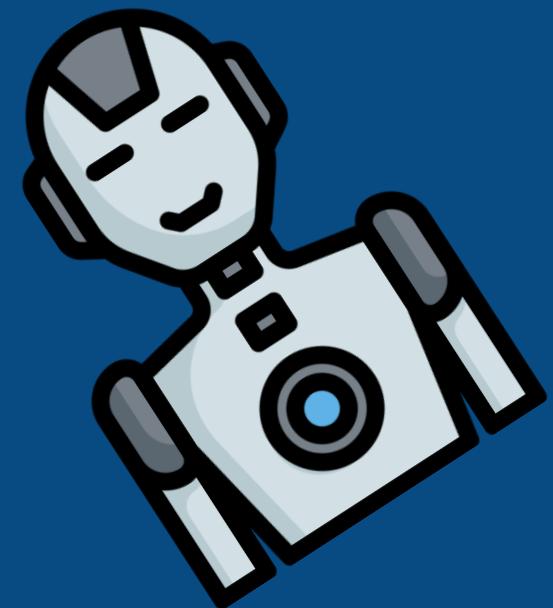


이청수와 아이들 최종 발표

신개념 토론 싱크탱크 방구석 대한민국



✓ Role of Each Member

김석: 백엔드 서버 개발

박진우: 대화요약 모델 개발

이청수: 프론트엔드 개발

장채윤: 비방성표현 마스킹 모델 개발

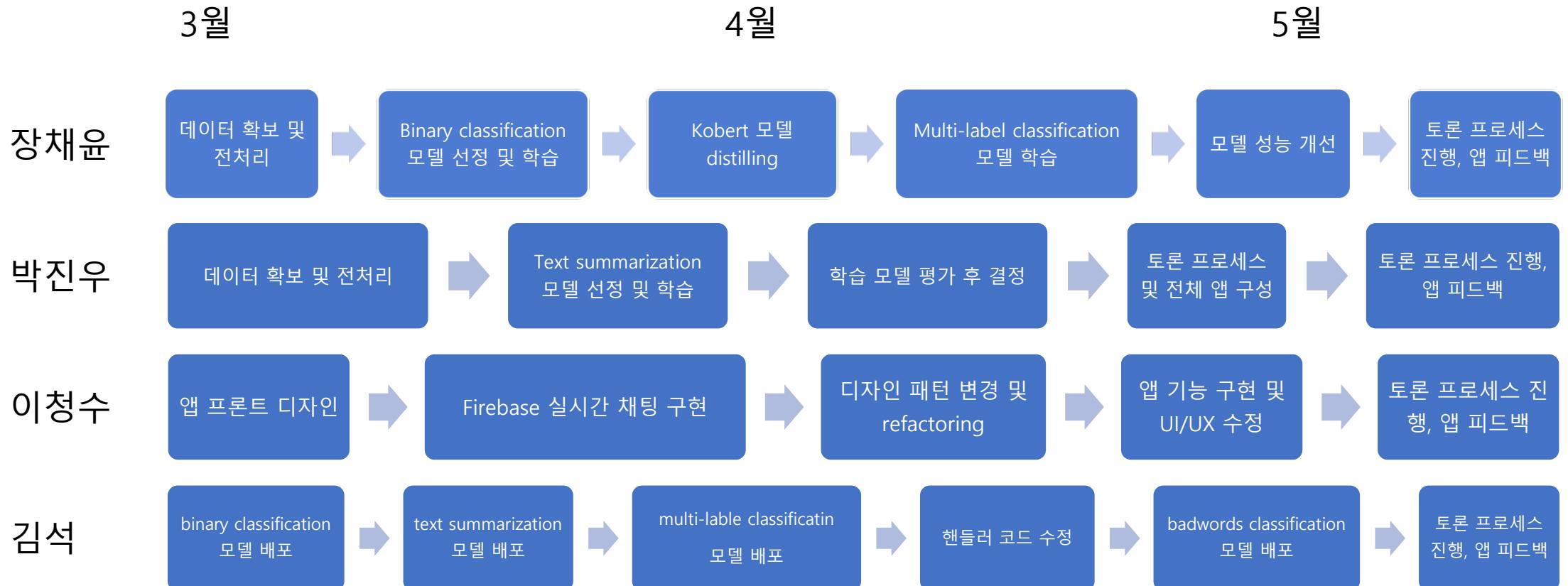
● Schedule

- 매주 일요일 3시 정기 미팅 진행
 - 앱 완성 후 매주 정기 토론 진행 및 피드백

Project Progress

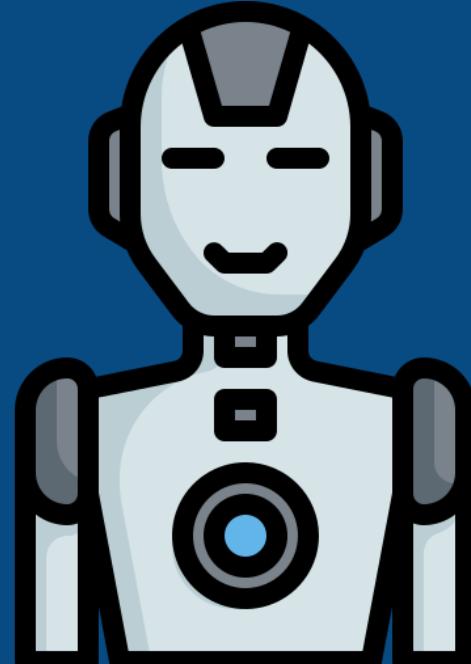
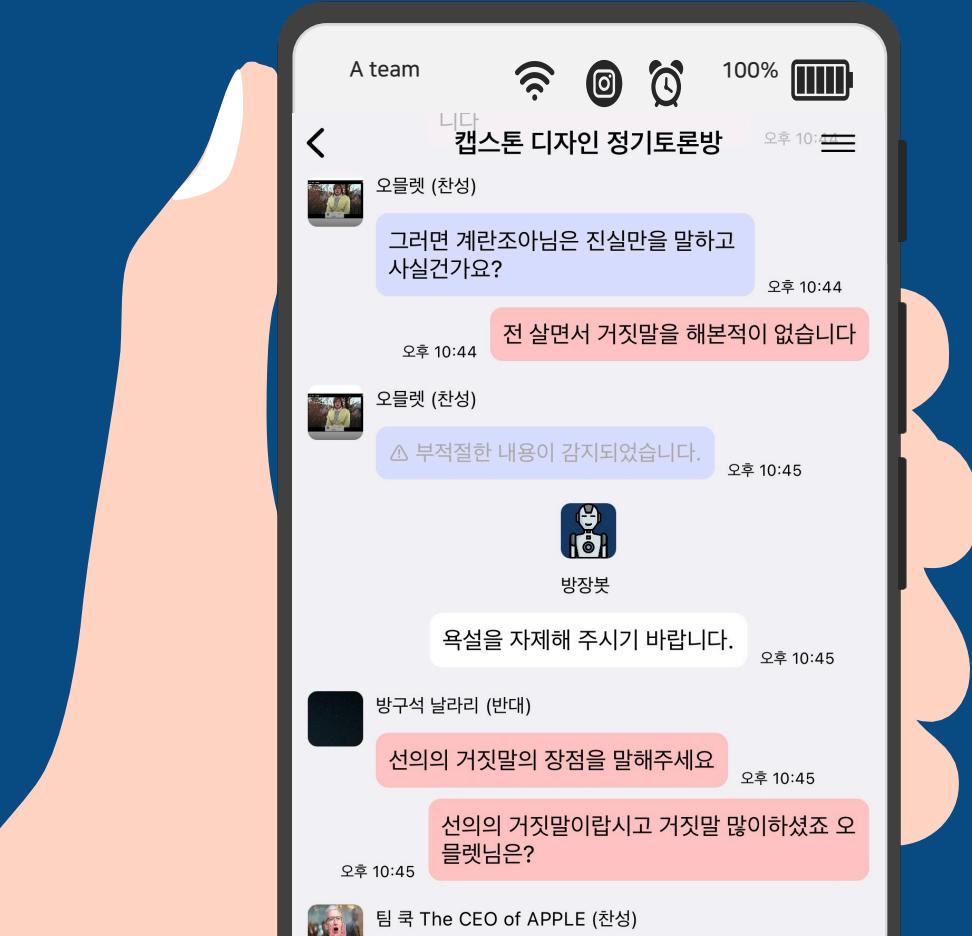
A Team ⌂ ⌂ ⌂ 100% 🔋

Milestone



방구석 대한민국이란?

- ✓ 기획 동기
- ✓ 강점
- ✓ 목표



✓ 방구석 대한민국의 기획 동기

1. 현대사회 문제를 해결하기 위해 토론만큼 좋은 수단이 없음.
2. 고전식 토론의 문제를 보완한 새로운 구조의 토론 플랫폼이 없음.
3. 익명성과 건전성을 모두 갖춘 토론 플랫폼이 없음.

일반 사람들이 토론에 크게 재미를 느끼지 못함

기존 플랫폼들은 공정한 사회자가 없음.

방구석 대한민국의 강점

A Team 100%

✓ 방구석 대한민국의 강점

1. 공정한 플랫폼

- 인공지능 사회자가 토론을 진행하여 양측에 이해관계에 얹혀 있지 않음.

2. 유익한 플랫폼

- 다양한 의견을 낼 수 있도록 익명성을 도입하고 대신 비방성 표현을 억제하여 건전하고 자유로운 토론풍화를 형성함.

3. 재미있는 토론

- 기존 복잡한 토론 방식들을 간소화하고 토론 대화를 요약해주며 구조를 게임형식으로 구성하여 편리하고 재미있는 토론이 가능함.

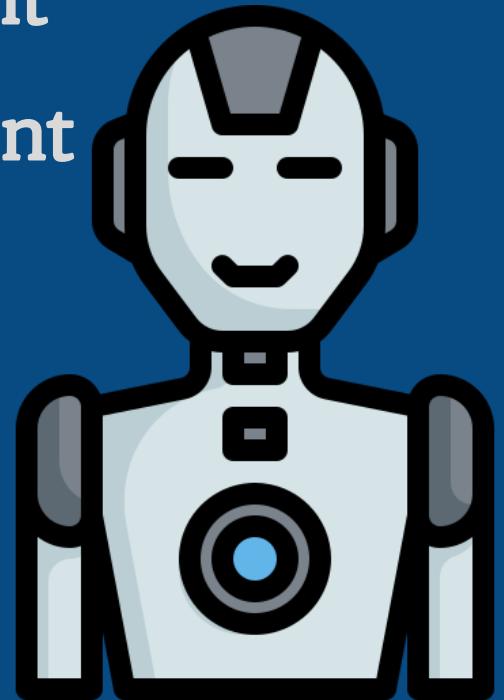
✓ 방구식 대한민국의 목표

AI 기술과 기존 토론 방식의 장점을 융합하여 사회에 만연해 있는 극단성을 극복할 수 있도록, 생산적이고 유의미한 토론풍화를 형성 해 나가는 것



기술 설명

- ✓ AI Model Development
- ✓ Back-End Development
- ✓ Front-End Development



✓ Text Summarization

- Dataset: [Dacon] 한국어 문서 생성요약 AI 경진대회 (40k)
[AI HUB] 문서요약 텍스트 (350k)
- Model description: [model name] Kobart (SKT에서 만든 한국어 BART 모델)
[Pretrained model] 'digit82/kobart-summarization' 사용
- Hyperparameters : beam, epoch, combination of datasets
- Evaluation metrics: 디베이팅 데이에서 크롤링한 토론 샘플 200개 human evaluation
- results: [model epoch] 10 | 20 | 30 | 40 | 50
[evaluation] 5.1 | 5.7 | 6.3 | 6.7 | 6.1

✓ Text Summarization

- metric example

model_version	3	4	5	6	7
노키즈존	5	4	4	7	7
인공지능	3	4	5	9	8
9시등교	7	7	5	8	5
컴퓨터계산	9	7	7	10	8
사드배치	8	9	6	7	7
sum	32	31	27	41	35

사람의 평가

✓ Offensive Comment Masking

- Dataset:

[Naver] 2021.01 ~ 2021.11 Top10 news comment 크롤링 (20K)

[욕설감지데이터셋] <https://github.com/2runo/Curse-detection-data> (5K)

[Korean-hate-speech] <https://github.com/kocohub/korean-hate-speech> (15k)

= Total : 약 40K Binary Classification Dataset

[Korean-unsmiled-dataset] https://github.com/smilegate-ai/korean_unsmile_dataset (18K)

= Total : 약 18K Multi-Label Classification Dataset

[Bad words dataset] <https://asus1004.tistory.com/241>

Model description: [model name] KLUE-Roberta Model

[Pretrained model] 'klue/roberat-base' 사용 두가지 버전으로 파인튜닝

- Evaluation metrics: F1 score, accuracy
- Hyperparameters : epoch, number of label, learning rate, type of model(2 type)

✓ Offensive Comment Masking

Challenges:

- 모델에 임베디드 하기 위해 distilling 작업을 했으나 결과가 F1 50점으로 좋지 않았다.
 - >> 임베디드 하지 않아도 서버에서 결과를 보내는 시간이 0.0X초로 짧아 상관없다.
 - >> 성능이 좋은 main 모델을 사용함.
- 같은 단어라도 다른 맥락을 가질 경우 비방성의 여부에 차이가 날 수 있다
- 비방의 정도나 카테고리도 분류할 수 있으면 방장봇이 사용자에게 경고를 주는데 적합하다
 - >> Multi-Label로 된 데이터셋을 활용

Limitation: Multi-Label의 경우 데이터의 개수가 대용량 언어코퍼스로 사전 학습된 모델을 완전히 Fine-Tuned 시키기에 한계가 있음

- >> Binary한 모델과 Multi-Label 모델을 양상별하여 최선의 성능을 도출할 것
- 새로 만들어진 비방성 표현에 대한 분류 제한이 생긴다.
- >> 딕셔너리를 만들어 지속적인 욕설 신조어 추가가 가능하도록 구성한다.

✓ Offensive Comment Masking

How to Fine-Tune BERT for Text Classification? (<https://arxiv.org/pdf/1905.05583.pdf>)

last layer of BERT gives the best performance.
Therefore, we use this setting for the following experiments.

Layer	Test error rates(%)
Layer-0	11.07
Layer-1	9.81
Layer-2	9.29
Layer-3	8.66
Layer-4	7.83
Layer-5	6.83
Layer-6	6.83
Layer-7	6.41
Layer-8	6.04
Layer-9	5.70
Layer-10	5.46
Layer-11	5.42
First 4 Layers + concat	8.69
First 4 Layers + mean	9.09
First 4 Layers + max	8.76
Last 4 Layers + concat	5.43
Last 4 Layers + mean	5.44
Last 4 Layers + max	5.42
All 12 Layers + concat	5.44

Table 3: Fine-tuning BERT with different layers on
IMDb dataset.

• 논문에서 제안하는 방식

- 일반적으로 뉴럴넷의 다른 layer들은 다른 정보를 capture 한다.
- 하위의 레이어는 좀 더 일반적인 정보를 담고 있다.
- 실제로 분류 모델에서 Bert의 12개 layer중 마지막 4개의 layer만을 사용하여 작업한 경우 성능이 가장 좋게 나온 사례가 있다.
- layer에 대한 parameter를 나누어 learning rate decay시키는 방법이 있다.

✓ Offensive Comment Masking

How to Fine-Tune BERT for Text Classification? (<https://arxiv.org/pdf/1905.05583.pdf>)

So What?

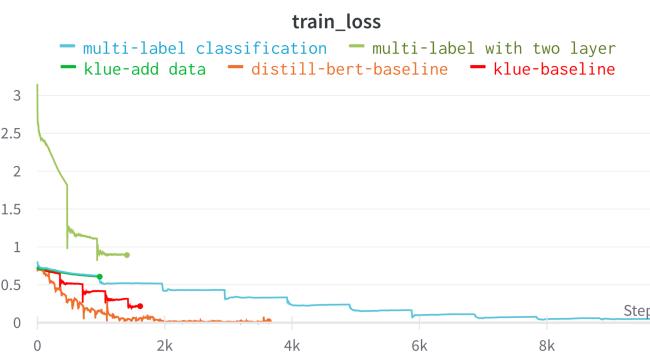
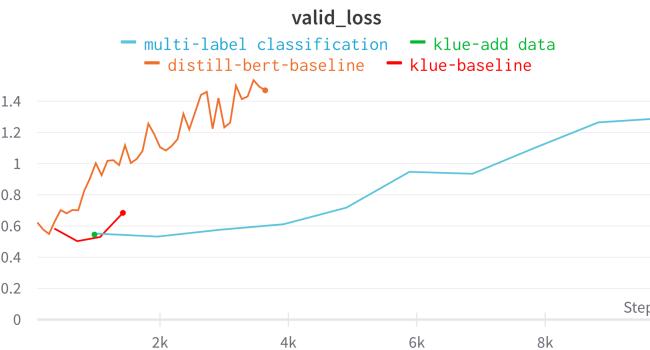
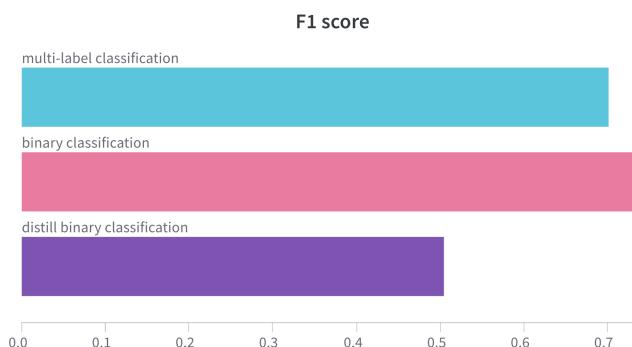
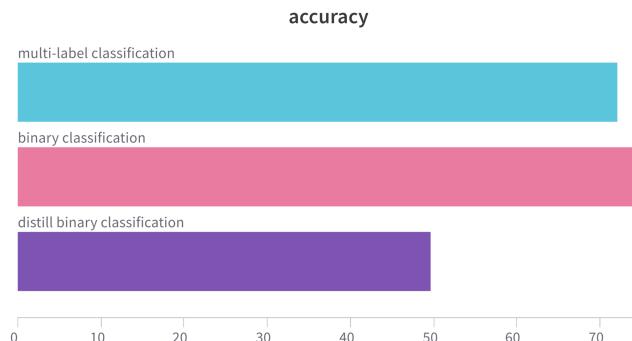
```
pooled_output = torch.cat(tuple([outputs.hidden_states[i] for i in [-4, -3, -2, -1]]), dim=-2)
pooled_output = torch.max(pooled_output, dim=-2).values
pooled_output = pooled_output[:, :]
pooled_output = self.dropout(pooled_output)
logits = self.classifier(pooled_output.view(-1, 768))
```

- In Forward Code,
- 마지막 네 개의 hidden states의 logits값들의 column 방향으로 max인 값을 최종 아웃풋을 위한 hidden state로 사용하였습니다.
- 모델 학습 초반 learning rate 를 2e-5로 매우 낮게 낮추었습니다.

위 두가지 방법은 기존에 사전 학습된 정보를 소실하지 않으면서 파인튜닝을 하는데 도움이 됩니다.

✓ Offensive Comment Masking

- Evaluation Result



- distill binary model
<multi-label < binary F1 약 $0.5 < 0.72 < 0.8$

Multi-label의 경우 성능은 떨어져도 욕설이 어느 카테고리에 해당하는지 까지 분류를 하므로 추가적인 이점이 있다.

따라서 두가지 모델을 양상별 하기로 결정함.

이에 신조어나 모델이 학습하지 못한 비방성 단어 또한 탐지하기 위해 욕설 사전을 구축.

욕설 사전 VS binary model VS multi-label 모델 세가지가 각각 투표를 하고 이긴 쪽으로 문장을 분류한다.

욕설 사전의 경우 사용자의 신고가 들어온 신조어 욕설에 대해 지속적으로 추가 및 수정이 가능하다는 장점.

✓ Offensive Comment Masking Example

비방성 탐지 AI

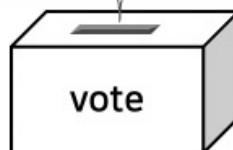
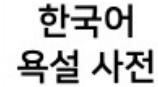
Input Text : 니미럴 드록바 (너희 어머니는 드록바다. 가족 비하 발언)

KLUE-Roberta

Linear[-1, 2]

KLUE-Roberta

Linear[-1, 11]



✓ Result : 86%의 확률로 여성/가족에 해당하는 비방성 표현입니다.

토론 시작

전반전

후반전

투표 및 종료

✓ Offensive Comment Masking Example

- Test Dataset에 있는 Sample을 이용해 확인한 결과입니다.

Type the sentence : 한남동에는 한남들이 많이 사나
해당 댓글은 72%의 확률로 남성에 해당하는 비방성 댓글입니다.

Type the sentence : 한남동에 카페가고 싶다
해당 댓글은 지역에 해당하는 표현이며, clean한 댓글입니다.

Type the sentence : 딱 전라도 마인드임 통수찌는 위선자들
해당 댓글은 82%의 확률로 지역에 해당하는 비방성 댓글입니다.

Type the sentence : 게이 새끼들 다 정상으로 돌아오게 치료받자
해당 댓글은 94%의 확률로 성소수자에 해당하는 비방성 댓글입니다.

Type the sentence : 개독교인들 꺼져라
해당 댓글은 94%의 확률로 종교에 해당하는 비방성 댓글입니다.

Type the sentence : 문재앙
해당 댓글은 75%의 확률로 개인지칭에 해당하는 비방성 댓글입니다.

✓ 비방성 표현 마스킹 모델 3개 배포

1. classification1: klue-roBERTa-base binary classification model

0: not toxic 1: toxic

2. classification2: Badwords checking model

0: not toxic 1: toxic



3. classification3: klue-roBerta-base multi-label classification model

["여성/가족", "남성", "성소수자", "인종/국적", "연령", "지역", "종교", "혐오", "욕설", "clean", "개인지침"]

→ Hard voting을 통해 toxic / not toxic 판단

✓ Badwords checking model

644개 욕설 단어 데이터

해당 단어가 문장에 포함되어 있을 경우 toxic, 아닐 경우 not toxic으로 판단

이 데이터만으로 비방성 표현 판단하기에는 정확성이 떨어지지만,
두 인공지능 모델과 함께 voting하는 방식으로 앙상블하면 비방성 마스킹 성능 향상.

343	"시바",
344	"시발",
345	"시방",
346	"시박",
347	"시벌",
348	"시부랄",
349	"시부럴",
350	"시부리",
351	"시불",
352	"시브랄",
353	"시이발",
354	"시팍",
355	"시팔",
356	"시펄",

<badwords 출처: <https://github.com/organization/Gentleman/blob/master/resources/badwords.json>>

Back-End Development

A Team

WiFi 📸 ⏳ 100% 🔋

✓ 비방성 표현 마스킹 모델 양상들

너같이 생긴놈들이 평균 외모 수준을 낮춰
1 0 육설
toxic

진정하시고 제 말을 좀 들어주세요
0 0 clean
not toxic

대화가 안되네 그냥 끄자
0 1 육설
toxic

어쨌든 수능이 가장 공평한 제도라고 할 수 있습니다.
0 0 clean
not toxic

토론에 집중을 좀 하세요
0 0 clean
not toxic

근데 그쪽은 진짜 수준이 너무 낮네요
1 0 육설
toxic

제발 내 말좀 들어봐 시발
1 1 육설
toxic

아까부터 자꾸 주제에서 벗어나시네요
0 0 clean
not toxic

✓ 대화 요약 모델 배포

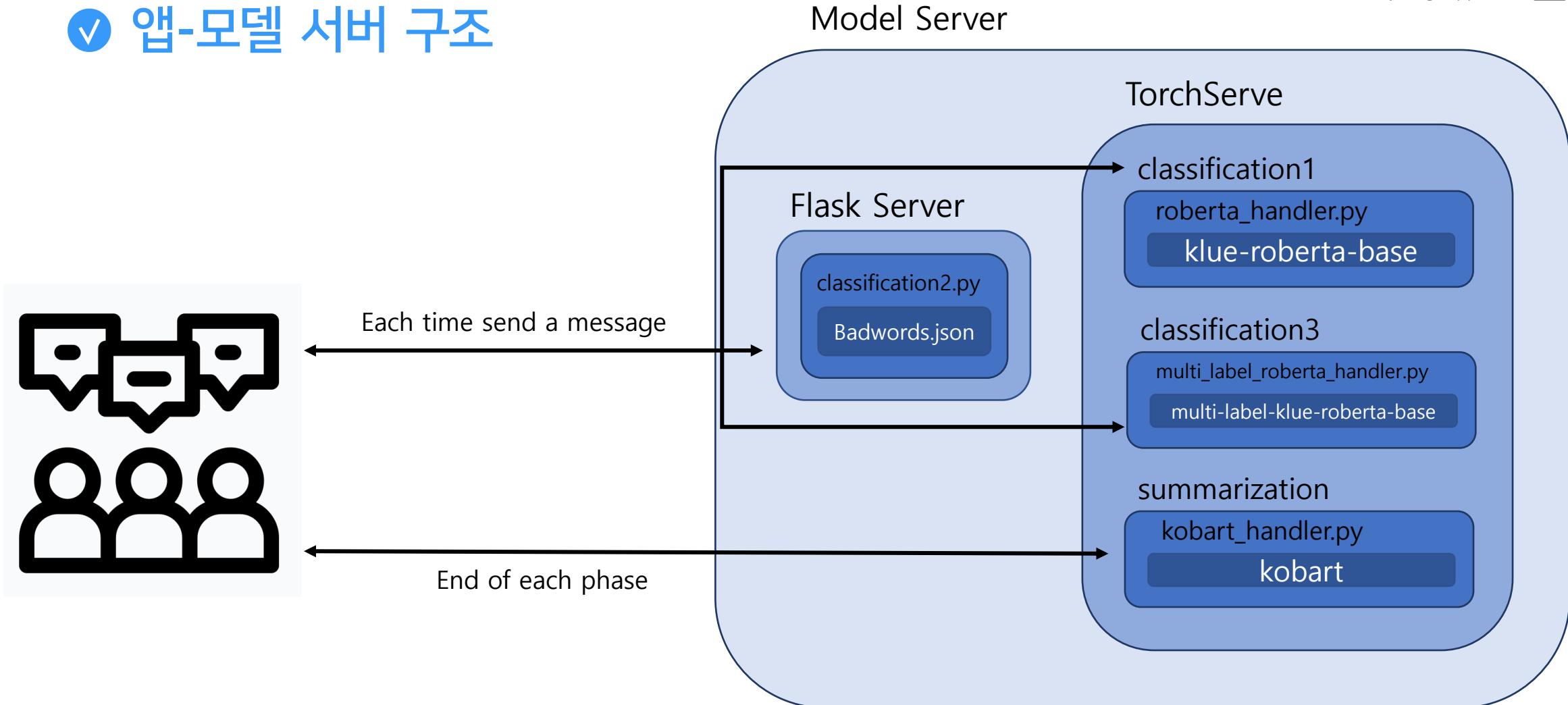
Ex) 반려동물 보유세 찬성 의견

```
> curl --header "Content-Type: application/json" --request  
POST --data '{"text":"사람들은 자유로운 선택으로 반려동물을 입양한다. 자기 삶에 긍정적인 영향을 주기 때문이다. 그렇게 반려동물 인구가 늘어나면서, 사회적 비용도 증가했다. 자신의 자유로운 행동이 사회적 비용을 유발했다면, 그에 대한 책임을 져야 한다. 반려동물로 인한 사회적 비용인 만큼, 반려동물을 키우는 가구에서 감당하는 것이 적합하다. 자유와 권리를 중요하게 생각하는 미국, 뉴질랜드, 네덜란드, 독일 등 선진국에서는 이미 애완견 등록세를 징수하고 있다. 자유를 누리되 그에 맞는 책임을 지는 것이 민주시민의 자세다. 반려동물 인구가 늘어나면서 관련 산업도 꾸준히 증가했다. 현재 3조 원에 이르는 시장은 2027년이 되면 6조 원을 넘어설 거라고 예상된다. 문제는 민간에서 사업을 이끌기 때문에 성장에만 힘쓸 뿐, 반려동물의 복지나 사고 예방 같은 제도적 장치가 전무하다는 점이다. 이러한 문제를 해결하려면 국가가 주도하는 제도적 변화가 필요하고, 그에 맞는 세금도 필요하다. 반려동물을 키우는 가구는 없었던 세금이 생겨 부담스러울 수 있지만, 자신의 반려동물을 위한 첫걸음인 만큼 감당해야 한다."}' http://119.194.17.59:8080/predictions/summarization  
반려동물 인구가 증가하면서 사회적 비용도 증가한 만큼 국가가 주도하는 제도적 변화가 필요하고 그에 맞는 세금도 필요한데 이미 선진국에서는 이미 애완견 등록세를 징수하고 있다.
```

Back-End Development

A Team ⚡ 100%

✓ 앱-모델 서버 구조



✓ 서버 response time

1. 비방성 표현 마스킹 모델

서로 다른 길이의 text 10개 추론 시간 평균

- | | |
|---|--------|
| 1. klue-roBERTa-base binary classification: | 0.092초 |
| 2. klue-roBERTa-base multi-labe classification: | 0.103초 |
| 3. Bad word classification: | 약 0.1초 |

2. 대화 요약 모델

서로 다른 길이의 text 10개 추론 시간 평균

- | | |
|------------|--------|
| 1. koBART: | 4.563초 |
|------------|--------|

✓ 채팅 기능 구현

Challenges: 앱의 핵심 기능인 실시간 채팅 기능 구현하기

별도 채팅 서버 및 이미지 서버 없이 채팅과 프로필 사진 기능을 구현하기 위해 각각

Firebase Realtime Database, Firebase Cloud Storage 사용

- Firebase Realtime Database: 실시간 클라우드 데이터베이스
- Firebase Cloud Storage: 사진, 동영상 등을 업로드 및 다운로드할 수 있는 클라우드 데이터베이스



Firebase
Realtime Database



Cloud Storage
for Firebase

✓ 채팅 기능 구현 - 프로필 사진

네트워크를 이용해 이미지 데이터를 다운받고 View에 띄우는 작업은 Cost가 높아 고려해야 할 요소가 많음

- 네트워크 비동기 요청 보내기
- (성능을 위해) 로컬 캐싱, 메모리 캐싱
- 캐시 데이터 혹은 다운로드한 데이터를 `UIImage`로 바꿔 View에 띄우기

직접 구현하려면 시간이 너무 많이 필요할 것으로 판단,

[Kingfisher](#)라는 이미지 처리 오픈소스 라이브러리를 활용하여 빠르게 구현함 (MIT License).



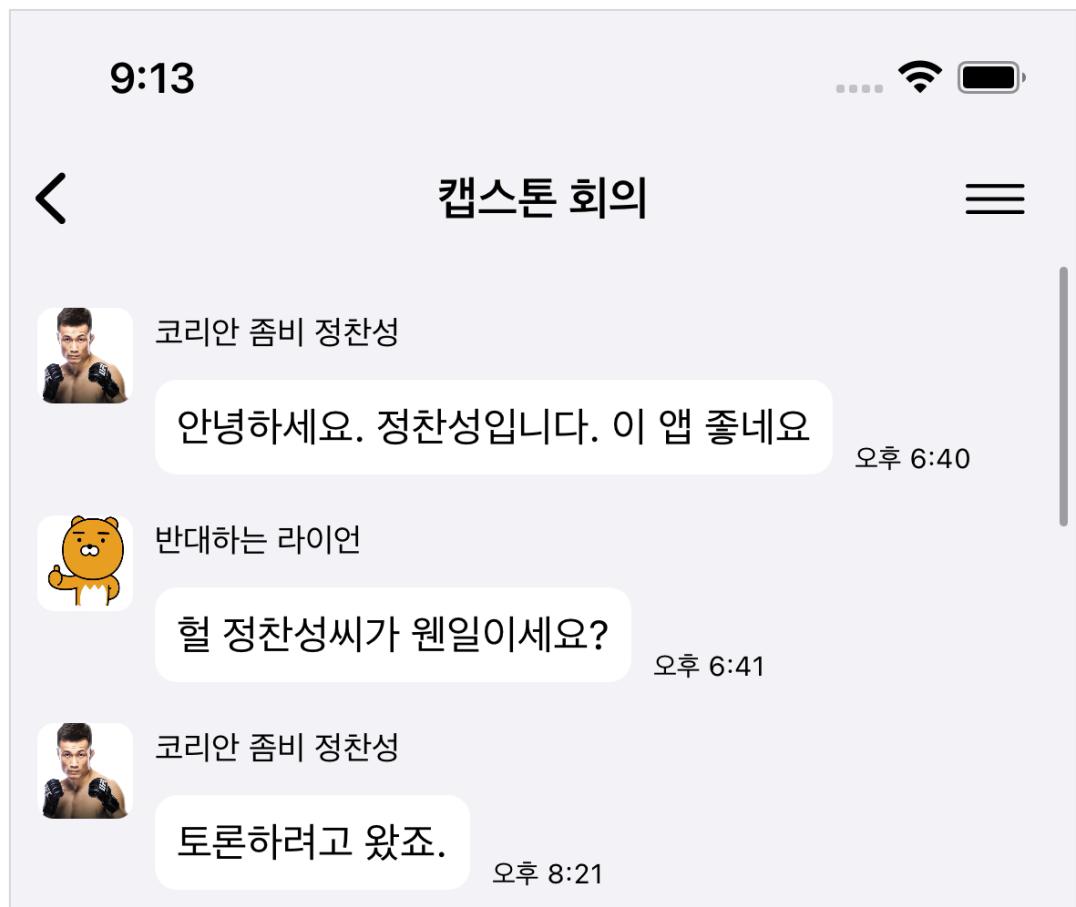
<https://github.com/onevcat/Kingfisher>

Front-End Development

✓ 채팅 기능 구현 - 프로필 사진

결과적으로, Firebase Cloud Storage + Kingfisher로 빠르게 구현 가능했음

A Team ⌂ ⌂ ⌂ 100% 🔋



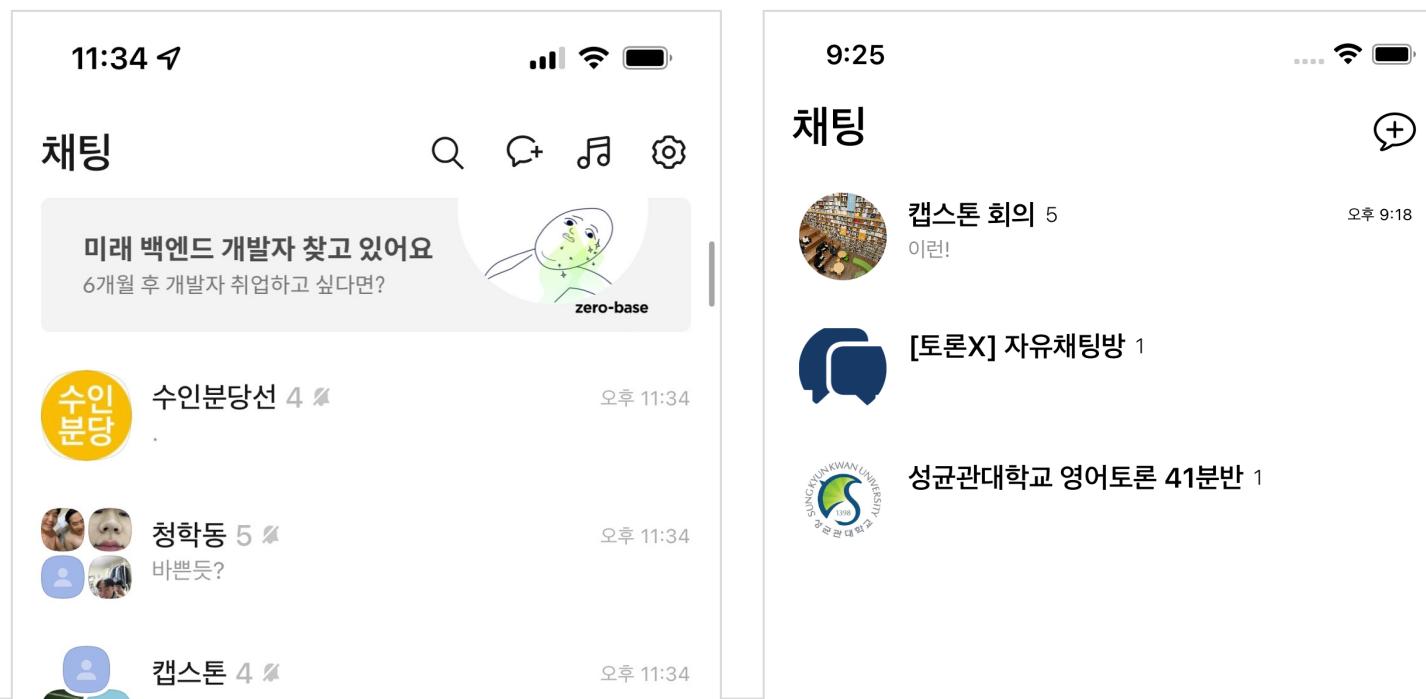
Front-End Development

A Team ⌂ ⓘ ⌚ 100% 🔋

✓ 채팅 기능 구현 - 채팅방

채팅방을 사용자가 추가하는 기능을 구현

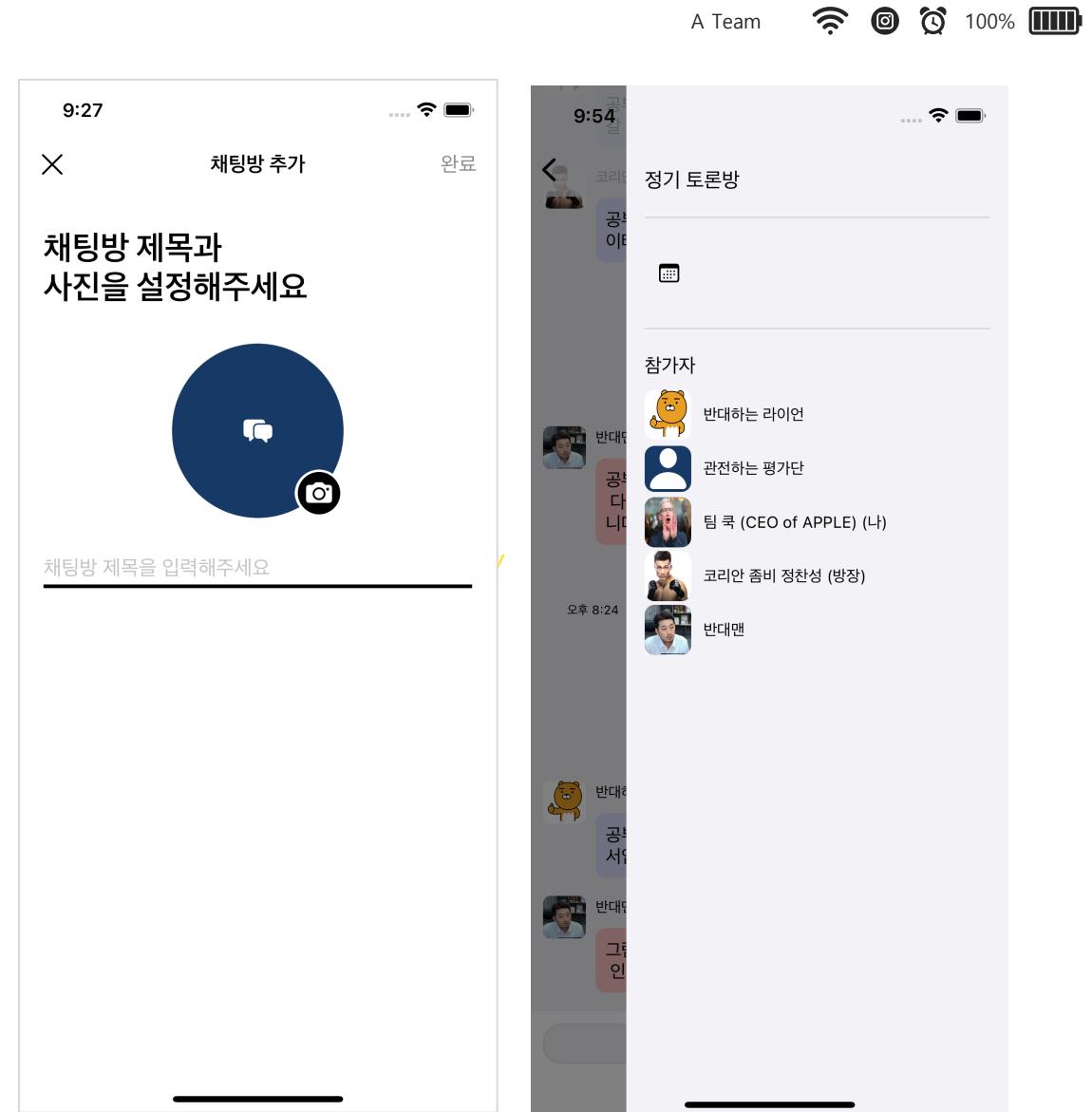
- 채팅방 프로필 사진도 사용자 프로필 사진과 동일한 방법으로 구현함
 - 전반적인 UI는 카카오톡을 많이 참고하여 깔끔하면서도 필요한 정보를 모두 담고자 했음



Front-End Development

✓ 채팅 기능 구현 - 채팅방

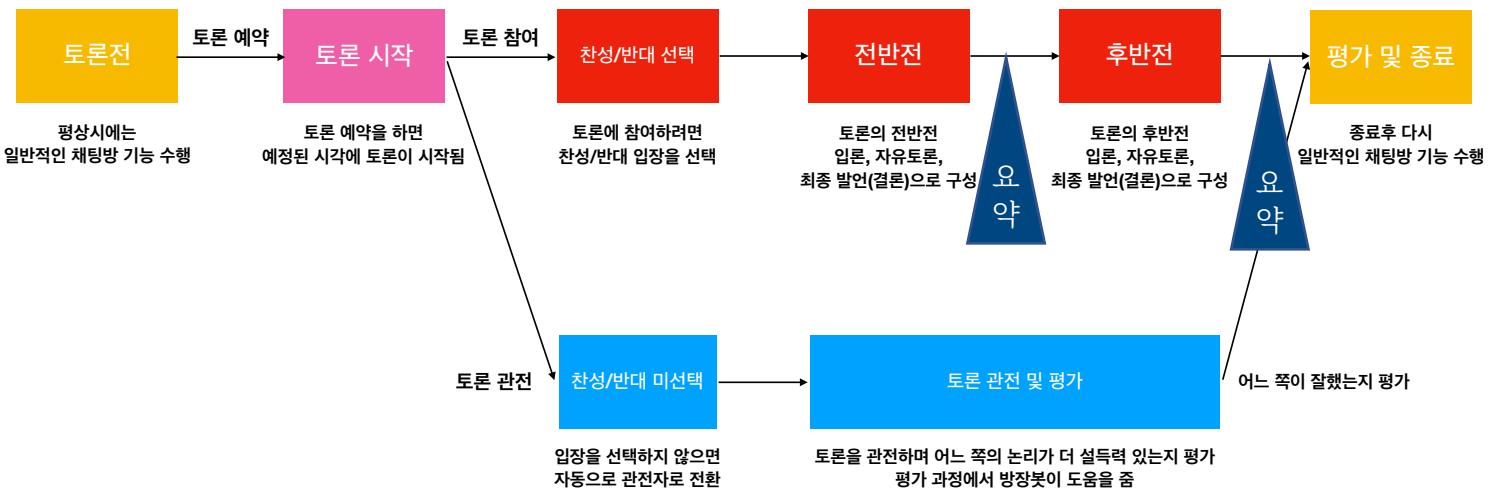
Example:



✓ 토론 기능 구현 - 최초 기획 완성

Challenges: 앱의 두 번째 핵심 기능인 토론 기능을 구현하기

- Firebase 실시간 데이터베이스에 토론의 상태 값을 나타내는 flag 값을 저장
- Flag를 추가하여 최초 기획한 전반전 & 후반전까지 구현 완료
- 기획대로 쉬는 시간 및 투표 시간 전에 참가자별 발언 요약까지 제공



Front-End Development

A Team ⚡ 100% 🔋

✓ 토론 기능 구현 - 최초 기획 완성

발언 요약:

- 진영별로, 참가자별로 요약해서 방장봇이 제공



방장봇

찬성측 "팀 쿡 The CEO of APPLE"님의
발언 요약입니다:

항상 거짓만 말하고 사는 존재였던
이 세상은 진작 멸망했을 것이므로
남을 속이는 것보다 거짓을 말하는게
상대에게 더 좋다면 거짓말을 하는
게 서로에게 이득입니다.

반대측 "계란조아"님의 발언 요약입니다:

성격이 우유 우유부단한 친구에게 "넌 참 성격이 좋구나"라고 하면 제발
결단력있게 살아라 라고 말을 해야
그 친구는 정신을 차릴 수 있고 지속
적인 칭찬으로 인해 그 청년은 본인
의 외모가 차운우 정도가 된다고 착
각하기 그려 서이이 까지마으 드 그



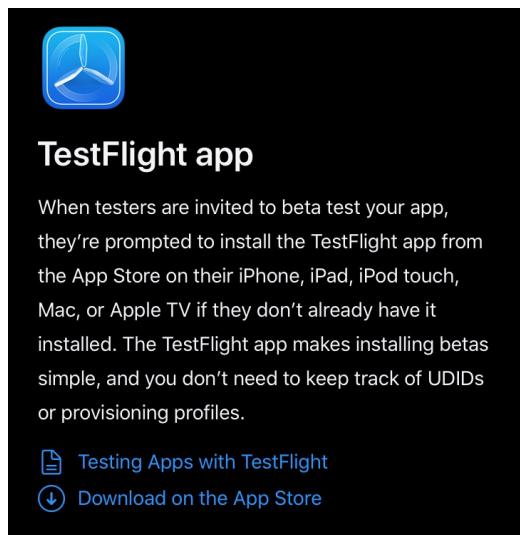
Front-End Development

A Team ⌂ ⌂ ⌂ 100% 🔋

✓ 사용성 개선 - 내부 테스팅을 통한 UI 개선

사용자에게 직접 보여지는 화면을 만든다는 Frontend 특성상 사용성을 높이기 위한 시도

- Private AppStore인 TestFlight를 통한 팀 내 배포 및 정기적인 테스트를 통해 사용성을 개선



The screenshot shows the Xcode Cloud TestFlight build dashboard for the project "방구석대한민국". The top navigation bar includes "App Store", "서비스", "TestFlight" (which is underlined), and "Xcode Cloud". The main section is titled "iOS 빌드" and contains a sub-section "버전 1.0.0". This section lists four build entries, each with a status of "제출 준비 완료" (Submitted ready) and a timestamp: "202205293", "202205292", "202205291", and "202205221". Each entry also includes a user icon and a group icon.

빌드	진행 상태	그룹	초대	설치	세션	총돌	피드백
202205293	제출 준비 완료 89일 후 무효화	👤	5	2	-	-	-
202205292	제출 준비 완료 88일 후 무효화	👤	5	4	-	-	-
202205291	제출 준비 완료 88일 후 무효화	👤	5	3	-	-	-
202205221	제출 준비 완료 81일 후 무효화	👤	5	5	-	-	1

Front-End Development

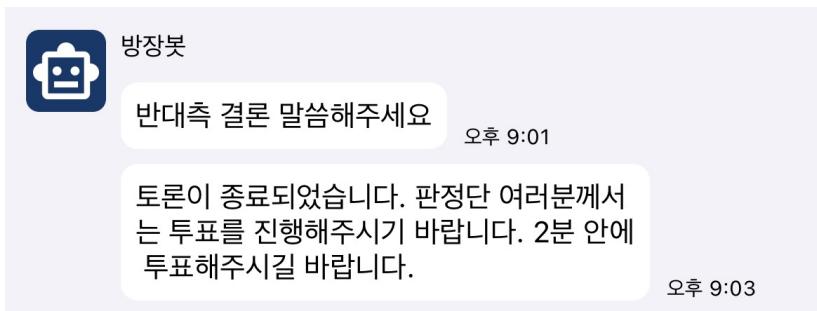
A Team ⌂ ⌂ ⌂ 100% 🔋

✓ 사용성 개선 - 내부 테스팅을 통한 UI 개선

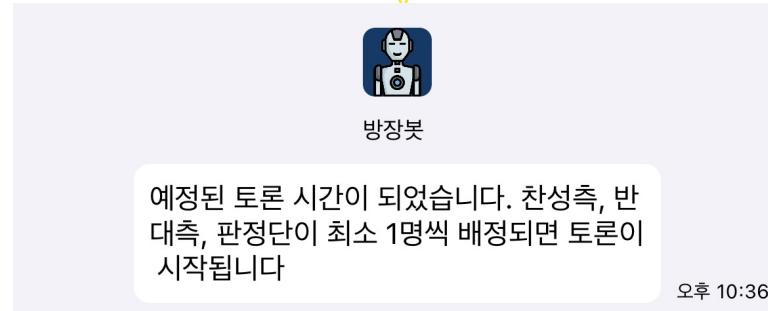
Feedback 예시 (1)

- “방장봇과 참가자 채팅을 구별하기 힘들다”👉 방장봇 채팅은 중앙에 두어 직관적으로 구분할 수 있게 변경

Before



After



Front-End Development

A Team

100%

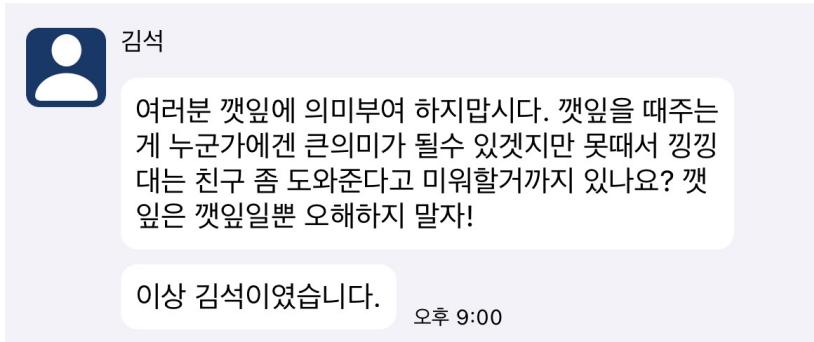
✓ 사용성 개선 - 내부 테스팅을 통한 UI 개선

Feedback 예시 (2)

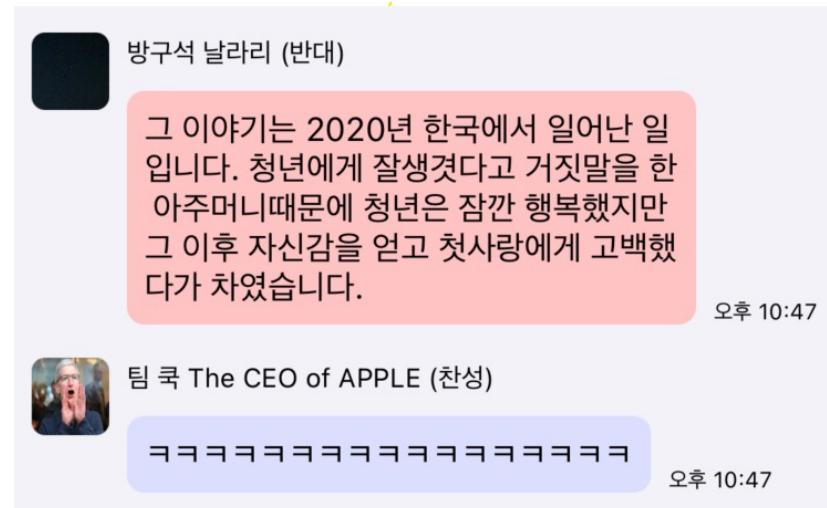
- “토론중 채팅이 찬성측 발언인지 반대측 발언인지 구분하기 힘들다”

👉 토론중 채팅에 한정하여 발언자의 진영에 따라 배경색을 바꾸고, 닉네임 옆에 진영을 표시하도록 변경

Before



After



✓ 사용성 개선 - HIG를 참고한 UI 개선

HIG란, Human Interface Guidelines의 줄임말로, Apple 플랫폼에서 동작하는 앱에 권장하는 UI 디자인 가이드라인을 말함.

The screenshot shows the Apple Developer website with the navigation bar at the top. The main content area is titled "Human Interface Guidelines". On the left, there's a sidebar with sections for iOS, macOS, tvOS, watchOS, and Technologies. The iOS section is expanded, showing sub-sections like Themes, Interface Essentials, App Architecture, User Interaction, System Capabilities, Visual Design, Icons and Images, Bars, Views, Controls, and Extensions. To the right of the sidebar, there are three devices (iPhone, iPad, and Mac) each displaying a grid of blue squares with dashed lines, representing a wireframe or layout. Below the devices, the text "iOS Design Themes" is visible.

✓ 사용성 개선 - HIG를 참고한 UI 개선

HIG를 적용한 예시 (1)

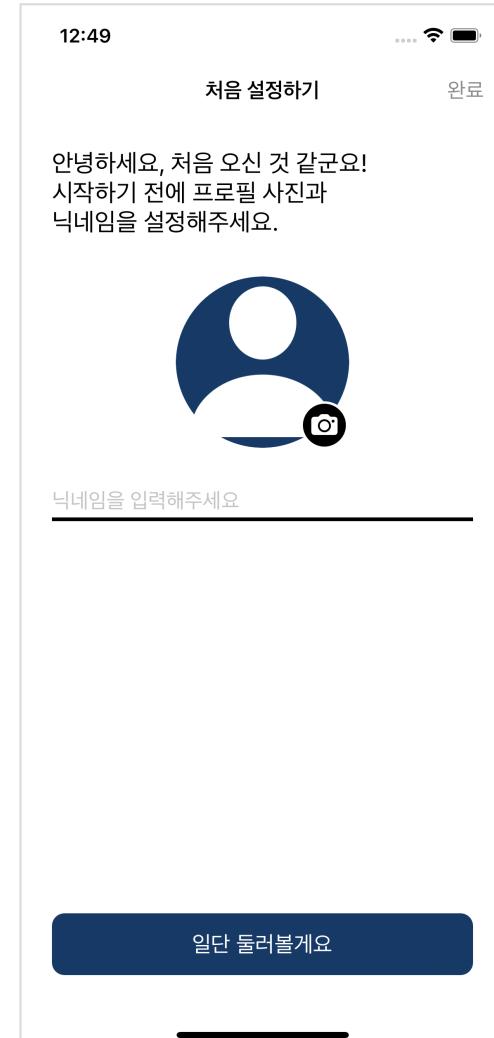
- Progress Indicator: 시간이 오래 걸리는 작업을 수행할 때 사용자에게 정적인 화면을 보여주지 말라.
- 사진을 업로드하는 작업은 시간이 오래 걸림👉 Progress Indicator가 돌아가도록 구현



✓ 사용성 개선 - HIG를 참고한 UI 개선

HIG를 적용한 예시 (2)

- 로그인 미루기: 로그인을 최대한 미루고 사람들이 앱을 둘러볼 수 있게 하라.
 - “*Give people a chance to appreciate your app before asking them to make a commitment to it.*”
- 로그인 기능은 없으나 앱을 처음으로 켜게 되면 비슷한 역할을 수행하는
프로필 설정 화면이 나옴
 - 우선 앱을 둘러보고 프로필을 설정할 수 있게 구현함
 - 출시 전 로그인을 구현하게 되면 동일한 흐름으로 동작할 것임



✓ 사용성 개선 - 접근성 Accessibility

접근성이란, 장애인, 고령자 등이 비장애인과 정보에 동등하게 접근하고 이해할 수 있도록 보장하는 것

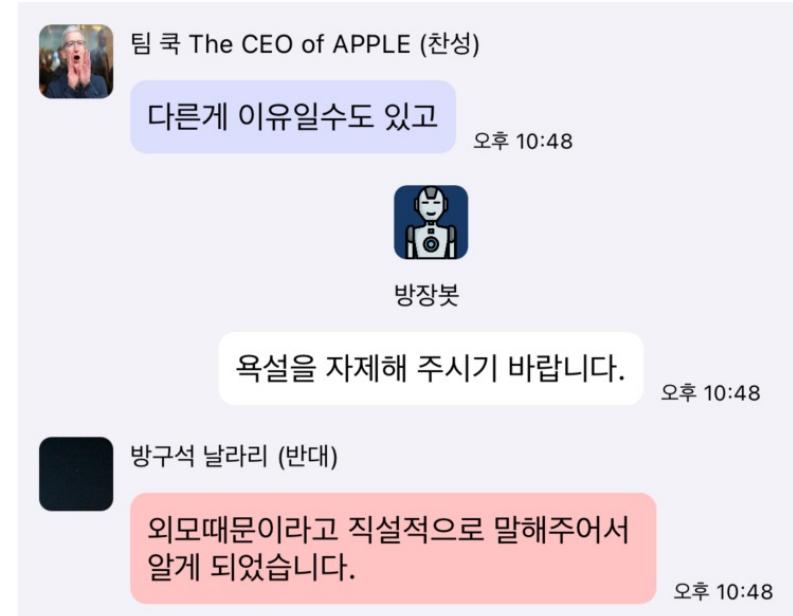
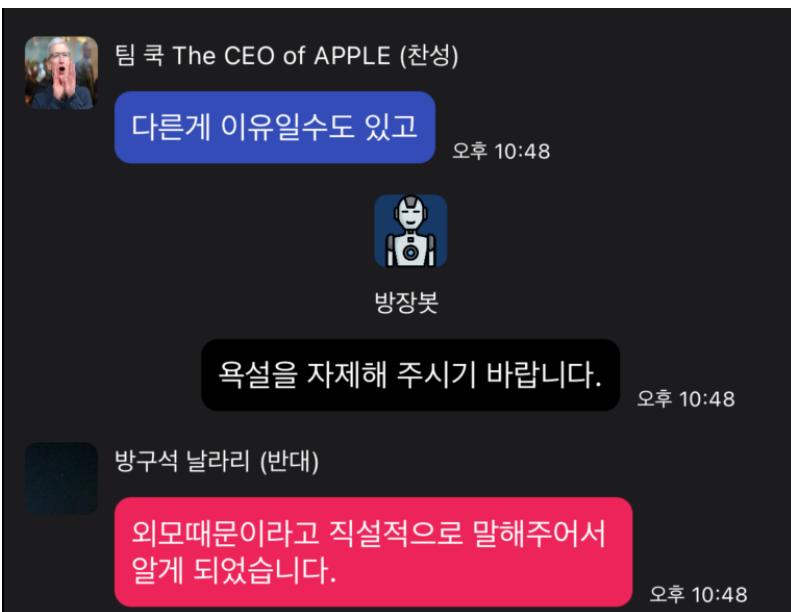
- 비장애인이라도 일시적으로 접근성이 떨어질 수 있음
 - 눈 관련 수술을 한 직후이거나, 운전중이거나, 조명이 어두운 상태 등
- Frontend 개발자에게 접근성 요소를 지원하는 것이 점점 중요해짐
- iOS에서 접근성 요소는 VoiceOver, Dark Mode 등이 있음



✓ 사용성 개선 - 접근성 Accessibility

접근성을 적용한 예시 (1)

- DarkMode: 어두운 환경의 UI를 말하며 밝은 빛에 예민한 사용자들, 조명이 어두운 환경에서의 사용자들의 접근성을 높일 수 있음



✓ 사용성 개선 - 접근성 Accessibility

접근성을 적용한 예시 (2)

- VoiceOver: 화면을 음성 비서 Siri가 읽어주는 기능을 말하며 저시력자 및 시각 장애인 사용자들의 접근성을 높일 수 있음
 - VoiceOver를 켜게 되면 조작이 일반적인 제스쳐와 완전히 다르게 동작함

✓ 사용성 개선 - 접근성 Accessibility

접근성을 적용한 예시 (2)

- VoiceOver 예시: Demo에서...

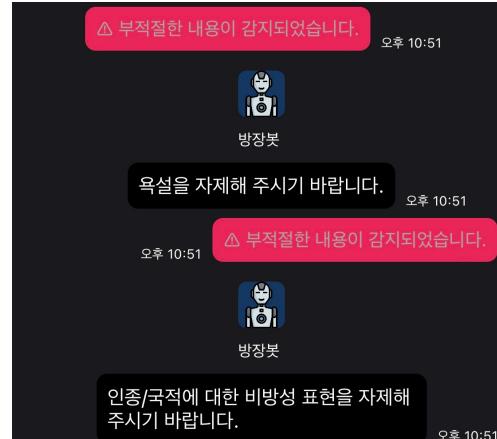
Limitation

A Team ⌂ ⓘ ⌚ 100% 🔋

✓ 모델의 한계 & 개선할 점

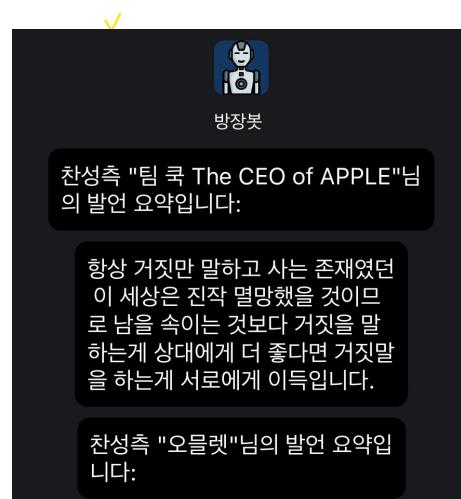
1. 비방성 표현 마스킹 모델

hard voting을 적용 하였는데도, 욕설이 아닌데
욕설이라고 판단하는 경우가 존재했다. 로직을 고치고,
더 많은 데이터를 추가하여 성능을 개선할 계획이다.



2. 대화 요약 모델

정제되고 잘 짜여진 글을 요약하는 것이 아닌,
토론 방에서 보내는 메세지들을 모아서
요약하기 때문에, 어색하게 요약하는 경우가 있다.



데모 시연

