Week	Course Schedule	Team Schedule & Individual Schedule
1	Orientation / Team Organization	공통: 2023 캡스톤디자인(2) 프로젝트 주제 선정, Proposal 작성, PPT 작성
2	**Proposal /** **1st Professor Feedback**	공통: 교수님 피드백 반영, 디자인 구상 및 사전조사, 노현진: 회원가입 화면 및 기능 구현 정현규: 로그인 화면 및 기능 구현
3	1st Company-Mentor Feedback	공통: 멘토링 피드백 반영 노현진: 설정 화면 및 기능 구현 정현규: 키보드 xml 및 기능 구현 이보림: 비속어 마스킹 모델 구현
4	1st Additional Professor Feedback	노현진, 정현규: 이전 기능 보완 이보림: 혐오표현 데이터셋
5	**2nd Professor Feedback**	노현진, 정현규: 포그라운드로 키보드 앱 동작 여부 확인 이보림: 혐오표현 모델 학습
6	2nd Company-Mentor Feedback	노현진, 정현규: 키보드 설정 기능 연결 이보림: 혐오표현 모델 학습
7	2nd Additional Professor Feedback	노현진, 정현규: 기록화면 구현 이보림: 비속어 대체어 사전 구축

8	Midterm period (No Feedback Session)	공통 : 중간고사 준비
9	**3rd Professor Feedback**	노현진, 정현규: 단계별 기능 구현 이보림: 대체어 사전 완성
10	3rd Company-Mentor Feedback	노현진, 정현규: 부족한 기능 구현 이보림: 비속어 마스킹 모델 수정
11	3rd Additional Professor Feedback	공통: 중간 발표 준비 노현진, 정현규: 기록 화면 구현 완성 이보림: 비속어 마스킹 모델 시간 개선
12	**Interim Project Demonstration**	공통 : 중간 Demo 준비
13	4th Company-Mentor Feedback	공통: 세부 기능 수정 이보림 : 비속어 마스킹 로직 변경
14	4th Additional Professor Feedback	공통: Final Report 작성
15	**Final Project Demonstration**	공통: 최종 Demo 준비
16	Submission for the Project Outcomes (+ Peer Review)	공통: 최종 documents 제출