JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 3, pp. 1-8, March 2025
https://doi.org/10.9708/jksci.2025.30.03.001

A Design and Implementation of Youth Profanity Prevention Application Based on LLM and Generative AI

Ho-joon Kim*, Hyun-dong Kim**, Seo-hee Son***, Sung-uk Bae****,
Ji-Won Ock*****, Sejong Lee†

*Student, Dept. of Computer Engineering, Seoul National University of Science and Technology, Seoul, Korea
**Student, Dept. of Computer Engineering, Hankuk University of Foreign Studies, Gyeonggi, Korea
***Student, Dept. of Computer Engineering, Ewha Womans University, Seoul, Korea
****Student, Dept. of Computer Engineering, Konkuk University, Chungbuk, Korea
*****Researcher, Agency for Defense Development (ADD), Seoul, Korea
†Asistant Professor, School of Computer Science and Engineering, Yeungnam University, Gyeongsan, Korea

[Abstract]

In this paper, we propose an AI-based service called Baleunmalssami aimed at fostering a positive language culture among adolescents and preventing cyberbullying. This service utilizes a classification model based on LLM and RAG to accurately detect offensive language and provide real-time suggestions for alternative expressions or emojis suited to the context, thus protecting the privacy of minors and reducing the use of profanity. The service consists of a keypad and an app. The keypad replaces offensive language with emojis that match the tone and provides a real-time risk rating for unethical expressions. The app uses accumulated text data analyzed by the LLM to automatically generate reports on language habits and cyberbullying for both students and parents. The KoSim-CSE-BERT-multitask model used in the keypad delivers fast and accurate results on-device without exposing data, while reports generated with LLM and RAG include only select portions of actual conversations to protect the privacy of minors. As a result, this service provides students with real-time language correction and self-awareness opportunities, while offering parents insightful information about their children's language habits. Baleunmalssami will play an important role in fostering healthy language habits and contributing to a safer cyber environment.

▸ **Key words**: Generative AI, Cyberbullying, Verbal abuse, LLM, RAG

• First Author: Ho-joon Kim, Corresponding Author: Sejong Lee
*Ho-joon Kim (hojoon00905@gmail.com), Dept. of Computer Engineering, Seoul National University of Science and Technology
**Hyun-dong Kim (ksglsg1350@gmail.com), Dept. of Computer Engineering, Hankuk University of Foreign Studies
***Seo-hee Son (ligerhearts@gmail.com), Dept. of Computer Engineering, Ewha Womans University
****Sung-uk Bae (qorlwk324@gmail.com), Dept. of Computer Engineering, Konkuk University
*****Ji-Won Ock (jwock@add.re.kr), Agency for Defense Development (ADD)
†Sejong Lee (kingsaejong@yu.ac.kr), School of Computer Science and Engineering, Yeungnam University
• Received: 2024. 11. 26, Revised: 2025. 01. 13, Accepted: 2025. 01. 13.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 논문에서는 청소년의 건전한 언어문화 형성과 사이버 폭력 예방을 위한 AI 기반의 바른말ㅆ미 서비스를 제안한다. 이 서비스는 LLM(Large Language Model) 기반의 분류 모델과 RAG(Retrieval Augmented Generation)를 활용하여 비속어를 정확히 감지하고, 상황에 적합한 대체 표현이나 이모티콘을 실시간으로 제공함으로써 자녀의 사생활을 보호하고 비속어 사용을 줄일 수 있도록 한다. 이 서비스는 키패드와 앱으로 구성한다. 키패드는 비속어를 감정에 맞는 이모지로 대체하고, 비윤리적 표현의 위험등급과 올바른 표현 제안 기능을 실시간으로 제공한다. 앱은 LLM으로 누적된 텍스트를 분석하여 언어 습관과 사이버 폭력에 관한 리포트를 자동 생성하여 학생과 부모에게 각각 전달한다. 키패드에 사용한 KoSim-CSE-BERT-multitask 모델은 온디바이스에서 데이터 노출없이 빠르고 정확한 결과를 내고, LLM과 RAG를 활용한 리포트에는 부모에게 자녀의 실제 대화 내용의 일부만을 제공한다. 그 결과 학생들에게는 실시간 언어 교정과 자기 인식 기회가 주어지고, 부모는 자녀의 언어 습관에 대한 통찰력 있는 정보를 얻는다. 바른말ㅆ미 서비스는 올바른 언어 습관을 형성하고, 건전한 사이버 환경 조성에 중요한 역할을 할 것이다.

▸ **주제어**: 생성형 AI, 사이버폭력, 비속어, LLM, RAG

I. Introduction

최근 청소년의 스마트폰 보유율이 지속적으로 증가함에 따라 카카오톡, SNS 같은 사이버 공간에서 청소년들이 비속어와 자극적인 표현을 사용하는 빈도수도 함께 증가하고 있다[1]. 이러한 현상은 교육 현장에서 학교 폭력의 큰 요인 중에 하나이고, 비속어를 접하는 청소년의 연령이 점점 낮아지는 주요 원인이 되고 있다. 특히, 2021년 원주시와 2024년 국립국어원의 조사 결과에서 초등학생의 비속어 사용 비율이 79%에서 90%로 큰 폭으로 증가하였음을 볼 수 있다[2,3].

교육부의 2023년 학교폭력 실태조사에서 학교폭력 피해 응답률이 초등학생 3.9%, 중학생 1.3%, 고등학생 0.4%로 나타났다. 이는 초등학생의 피해 응답률이 중학생의 3배, 고등학생의 약 9.75배에 달하는 수준이다[4]. 특히 학교폭력 중 사이버 폭력의 형태가 많아지고, 그중 언어폭력이 37.1%로 가장 높은 비율을 차지한다. 즉, 사이버 공간에서 무분별한 비속어 사용이 많아지고 있음을 의미한다[4]. 이러한 비속어 사용에 따른 사이버 및 학교 폭력이 증가하는 추세이다. 또한 초등학생의 정서 발달과 자아 존중감에 심각한 악영향을 미칠 수 있다.

디지털 기기 사용으로 가족 간 대화가 감소함에 따라 부모가 자녀의 언어습관 교육에 어려움을 겪고 있는 상황이다. 이는 부모가 자녀의 온라인 활동을 직접 모니터링하기 어렵고, 디지털 환경에서 언어 사용에 대한 이해가 부족한 경우가 많기 때문이다. 따라서 언어습관이 성되는 초등학생을 대상으로 사이버 공간에서의 언어폭력 방지를 위한 효과적인 대책 마련이 시급하다. 언어폭력 방지를 위해 현재 SK텔레콤의 T 청소년 안심팩 서비스가 제공되고 있으나, 정상적인 문장도 비속어로 오탐지하고, 해당 내용을 부모에게 전송하는 문제점이 있다. 이는 사생활 침해 우려와 부모-자녀 간의 갈등을 일으키는 요인이 될 수 있다.

이러한 문제를 해결하기 위해 본 논문에서는 바른말ㅆ미 서비스를 제안한다. 이 서비스의 목적은 자녀의 건전하고 올바른 언어 사용 습관을 길러주는 기능을 제공함으로써 정서적 안정과 건강한 의사소통 능력을 향상시키고, 사이버 폭력을 예방하는 것이다. 이 서비스의 특징은 LLM(Large Language Model) 기반의 분류 모델과 RAG(Retrieval Augmented Generation)를 활용하여 비속어를 정확히 감지하고, 상황에 적합한 대체 표현이나 이모티콘을 실시간으로 제공한다. 또한, 자녀의 언어 사용 습관에 대한 리포트를 자녀와 부모에게 제공함으로써 자기 인식과 언어 사용 습관을 개선할 기회를 제공한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 서비스의 특징과 한계점을 분석한다. 3장에서는 제안하는 바른말ㅆ미 서비스의 주요 기능과 적용한 AI 기술에 대하여 상세히 설명한다. 4장에서는 바른말ㅆ미 서비스 구현에대하여 기술한다. 마지막으로 5장에서 결론과 향후 연구 방향을 제시한다.

II. Preliminaries

1. Related works

기존 서비스를 벤치마킹한 결과는 표 1과 같다.

Table 1. Bench-marking services

Service	Advantage	Disadvantage
Baleunmal Keypad	• Profanity substitution • Gamification elements	• Limited context understanding • Difficulty detecting modified profanity • Fixed conversion methods
T Youth Safety Pack	• Integrated youth protection features • Real-time bullying alerts	• False detection due to context misinterpretation • Privacy concerns

1) Barunmal Keypad

바른말 키패드(Barunmal keypad) 앱은 청소년의 비속어 사용을 줄이기 위해 개발된 키보드 앱으로, 비트 바이트라는 고등학생 스타트업의 사회공헌 활동으로 시작하였다. 이 앱은 사용자가 입력하는 비속어를 감지하여 이를 순화된 표현이나 이모티콘으로 변환함으로써 비속어 사용을 자연스럽게 최소화하도록 돕는다[5]. 사용자는 앱을 통하여 자신의 비속어 사용량을 기록하고, 이를 그래프로 확인할 수 있으며, 바른말 점수와 함께 랭킹 시스템을 통해 친구들과 비교할 수 있다. 이러한 게이미피케이션 요소 덕분에 사용자들이 앱을 더 재미있게 이용할 수 있도록 만들며 그 결과, 출시 이후 13만 건 이상의 다운로드를 기록했다. 사용자 중 90% 이상이 청소년이었다는 결과는 청소년층에서 높은 인기를 얻은 앱임을 알 수 있다[5, 6]. 이러한 바른말 키패드는 기능적 장점을 가지고 있지만, 몇 가지 한계점도 존재한다. 문맥과 감정 파악 능력의 부재로 비속어의 의도나 사용 맥락을 파악하지 못하기 때문에, 비속어를 단순히 이모티콘으로 변환하는 방식을 채택했다[6]. 이는 텍스트를 입력하는 사용자의 목적을 왜곡하는 등 부적절한 결과를 초래할 수 있다. 또한 변형된 비속어에 대한 대응이 어려워, 새로운 비속어나 은어를 자동으로 감지하고 변환하는 데에 한계가 있다는 것이다. 바른말쓰미 서비스와 바른말 키패드를 비교하면, 바른말쓰미는 인공지능을 통해 문맥과 감정을 파악하여 상황에 따라 적절한 이모티콘으로 변환할 수 있다는 것이 차별점이다. 바른말 키패드가 고정된 방식으로 비속어를 변환하는 반면, 바른말쓰미는 실시간 AI 분석을 통해 더 정교한 대응이 가능하다. 또한, 사이버 폭력 탐지 기능과 부모 계정과의 연동으로 보다 포괄적인 언어 사용 관리 서비스를 지원하여 청소년뿐만 아니라 학부모들에게도 유용한 서비스를 제공한다.

2) T Youth Safety Pack

T 청소년 안심 팩(Youth Safety Pack) SK텔레콤에서 제공하는 청소년 보호 모바일 앱이다. 스마트폰 중독 예방, 학교 폭력 예방, 유해 콘텐츠 차단, 자녀 위치 조회 등 다양한 청소년 보호 기능을 하나의 통합 앱으로 제공한다. 이 중 학교 폭력 예방 기능은 자녀가 송•수신한 메시지에서 학교 폭력이나 비행이 의심되는 위협적이고 불건전한 어휘가 존재할 경우, 부모에게 즉시 해당 내용을 알려준다. 하지만 이 과정에서 발생하는 주요 문제는 룰베이스 기반의 비속어 탐지가 문맥을 이해하지 못해 오탐지를 자주 일으킨다는 것이다. 즉, 자녀가 친구와 나눈 일상적인 대화조차 학교 폭력으로 잘못 인식되어 부모에게 전송되며, 이것으로 인해 자녀의 사생활이 과도하게 침해된다는 것이다. 실제 사례로, 한 초등학생이 친구에게 “니 미술 준비물 좀 빌려줘”라는 메시지를 보냈을 때, 부모에게는 자녀가 학교 폭력에 노출되었다는 알림이 전송되었다[7]. 이는 ‘니 미’와 ‘빌려’라는 단어를 각각 비속어와 갈취로 판단한 결과이다. 이처럼 오탐지된 내용이 부모에게 그대로 전송되면서, 자녀의 일상 대화까지 부모가 감시하는 상황이 발생하여 사생활 침해의 우려가 커지고 있다는 것이다. 이러한 문제점을 해결할 수 있는 서비스가 필요하다.

바른말쓰미 서비스는 LLM과 RAG를 기반으로 문맥을 파악하고, 자녀에게 언어 교정 대체 표현을 제시한다. 이를 통해 자녀가 비속어를 사용하지 않도록 유도하고, 부모에게는 전체 대화 내용을 전송하는 대신 주간 리포트 형식으로 자녀의 비속어 사용 현황을 요약하여 제공한다.

III. The Design of 바른말쓰미 Service

본 논문에서는 LLM과 RAG(Retrieval Augmented Generation)를 기반으로 하는 바른말쓰미 서비스를 설계한다. 제안하는 서비스는 키패드와 앱 부문으로 나뉜다. 키패드에는 비속어를 감정에 맞는 이모지 대체, 비윤리적 표현의 위험등급 실시간 표현 기능을 제공한다. 앱에서는 언어 습관 점수, 사용자와 보호자 각각에 맞는 언어 습관 및 사이버 폭력에 관한 리포트가 제공이 된다. 앱 개발 및 구현을 위한 프레임워크는 그림 1과 같다.

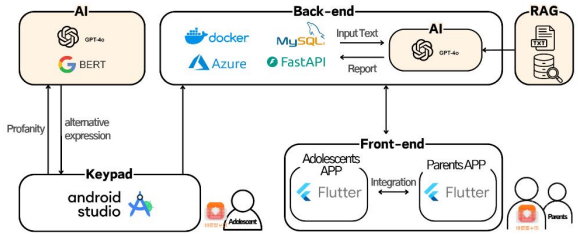


Fig. 1. Service Architecture

그림 1에서 제공하는 키패드 서비스는 BERT[8] 모델을 활용하여 사용자가 입력한 텍스트를 기반으로 문장의 비윤리 정도, 비속어 사용 여부와 감정에 대해 파악할 수 있다. 이를 기반으로 키패드 내에서 비윤리적 표현은 LLM을 통해 다른 표현으로 대체되도록 구현한다. 또한 키패드에 입력된 텍스트들은 모두 저장하여, 주 단위로 리포트 요청 시 Azure 서버로 전송하여 저장하고, 사용자용과 보호자용 RAG 기반의 리포트를 생성하여 전달하도록 구현한다.

앱의 ERD(Entity Relationship Diagram)은 그림 2와 같다.

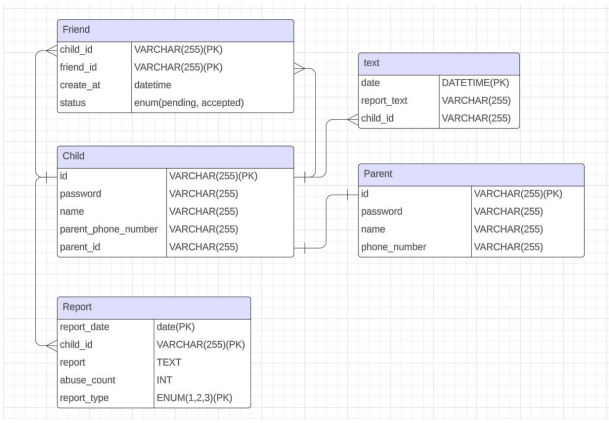


Fig. 2. ERD Diagram

그림 2 ERD는 DB 최적화를 위해 Friend Child, Report, Text, Parent 등 5개의 테이블로 구성한다.

1. Artificial Intelligence

1) Keypad AI

키패드에는 두 가지 AI 모델을 사용하며 워크플로우는 그림 3과 같다.

첫 번째로, 비윤리와 비속어 및 감정 분류 모델로 KoSimCSE-bert-multitask[9]를 사용한다. SimCSE[10] 모델은 사전학습된 BERT로 문장을 벡터 공간에 임베딩한다. 그 후 유사한 문장은 최근 위치에 다른 문장은 최대 위치에 존재하도록 하며, 한글에 특화된 대조 학습 (Contrastive learning) 기반 오픈소스 모델이다.

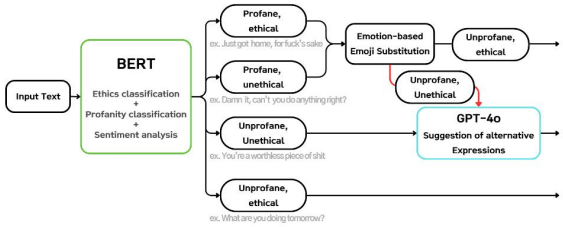


Fig. 3. Keypad Work Flow

이 모델을 활용하여 사용자가 키패드에 입력하는 텍스트를 음절 단위로 입력마다 추론한다. 이때 "나 몇 호로 가?" 처럼 비속어로 오인될 수 있는 표현이 이모지로 대치되지 않도록, 임계값을 넘으면 비속어를 이모지로 대체한다. 이모지 대치는 데이터베이스에 있는 비속어 리스트를 기반으로 한다. 사용자의 감정을 대체하여 효과적으로 전달할 수 있도록 한다. KoSimCSE-bert-multitask은 파라미터 수가 110M이다. 이는 최신 온디바이스용 LLM의 파라미터 수보다 20~80배 정도 적다[11]. 따라서 온디바이스에서 실행이 가능하며, 이를 통해 개인정보 유출을 최소화할 수 있다.

두 번째로, GPT-4o를 활용하여 비윤리 문장을 대체한다. 비윤리 문장이지만 비속어가 사용되지 않는 문장 대체를 위해 LLM을 사용한다. 예를 들어 "너 쓰레기구나" 같은 문장이다. LLM에 입력하는 문장은 완성된 문장이어야 한다. 따라서 사용자의 선택으로 대치할지 결정하도록 한다. 적절한 문장 추천을 위해 프롬프트 엔지니어링(Prompt Engineering)을 한다. 또한 무분별한 API 호출을 막기 위해 임계치를 넘었을 때만 대치하도록 한다.

2) Report AI

본 앱은 객관적이고 정확한 리포트 제공을 위해 RAG 기술을 활용한다[12]. RAG 시스템의 주요 구성 요소와 작동 과정은 다음과 같다.

❶ 문서 등록: '자녀의 누적된 키패드 입력 내용'을 참고 문서로 등록한다.

❷ 유사성 검색: 입력된 쿼리와 등록된 문서 간의 유사성을 검색한다.

❸ 프롬프트 엔지니어링: 사이버 폭력의 기준과 대응 방법을 포함한 프롬프트를 설계하여 객관적 사실에 기반한 리포트 생성을 유도한다. 카카오톡의 운영정책과 푸른 코끼리의 사이버 폭력 관련 내용을 참고했다[13,14].

❹ 정보 기반 생성: 검색된 정보를 바탕으로 인사이트와 리포트를 생성한다.

RAG 기반 리포트 생성 시스템의 주요 특징은 다음과 같다. 첫째, 사생활 보호 기능이다. 자녀의 실제 발화 내용 중 비속어를 사용한 경우에만 분석을 진행하여 불필요한 사생활 침해를 방지한다. 둘째, 문맥 파악 능력이다. 자녀의 전체 발화 내용을 바탕으로 문맥을 정확히 파악함으로써 오탐지를 최소화한다. 이러한 접근 방식을 통해, 본 Report AI는 자녀의 언어 사용 패턴에 대한 정확하고 유용한 인사이트를 제공하면서도 자녀의 개인 보호를 강화할 수 있다.

IV. The Implementation of 바른말쓰미 Service

1. Keypad AI Experiment

데이터는 AI-hub에 존재하는 텍스트 윤리 검증 데이터를 사용하였다[15]. 이 데이터는 한국어 비윤리 텍스트 453,320 문장에 대하여 비속어, 비난, 범죄, 차별, 혐오, 선정, 폭력에 대한 라벨링 되어 있고, 추가적으로 KOTE (Korean Online That-gul Emotions) Dataset[16]으로 학습한 KcELECTRA[17] 모델을 활용하여 추가적인 감정 라벨링을 수행한다. 감정값은 0부터 1 사이의 값으로 도출되었으며, 0.5 이상은 1로, 0.5 미만은 0으로 재분류한다. 결과적으로 총 52개의 Feature 값이 있는 데이터를 완성한다.

실험은 NVIDIA Geforce TitanX 8대 장착한 시스템과 Pytorch를 사용하여 수행한다. 모든 모델은 동일한 시스템에서 훈련 및 테스트한다. 모든 훈련 과정은 동일한 하이퍼 파라미터에서 실행된다. 평가 지표는 F1-score을 사용한자. F1-score는 모델의 정밀도, 재현율을 종합한 성능을 수치화하고 싶을 때 주로 사용되며, 모델의 균형 잡힌 성능을 측정하는 데 사용된다. 텍스트 윤리 검증 데이터의 검증 세트를 기준으로 측정한다. 시뮬레이션 파라미터는 표 3과 같다.

Table 3. Simulation Parameter

Parameter	Value
Epoch	3
Learning rate	0.001
Batch size	32

2. Keypad AI Evaluation

기존의 윤리 검증 데이터뿐만 아니라 KcELECTRA를 사용한 감정 라벨을 추가하더라도, 표 4에서 확인할 수 있듯이 성능의 차이는 크지 않았다. 하지만 동음이의어와 같은 예외 사항을 테스트하였을 때, 윤리 검증 데이터를 사용하여 학습한 경우보다 윤리 검증과 감정 데이터를 결합하여 사용한 경우가 우수함을 알 수 있다.

Table 4. F1-score trained KoSimCSE-BERT-multitask

	Moral	Moral + Emotion
immoral	0.71	0.71
ABUSE	0.07	0.07
DISCRIMINATION	0.06	0.06
HATE	0.05	0.11
CENSURE	0.51	0.25
VIOLENCE	0.03	0.06
CRIME	0.03	0.03
SEXUAL	0.08	0.08

Table 5. Immoral score trained KoSimCSE-BERT-multitask

Test Sentence	KOTE	KOTE+Emotion
쓰레기 좀 치워라	0.79	0.34
몇 호로 가?	0.90	0.09
올해는 병신년의해야	0.93	0.09

또한, 표 5와 같이 비윤리 점수가 낮아야 하는 중의적인 표현 같은 예외 사항에서 우수한 성능을 보였기 때문에 기존 서비스인 바른말 키패드, T 청소년 안심팩보다 우수한 성능을 도출할 것으로 예상한다.

3. Implementation of Front-end

본 논문의 프론트엔드는 Flutter 프레임워크를 기반으로 설계되었으며, 학생용과 부모용의 두 가지 앱으로 구현한다. 각 앱은 학생과 부모의 사용 목적에 따른 UI/UX를 구현한다. 또한, 전체적인 디자인은 사용자의 경험을 간소화하고 접근성을 높이기 위해 직관적인 인터페이스와 시각적 요소를 고려한다.

1) Report application for student

학생용 앱은 사용자의 언어 습관을 개선하기 위해 개발된 키패드와 보고서 기능을 중심으로 구현한다. 사용자는 앱에서 자신의 비속어 사용 패턴을 확인할 수 있으며, 이를 기반으로 언어 습관을 개선할 수 있는 리포트를 제공받는다. 메인 페이지는 그림 4와 같다.

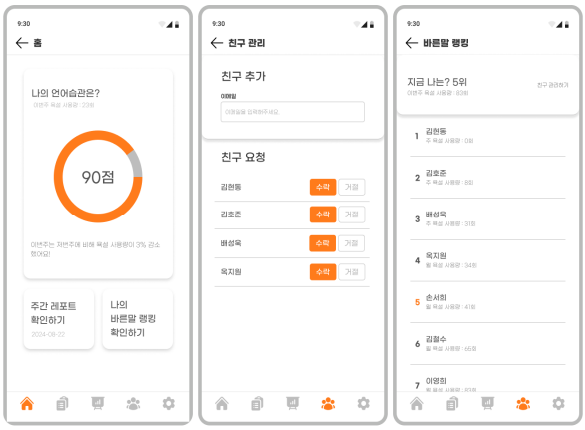


Fig. 4. Main page & Ranking page for student

학생용 메인 페이지는 사용자의 금주와 지난주의 비속어 사용량을 비교하여 시각화한다. 사용자는 자신의 점수를 확인하고, 주간 리포트 및 바른말 랭킹을 확인할 수 있는 페이지로 이동할 수 있다. 랭킹 페이지는 비속어 사용량을 순위로 표현한 페이지이다. 사용자는 자신과 친구의 랭킹을 확인하고, 친구 관리 기능을 통해 새로운 친구를 추가하거나 친구 요청을 관리할 수 있다. 친구 랭킹은 리스트 형식으로 제시되며, 각 친구의 비속어 사용량과 순위가 표시된다.

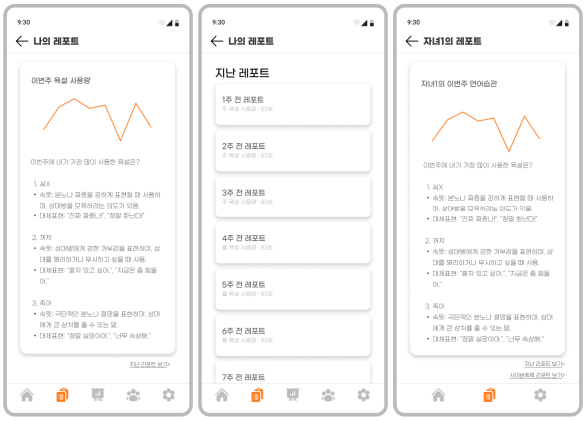


Fig. 5. Report page for student

리포트 페이지는 사용자의 언어 사용 습관에 대한 상세한 분석 결과를 나타낸다. 이번 주 사용량을 지난 5주와 비교하여 가장 빈번하게 사용된 비속어의 대체 표현 등이 그림 5와 같이 꺾은선 그래프와 함께 제시된다. 해당 보고서는 사용자가 자신의 언어 습관을 분석하고 개선할 수 있도록 도와준다.

2) Report application for parent

부모용 앱은 자녀의 언어 사용 습관을 모니터링하고, 자녀의 언어 습관 리포트와 사이버 폭력 보고서를 확인할 수 있는 기능을 지원한다. 부모는 자녀의 언어 사용 패턴을 주기적으로 확인하며, 이를 기반으로 자녀의 언어 습관 개선을 돕는 역할을 할 수 있다. 주요 페이지는 그림 6과 같다.

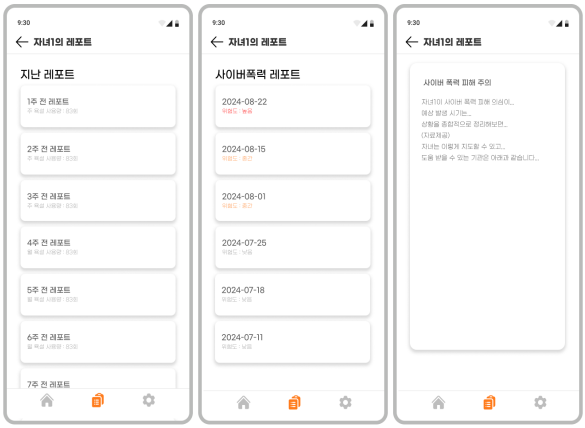


Fig. 6. Cyberbullying report page for parent

부모용 메인 페이지는 자녀의 금주와 전주의 비속어 사용량을 비교하여 시각적으로 확인할 수 있다. 이를 통해 자녀의 언어 습관을 모니터링하고, 필요에 따라 피드백을 제시할 수 있다. 부모용 리포트 페이지는 자녀의 언어 사용 습관을 확인할 수 있도록 구현한다. 이 화면에서는 자녀의 주간 비속어 사용량과 지난 다섯 주와의 비교, 그리고 사용된 비속어의 패턴을 꺾은선 그래프와 함께 보여준다. 또한 자녀의 언어 습관을 지도할 방법을 제안받을 수 있다. 이를 통해 부모는 자녀의 언어 습관 변화를 한눈에 파악할 수 있으며, 필요에 따라 자녀에게 피드백을 제공하거나 더 심층적인 모니터링을 할 수 있다. 사이버 폭력 리포트는 자녀가 사이버 폭력에 연루된 정도를 파악할 수 있는 보고서를 제공한다. 해당 보고서는 자녀가 사용하는 비속어과 그 심각성을 분석하여 부모에게 알린다. 부모는 이를 통해 자녀의 언어 습관을 개선할 수 있도록 지도할 수 있다.

4. Implementation of Back-end

앱의 주요 기능을 실행하기 위한 End Point는 표 6과 같다.

Table 6. Backend End Point

End Point	HTTP-Method
/signup	POST
/save_txt	POST
/reports/generate_report	POST
/reports	GET
/friend/ranking	GET

Child와 Parent 테이블은 각각 자녀와 부모의 계정 정보를 저장하는 엔티티로 아이디, 비밀번호, 이름을 포함한다. Child 테이블에는 부모의 휴대폰 번호와 부모 아이디가 추가로 저장되며, Parent 테이블에는 부모 본인의 휴대폰 번호가 저장된다. Report와 Text 테이블은 리포트 생성과 관련된 데이터를 저장한다. Report 테이블은 생성된 리포트를, Text 테이블은 입력된 문장 단위 데이터를 저장한다. Friend 테이블은 자녀의 친구 정보를 저장한다. 제안된 앱은 리포트 생성, 친구 간 비속어 랭킹 조회, 부모-자녀 간 연동의 기능을 제공한다. 이를 구현하기 위해 백엔드에 주요 엔드 포인트를 설정한다.

리포트 생성 기능은 “/save_txt”를 통해 키패드에 입력된 문장 단위를 데이터베이스에 저장한다. “/reports/generate_report” 통해 저장된 문장 데이터를 주간 단위로 합쳐 주간 리포트를 생성한다. 생성한 리포트는 “/reports/type1”, “/reports/type2”, “/reports/type3”를 통해 각각 부모용 리포트, 자녀용 리포트, 사이버 폭력 리포트로 조회한다.

친구 간 비속어 랭킹 기능은 “/friend/send_request”, “/friend/accept_request”를 통해 친구 요청 및 수락 과정을 거친 후, “/friend/ranking”을 통해 로그인된 사용자와 친구 간 비속어 랭킹을 조회한다.

부모-자녀 간 연동은 회원가입 시 동시에 이루어진다. 자녀가 “/signup/child”를 통해 회원 가입하면 입력된 부모 휴대전화 번호로 SMS 인증 번호가 전송된다. 이때 SMS 전송은 coolSMS API를 사용하며 인증 번호는 랜덤 난수로 생성하여 발송한다. 부모는 “/signup/parent”를 통해 계정을 생성하며 부모가 입력한 인증 번호와 생성된 인증 번호가 일치할 때 부모-자녀 계정이 연동된다.

V. Conclusions

본 논문에서는 청소년의 건전한 언어문화 형성과 사이버 폭력 예방을 위한 AI 기반의 바른말ㅆ미 서비스를 설계하고 구현하였다. 이 서비스는 자녀의 사생활을 보호하면서 비속어 사용을 줄일 수 있는 기능을 제공한다. 또한 학생과 부모의 요구에 맞는 기능을 지원하기 위해 학생용 앱과 부모용 앱을 별도로 설계하고 개발하였다. 키패드에 사용한 KoSim-CSE-BERT-multitask 모델은 온디바이스에서 데이터 노출없이 빠르고 정확한 결과를 제시하고, LLM과 RAG를 활용한 리포트에는 부모에게 자녀의 실제 대화 내용의 일부만을 공개한다. 그 결과 학생들에게는 실시간 언어 교정과 자기 인식 기회를 부여하고, 부모에게는 자녀의 언어 습관에 대한 통찰력 있는 정보를 제공한다. 초등학교 선생님과 학생들을 대상으로 바른말ㅆ미 서비스의 기능을 설명하고, 인터뷰한 결과 비속어 사용에 따른 사이버 및 학교 폭력을 줄일 수 있다는 긍정적인 답변을 얻었다. 바른말ㅆ미 서비스는 올바른 언어 습관을 형성하는 조기 교육 도구로 사용되며, 안전한 온라인 환경 조성에 중요한 역할을 할 것으로 기대한다.

ACKNOWLEDGEMENT

This work was supported by the SK Telecom's FLY AI Challenger program, conducted in collaboration with the Ministry of Employment and Labor and the Korean Skills Quality Authority as part of the 2024 K-Digital Training.

REFERENCES

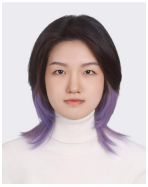
[1] Chanho Kim, "Children Learning Profanity from YouTube and SNS... Cause of Classroom Collapse?," https://www.newsis.com
[2] Suhee Park, "Profanity, First Encountered by 79% in Upper Elementary Grades," http://m.wonjutoday.co.kr
[3] Jinkyu Park, "Social Linguistics: Slang, Profanity, and Abbreviations Commonly Used by Adolescents," https://www.reportworld.co.kr
[4] Eunkyung KIM, "Announcement of Results from the 1st School Violence Survey of 2023," https://www.moe.go.kr
[5] Samsung Electronics, "Smarter and Prettier: The Barunmal Keyboard Now Unstoppable," https://news.samsung.com/kr
[6] Eunkyung Ahn, "Communicate with Grace, and Make Society Warmer," https://press.kookmin.ac.kr
[7] Siji Yoon, "A youth protection app that writes protection and reads surveillance," https://www.sisajournal-e.com
[8] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[9] https://huggingface.co/BM-K/KoSimCSE-bert-multitask
[10] Gao, T., Yao, X., & Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings.arXiv preprint arXiv:2104. 08821. https://doi.org/10.48550/arXiv.2104.08821
[11] Kim, G., Yoon, K., Kim, R., Ryu, J. H., & Kim, S. C. (2024). "Technical Trends in On-device Small Language Model Technology Development," Electronics and Telecommunications Trends," ETRI, Vol. 39 No. 4, pp. 82-92. Aug. 2024 https://doi.org/10.22648/ETRI.2024.J.390409
[12] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). "Retrieval augmented generation for knowledge intensive NLP tasks," Advances in Neural Information Processing Systems, arXiv:2005. 11401 Apr. 2021 https://doi.org/10.48550/arXiv.2005.11401
[13] https://kakao.com/
[14] https://www.bepuco.or.kr/
[15] Crowdworker, "Text Ethics Verification Data", https://www.aihub.or.kr/aihubdata/data/
[16] JDuyoung Jeon, Junho Lee, Cheongtag Kim, "User Guide for KOTE: Korean Online That-gul Emotions Dataset," In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 17254-17270, May 2024 https://aclanthology.org/2024.lrec-main.1499
[17] https://github.com/Beomi/KcELECTRA

Authors

Ho-joon Kim will receive the B.S. degrees in Computer Engineering from Seoul National University of Science and Technology, Korea, in 2026.

Hyun-dong Kim will receive the B.S. degrees in Computer Engineering from Hankuk University of Foreign Studies, Korea, in 2025.



Seo-hee Son will receive the B.S. degrees in Computer Engineering and Content Convergence from Ewha Womans University, Korea, in 2026. Seo-hee Son joined the student of the Department of Computer Engineering at Ewha Womans University, Seoul, Korea, in 2021. She is double majoring the Department of Contents Convergence at Ewha Womans University. She is interested in internet and mobile computing, game development, and contents planning.

Sung-uk Bae received the B.S. degrees in Computer Engineering from Konkuk University, Korea, in 2024.

Ji-Won Ock received the B.S. in Convergence Security Engineering in 2022 and the M.S. in Future Convergence Technology Engineering in 2024, both from Sungshin Women's University, Korea. Ji-Won Ock is currently a Researcher at the ADD(Agency for Defense Development). She is interested in information security, artificial intelligence, and mobile computing.

Sejong Lee received a B.S. degree in Computer Science and Engineering from Jeju National University, South Korea, in 2018. He received a Ph.D. degree in Computer Science and Engineering from Hanyang University, South Korea, in 2024. Dr. Lee joined the faculty of College of Digital Convergence, School of Computer Science and Engineering at Yeungnam University, Korea, in 2025. His research interests include IoT security and Blockchain-based medical data-sharing systems, Artificial intelligence, Cloud platform.