

# 시스템 로그 분석을 도입한 욕설 감지 프로그램

팀 인절미

지도교수 : 양희경 교수님

Youtube 영상 링크 : [https://www.youtube.com/watch?v=-srPIL6W064&ab\\_channel=%E1%84%86](https://www.youtube.com/watch?v=-srPIL6W064&ab_channel=%E1%84%86)

# 목차

- 서비스 소개
- 필요성
- 프로토타입
- 개발 전략 : 데이터 추출 & 알고리즘
- 사업화 가능성

# 서비스

- 서비스 이름 : '착하게 말해요'
- 기존 서비스 - 채팅에서 욕설 단어를 감지하여 욕설 필터링
- 제안하는 서비스 - 채팅에서 욕설 단어를 감지하여 욕설 필터링 + 맥락적 단서를 기반으로 필터링 기능 개선

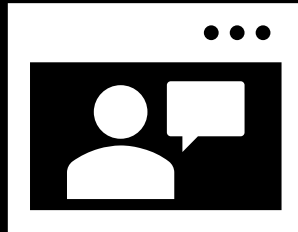
# 서비스

- 제안하는 서비스 - 채팅에서 욕설 단어를 감지하여 욕설 필터링 + 맥락적 단서를 기반으로 필터링 기능 개선

case01 : 시스템 로그

case02 : 타임 스탬프

case03 : 플레이 내적인 요소



시스템 로그

(단서) 시스템 로그 자체  
(선행 조건) 차단, 신고 등에 대한 로그가 채팅 로그에 표시되도록 해야함

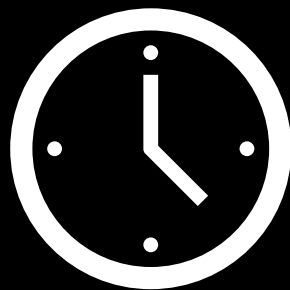
# 서비스

- 제안하는 서비스 - 채팅에서 욕설 단어를 감지하여 욕설 필터링 + 맥락적 단서를 기반으로 필터링 기능 개선

case01 : 시스템 로그

case02 : 타임 스탬프

case03 : 플레이 내적인 요소



타임 스탬프

(단서) 채팅 트래픽이 급증하는데 그 대화에 욕설이 포함되는 경우가 많다.  
(선행 조건) 타임 스탬프와 관련한 학습 전략 요구됨

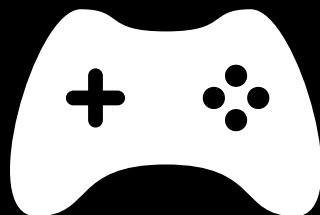
# 서비스

- 제안하는 서비스 - 채팅에서 욕설 단어를 감지하여 욕설 필터링 + 맥락적 단서를 기반으로 필터링 기능 개선

case01 : 시스템 로그

case02 : 타임 스탬프

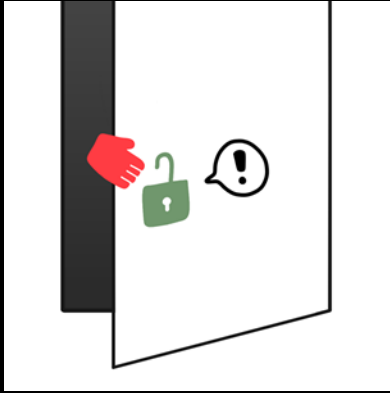
case03 : 플레이 내적인 요소



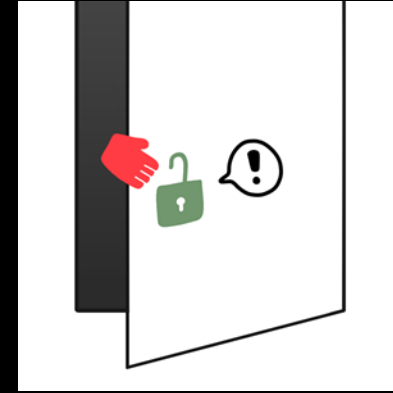
플레이 내적인 요소

플레이 내적인 요소를 채팅 로그에 시스템 로그 형태로 찍히도록 한다.

# 사용자 스토리



실시간으로 나쁜 언행을 막아주는 프로그램



(로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램

# 사용자 스토리

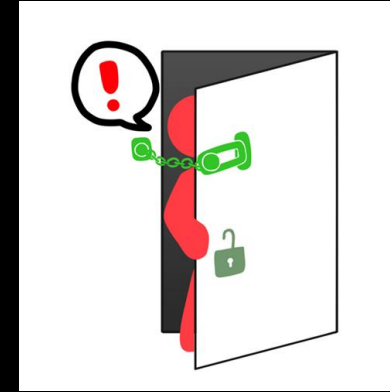


실시간으로 나쁜 언행을 막아주는 프로그램

## 기존 보안 장치

기존 비속어 필터 시스템

- 1) 직접 작성: 욕설
- 2) 특수 문자: 욕@설 / 욕 " 설 / 욕!설 / 욕1설
- 3) 단타: 욕  
설
- 4) 욕설욕설
- 5) 우회 및 변형: ㅇㅈㅅㅈㅅㅈㅅ



(로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램

## 기존 보안 장치

기존 비속어 필터 시스템

- 1) 직접 작성: 욕설
- 2) 특수 문자: 욕@설 / 욕 " 설 / 욕!설 / 욕1설
- 3) 단타: 욕  
설
- 4) 욕설욕설
- 5) 우회 및 변형: ㅇㅈㅅㅈㅅㅈㅅ



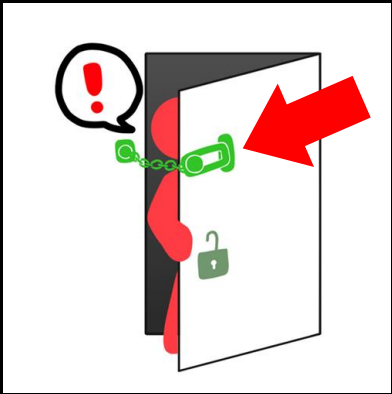
## 2중 보안 장치

추가하고자 하는 기능 :

채팅이 이뤄지고 난 후의  
전후 상황을 파악



# 개선안



## 2중 보안 장치

추가하고자 하는 기능 :

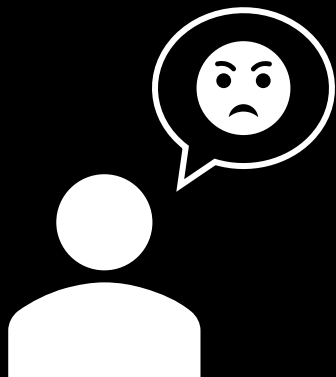
1. 특정 어휘를 사용한 이용자가 다른 유저에게 차단을 당하는 경우 (신고를 당한 경우)
2. 특정 어휘가 사용된 이후 채팅량 급증
3. 특정 행동을 유도 (어떤 시스템에 사용될지\_이식성 문제 해결 필요)
4. 해당 어휘를 본 다른 사람이 채팅방에서 퇴장하는 경우
5. 특정 어휘를 사용한 사람이 강퇴를 당하는 경우

# 필요성

- 성인에 비해서 판단력이 부족한 어린이나 청소년이 무분별하게 콘텐츠와 사람들의 반응에 노출되는 걸 막을 수 있다.
- 은어와 비속어는 변형이 쉽고 유행에 따라 등장하고 사라지는 주기가 짧기 때문에  
직접 단어를 추가해서 관리하는 방식은 장기적인 관점에서 어려움에 처할 가능성이 높아진다.
- 텍스트(채팅)만을 이용해서 필터링하는 것은 그 한계가 뚜렷하다.
- 우리가 사용할 수 있는 채팅 외적인 상황을 파악하면 이를 더 효과적으로 막아낼 수 있을 것이다.

# 주안점

- 궁극적인 목표: 부적절한 사용자를 빠짐없이 잡아내는 것



억울하게 채팅 권한을 상실하는 이용자가 발생할 수 있다.



그로 인해 억울한 사람이 발생하더라도 강하게 대처하자.

# 주안점

- 궁극적인 목표: 부적절한 사용자를 빠짐없이 잡아내는 것



억울한 제재를 점차 줄여나가기 위해서  
욕설하는 사람들이 발생했을 때 전후 상황을 정확하게 파악한다.

# 주안점

- 궁극적인 목표: 부적절한 사용자를 빠짐없이 잡아내는 것



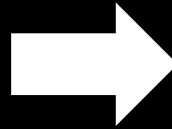
이러한 운영 방침이 장기적으로 유지되면 불만을 품는 이용자가 많이 생길 수 있기 때문에  
억울하게 불이익을 받는 이용자를 줄이기 위해서 다양한 단서를 활용하도록  
잘 디자인하고 학습시켜야 하며 꾸준히 유지보수를 할 필요가 있다.

# 개발 전략\_학습 데이터 추출

예시) LoL

기존 시스템  
오롯이 채팅만을 이용해서 비속어 탐지

```
**** (애니비아) : 야  
**** (애니비아) : 블루 먹지말라고  
**** (애니비아) : 청  
**** (애니비아) : 포  
**** (애니비아) : 도야  
**** (티모) : 왜  
**** (애니비아) : 저 파  
**** (애니비아) : 인  
**** (애니비아) : 애  
**** (애니비아) : 1플이  
**** (애니비아) : 아  
**** (티모) : 버섯깎아야대
```



제안하는 시스템  
타임스탬프를 적극적으로 활용

```
[05:01] **** (애니비아) : 야  
[05:02] **** (애니비아) : 블루 먹지말라고  
[05:03] **** (애니비아) : 청  
[05:03] **** (애니비아) : 포  
[05:04] **** (애니비아) : 도야  
[05:06] **** (티모) : 왜  
[05:08] **** (애니비아) : 저 파  
[05:09] **** (애니비아) : 인  
[05:09] **** (애니비아) : 애  
[05:11] **** (애니비아) : 1플이  
[05:13] **** (애니비아) : 아  
[05:17] **** (티모) : 버섯깎아야대
```

# 개발 전략\_학습 데이터 추출

예시) LoL

기존 시스템  
오롯이 채팅만을 이용해서 비속어 탐지

채팅로그

[12:25] : 킹카림이다  
[13:45] : 한심한코끼리님  
[13:47] : 진짜  
[13:48] : 무지성으로  
[13:50] : 가는걸보니  
[13:53] : 한심하시긴하네요  
[19:13] : 어디가누  
[21:01] : 잘가고  
[21:02] : 청  
[21:02] : 포도  
[21:05] : 한심한  
[21:06] : 코끼리

제안하는 시스템  
차단, 신고 내역을 시스템 로그에 추가하여 활용

채팅로그

[12:25] : 킹카림이다  
[13:45] : 한심한코끼리님  
[13:47] : 진짜  
[13:48] : 무지성으로  
[13:50] : 가는걸보니  
[13:53] : 한심하시긴하네요  
[13:58] : '한심한코끼리'님이 '비매너 사용자'님을 차단하셨습니다.  
[19:13] : 어디가누  
[21:01] : 잘가고  
[21:02] : 청  
[21:02] : 포도  
[21:05] : 한심한  
[21:06] : 코끼리

# 개발 전략 \_ 학습 데이터 추출

- 데이터 수집 :

**채팅 로그 분석**



**시스템 로그 (차단 횟수, 신고 횟수 등)  
또한 feature 중 하나로 설정**

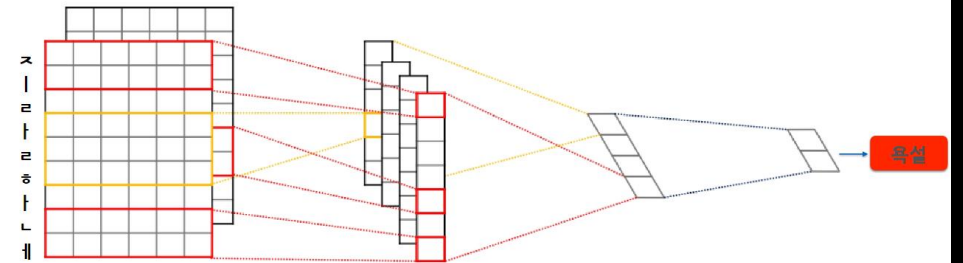


# 개발 전략

## 자연어 처리 + CNN

기존의 욕설 탐지 알고리즘 + 시스템 로그 분석

1. 입력 2. Embedding 3. Convolution 4. Pooling 5. 출력 레이어



넥슨 코리아 인텔리전스랩스 어뷰징탐지팀 - '딥러닝으로 욕설 탐지하기'

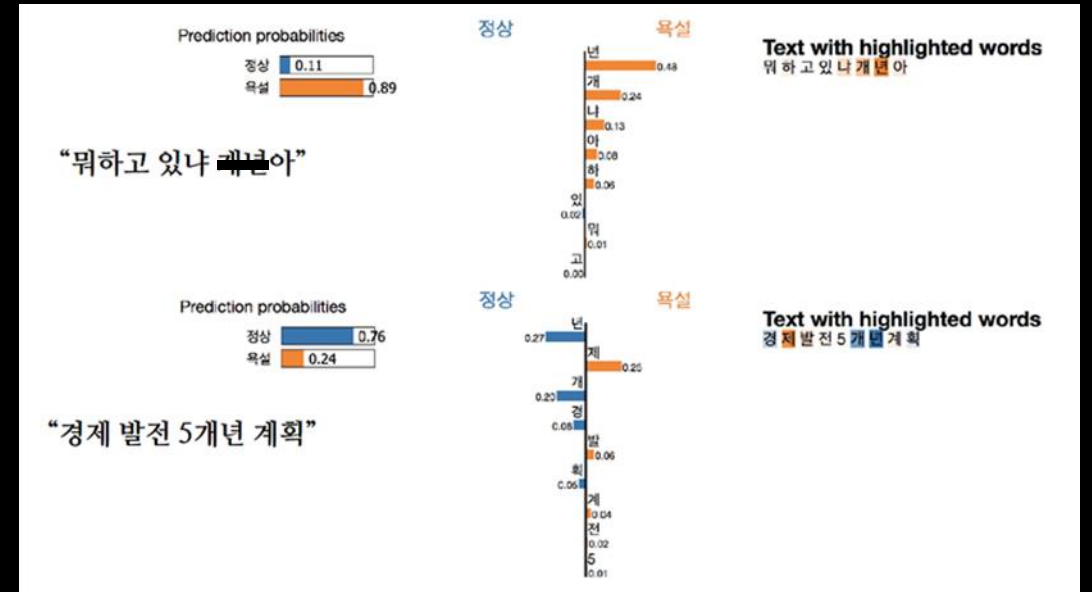
Kim, Yoon. "convolutional neural networks for sentence classification." (2014)

인벤. "넥슨 코리아 인텔리전스랩스 어뷰징탐지팀 - '딥러닝으로 욕설 탐지하기'". <https://www.inven.co.kr/webzine/news/?news=198156>. (2021.11.16)

# 개발 전략

## 자연어 처리 + CNN

기존의 욕설 탐지 알고리즘 + 시스템 로그 분석



넥슨 코리아 인텔리전스랩스 어뷰징탐지팀 - '딥러닝으로 욕설 탐지하기'

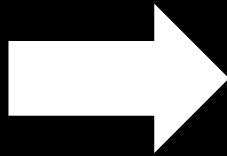
Kim, Yoon. "convolutional neural networks for sentence classification."(2014)

인벤. "넥슨 코리아 인텔리전스랩스 어뷰징탐지팀 - '딥러닝으로 욕설 탐지하기'". <https://www.inven.co.kr/webzine/news/?news=198156>. (2021.11.16)

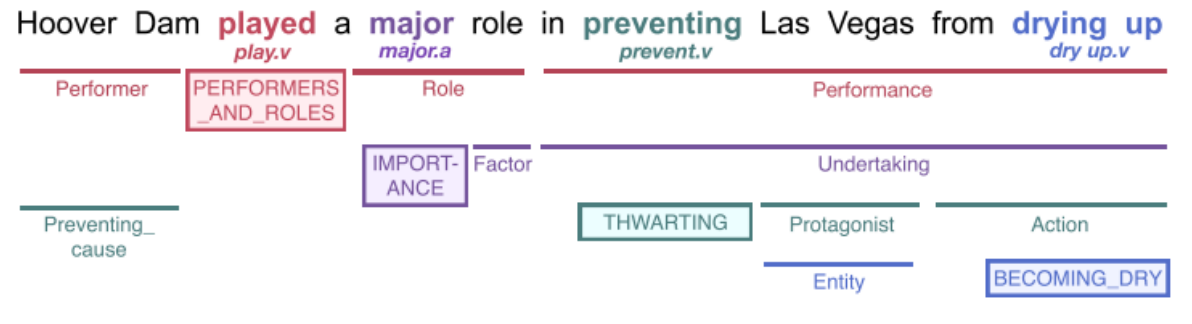
# 개발 전략

- RNN이 아닌 CNN을 택한 이유

CNN



이미지 영상

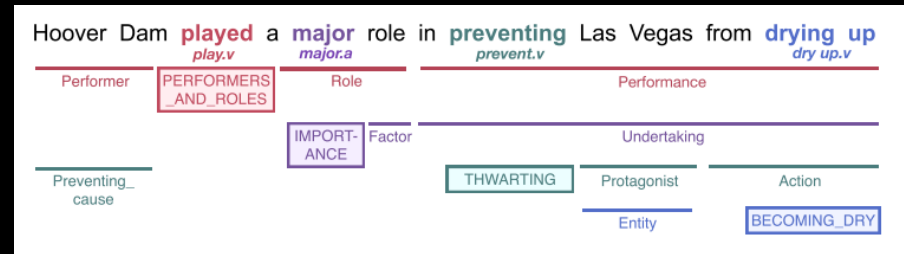


# 개발 전략

- RNN이 아닌 CNN을 택한 이유



이미지 & 영상  
2D Convolutionals



자연어 처리  
1D Convolutional

# 개발 전략

## • 모델 설명

$$c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b).$$

비선형 function 을 통한 필터

필터 사용

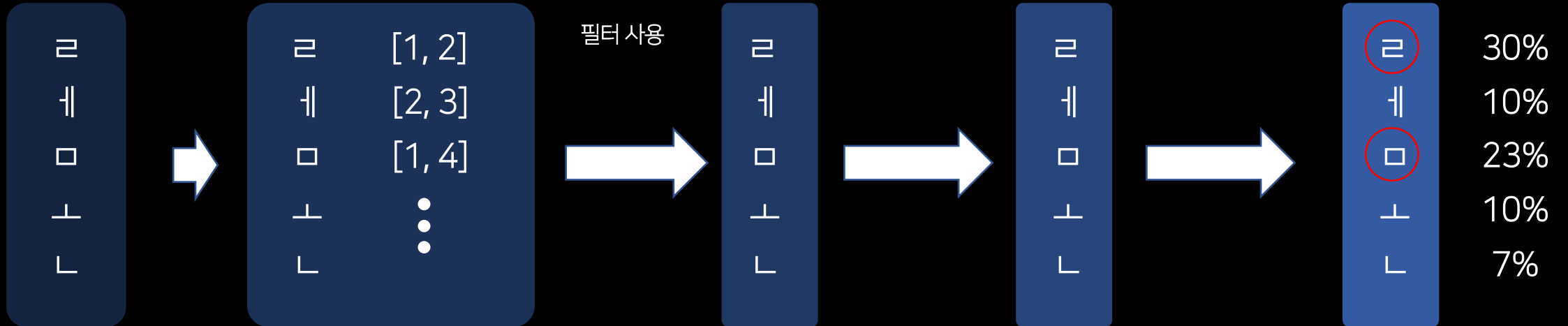
1. 입력 :  
자모음을 나눠서 입력  
- 인간의 인식과 데이터가  
다르기 때문에

2. Embedding :  
각 자모음에 벡터 수치 부여  
ex) ㄱ = [1, 2]  
Word Vector 생성

3. Convolution :  
필터를 통해 Feature를 추출하여  
Feature map 생성

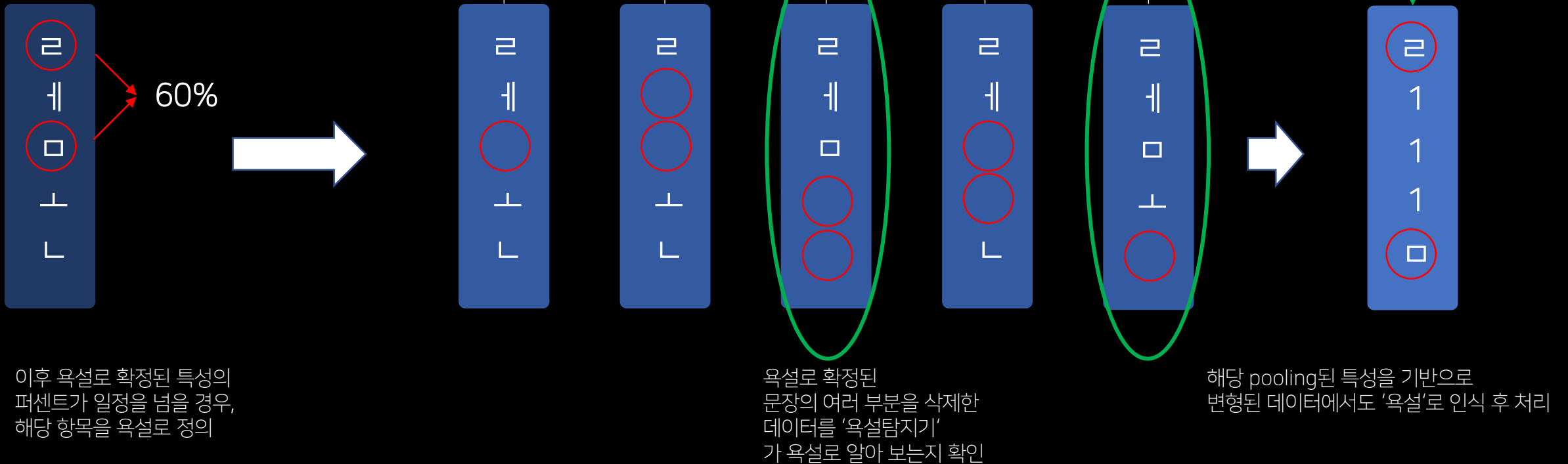
4. Pooling :  
가장 필요한 Feature제외  
나머지 버림

5. 해당 특성으로  
욕설일 확률 계산  
ex) 'ㄱ'과 'ㄴ'이 연속적으로  
사용 될 경우  
욕설일 확률이 높음



# 개발 전략

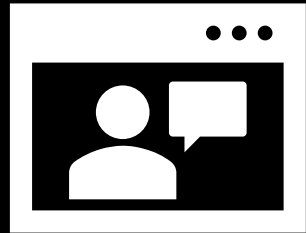
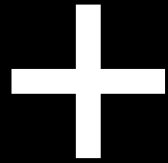
ex) 욕설로 인지 사유 : 연속된 자음 'ㄹ'과 'ㄹ'



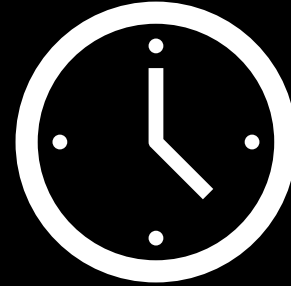
# 개발 전략



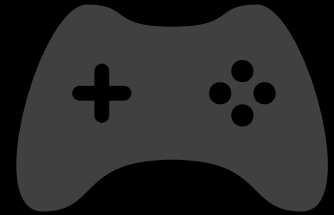
기존 욕설탐지 시스템



시스템 로그



타임 스탬프



플레이 내적인 요소

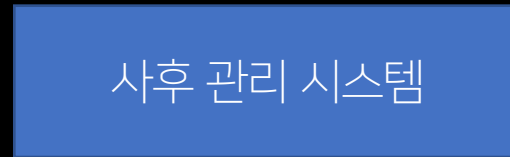
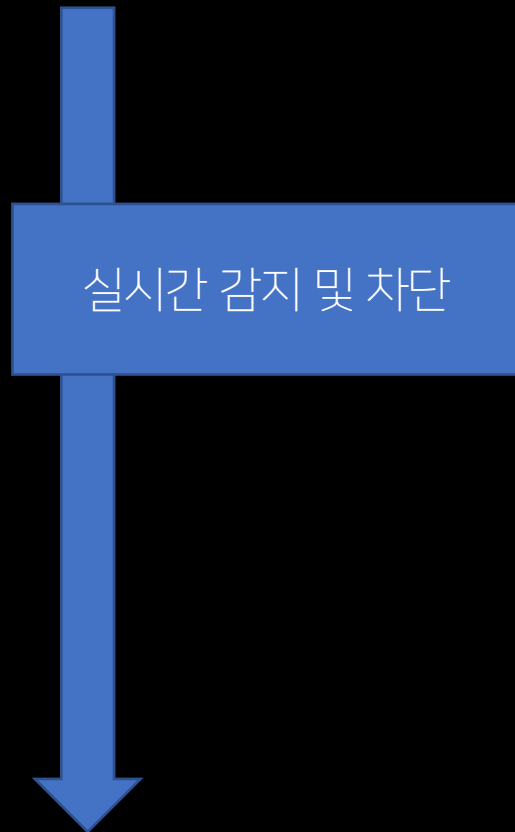
# 프로토타입

실시간 감지 및 차단

사후 관리 시스템



# 프로토타입



# 프로토타입



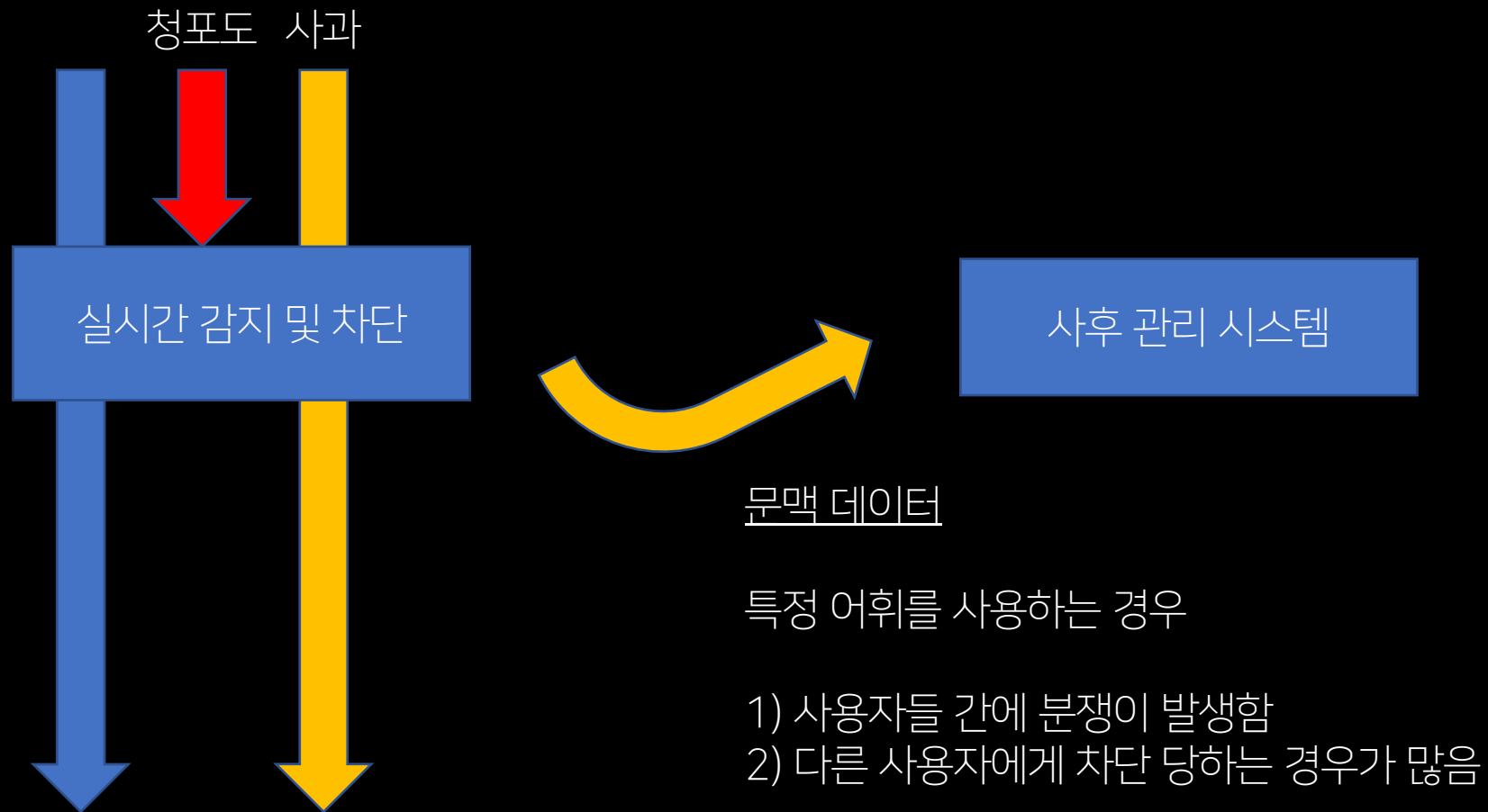
사후 관리 시스템

# 프로토타입

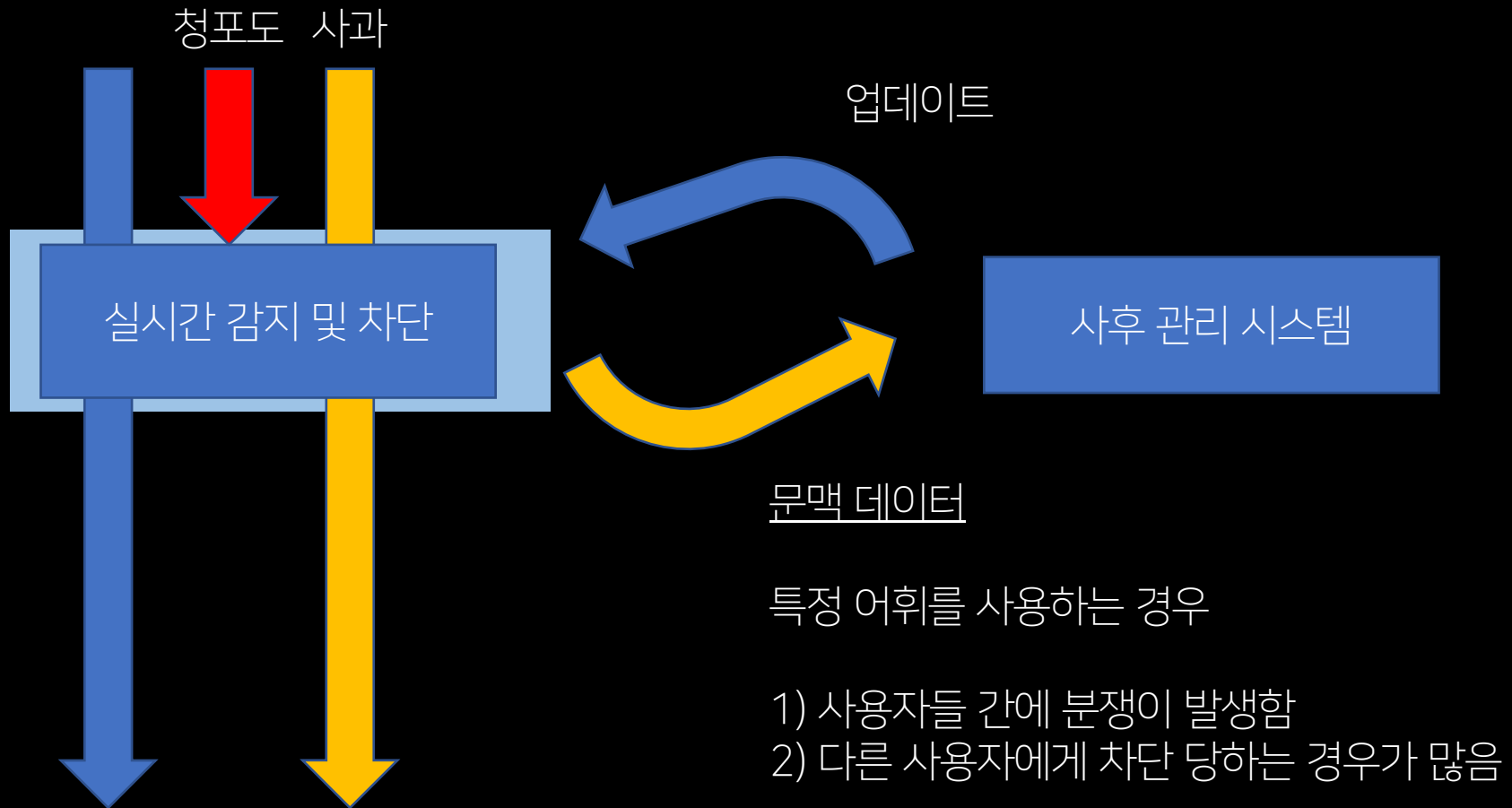


사후 관리 시스템

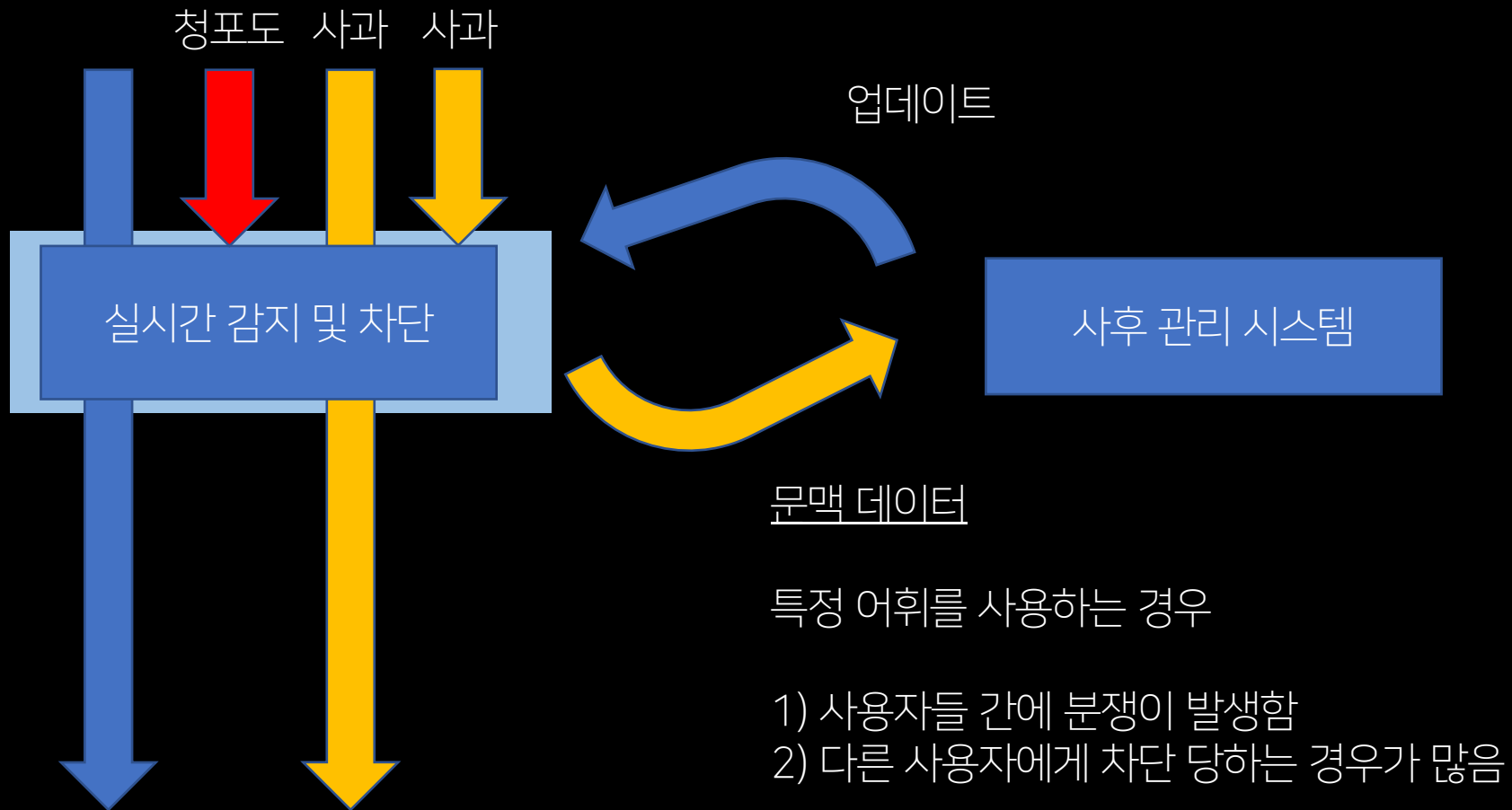
# 프로토타입



# 프로토타입



# 프로토타입



# 사업화 가능성

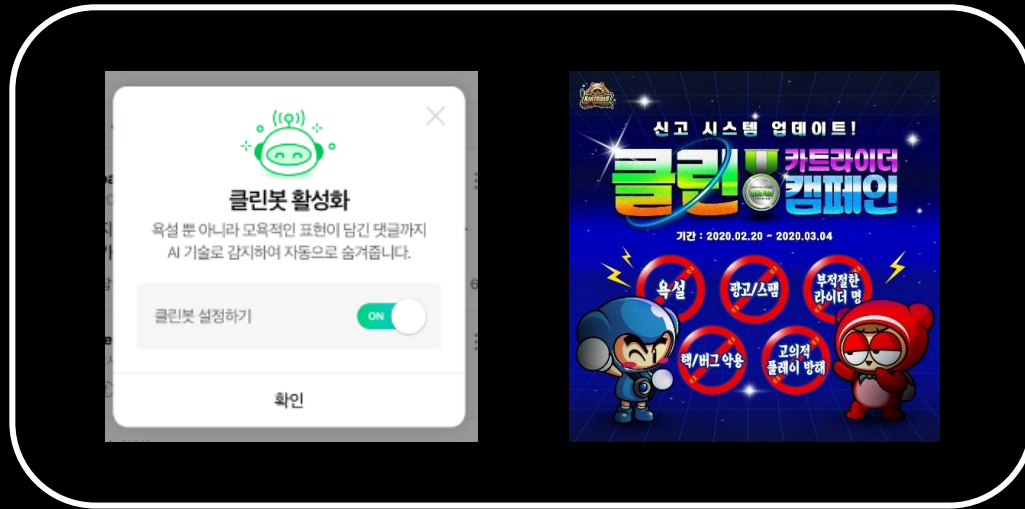
- 다른 서비스에 덧붙여 사용할 수 있는 이식성이 높은 프로그램
- 해당 프로그램을 기간제로 판매하여 이용기간에 따라 사용료를 지급하게함.
- 키즈 프로그램에 쓸 수 있을 정도로 깨끗한 채팅 환경을 원하는

관리자들에게 판매 가능

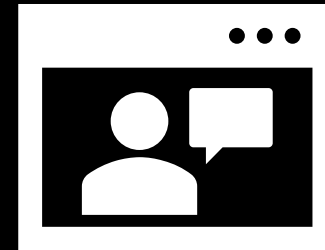
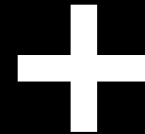
ex) 실시간 방송, 게임, 전 연령에게 노출될 수 있는 방송국·라디오 댓글

# 사업화 가능성

- 다른 서비스에 덧붙여 사용할 수 있는 이식성이 높은 프로그램



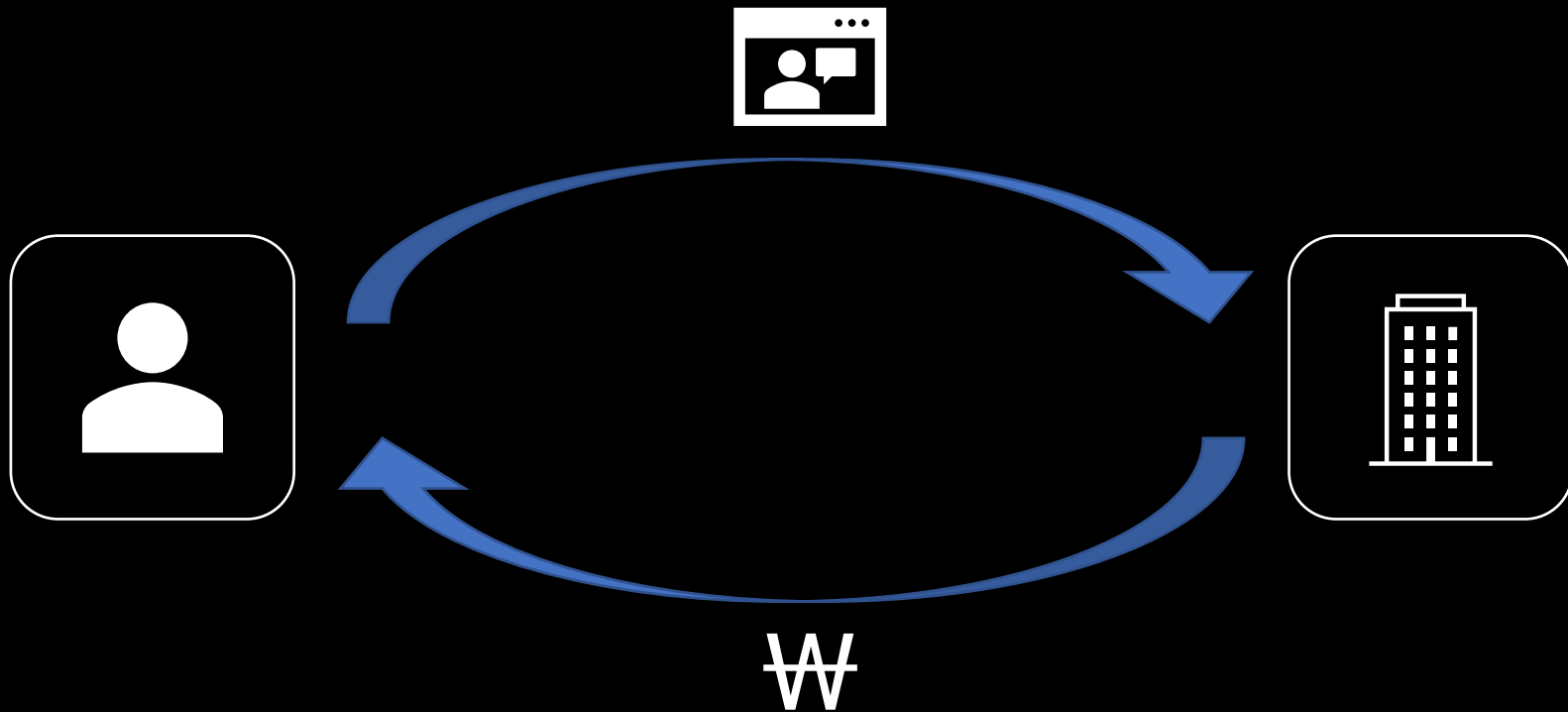
기존 프로그램





# 사업화 가능성

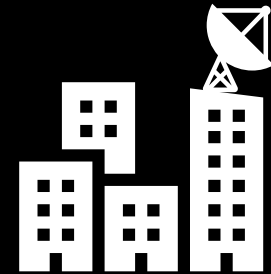
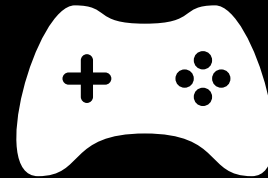
- 해당 프로그램을 기간제로 판매하여 이용기간에 따라 사용료를 지급하게함.



# 사업화 가능성

- 키즈 프로그램에 쓸 수 있을 정도로 깨끗한 채팅 환경을 원하는 관리자들에게 판매 가능

ex) 실시간 방송, 게임, 전 연령에게 노출될 수 있는 방송국·라디오 댓글



감사합니다!