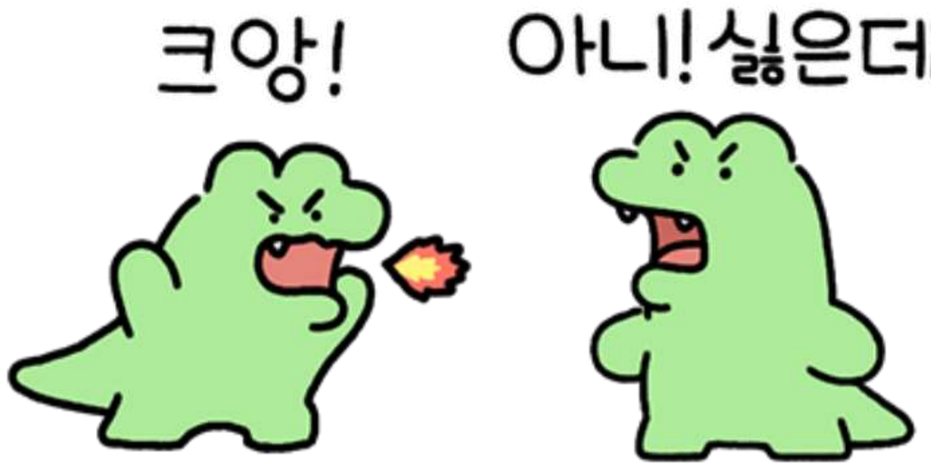
사용자 맞춤형 금지어 필터링을 적용한 실시간 온라인 채팅 시스템

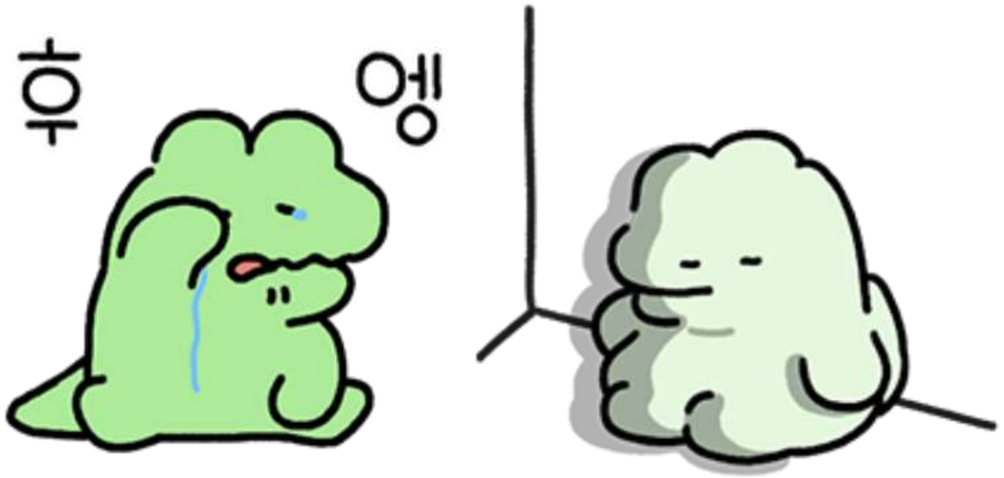
숭실대학교 소프트웨어학부
팀 : 창호공들

목 차

개발 배경

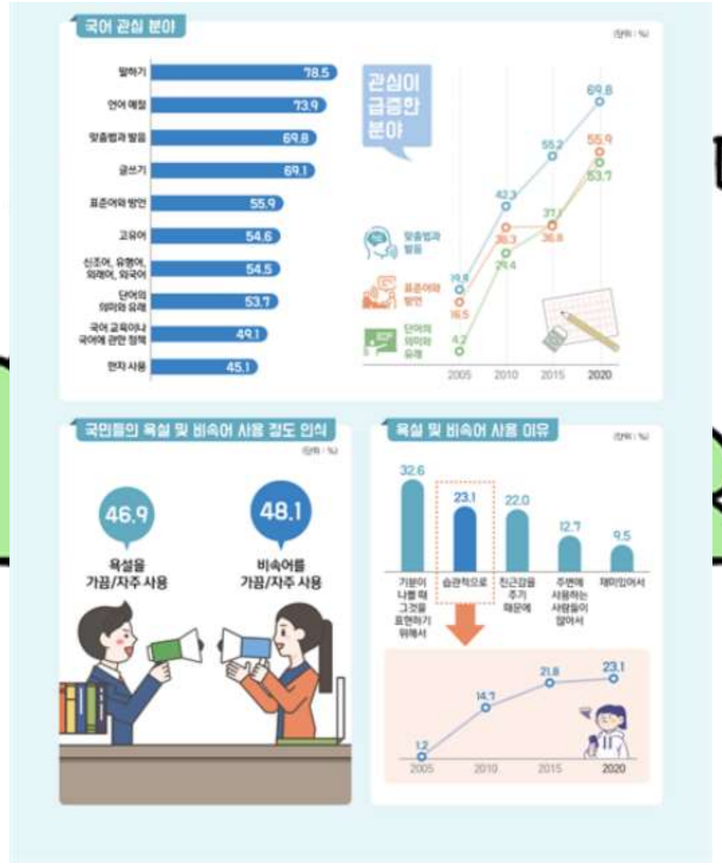


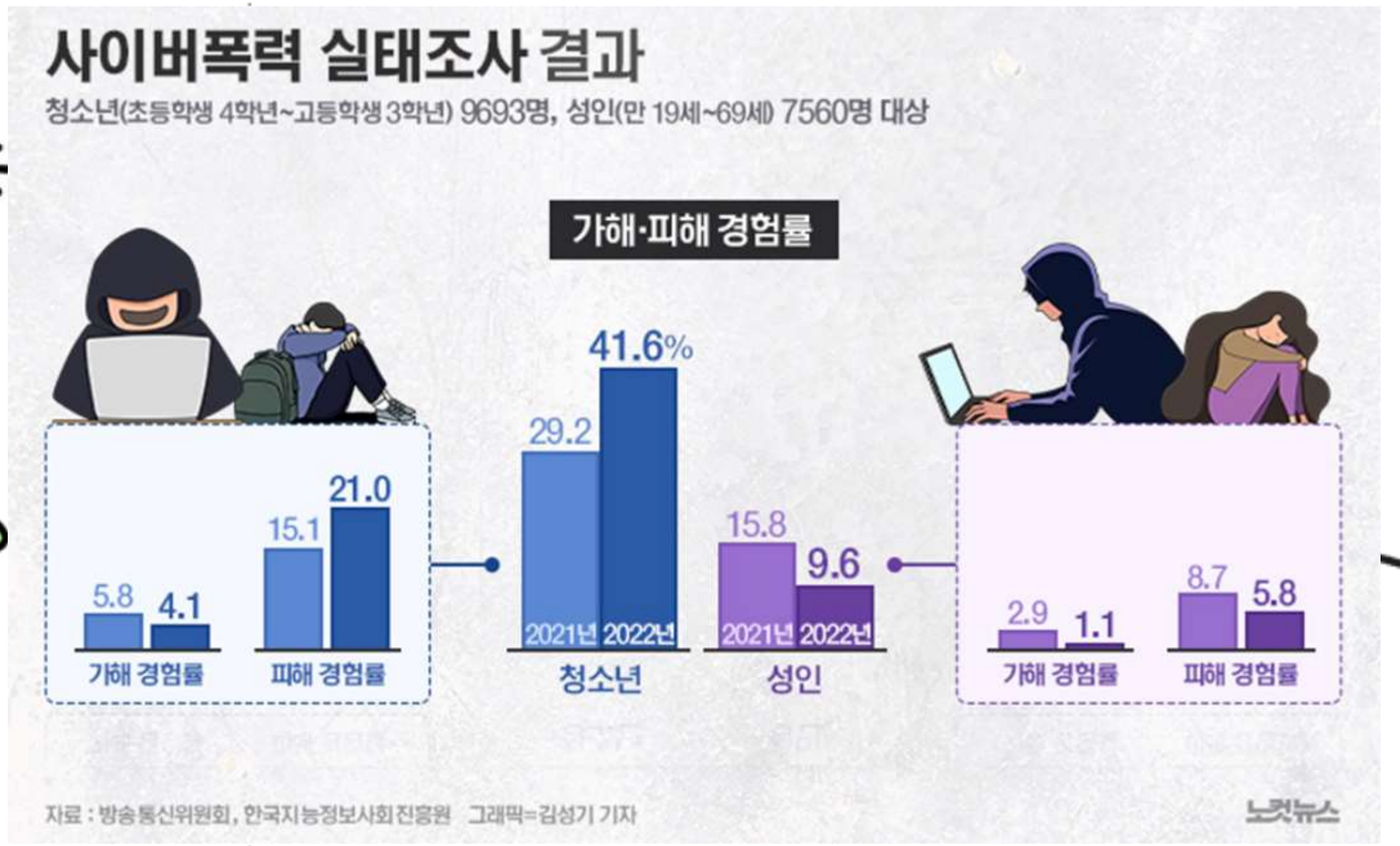
욕설 ⇧



사회적 문제 ⇧

개발 배경


국어 관심 분야
발하기 78.5
언어 예절 73.9
맞춤법과 발음 69.8
글쓰기 69.1
표준어와 방언 55.9
고유어 54.6
신조어, 유행어, 외래어, 외국어 54.5
단어의 의미와 유래 53.7
국어 교육이나 국어에 관한 정책 49.1
한자 사용 45.1
관심이 급증한 분야
맞춤법과 발음
표준어와 방언
단어의 의미와 유래
2005 2010 2015 2020
국민들의 욕설 및 비속어 사용 정도 인식
46.9
욕설을 가끔/자주 사용
48.1
비속어를 가끔/자주 사용
욕설 및 비속어 사용 이유
32.6
23.1
22.0
12.7
9.5
기분이 나쁠 때 그것을 표현하기 위해서
습관적으로
친근감을 주기 때문에
주변에 사용하는 사람들이 많아서
재미있어서
2005 2010 2015 2020


사이버폭력 실태조사 결과
청소년(초등학생 4학년~고등학생 3학년) 9693명, 성인(만 19세~69세) 7560명 대상
가해·피해 경험률
5.8
4.1
가해 경험률
15.1
21.0
피해 경험률
29.2
41.6%
2021년 2022년
청소년
15.8
9.6
2021년 2022년
성인
2.9
1.1
가해 경험률
8.7
5.8
피해 경험률
자료 : 방송통신위원회, 한국지능정보사회진흥원 그래픽=김성기 기자
노컷뉴스

개발 목표

- 욕설, 비속어 금지어 목록 생성
- 금지어 목록의 욕설, 비속어 실시간 필터링

-> 온라인 커뮤니케이션 환경 개선 & 언어 사용에 대한 인식 개선

관련 연구

1. 정규 표현식 기반 필터링
2. 사전 기반 키워드 필터링
3. 딥러닝 기반 메시지 필터링

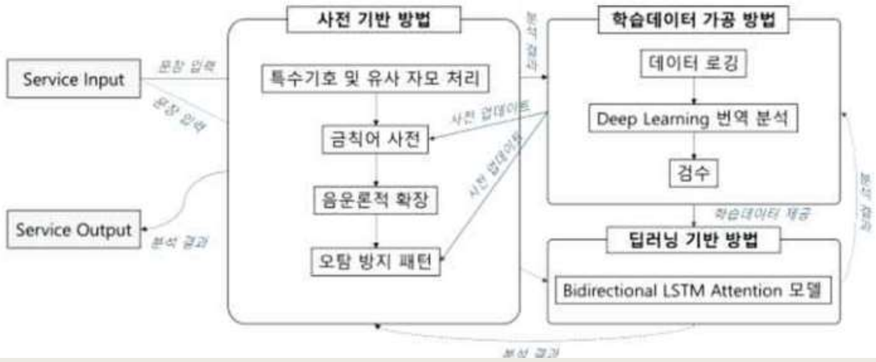
1. 정규 표현식 기반 필터링

- 미리 정의된 패턴을 기반으로 특정 단어 탐지
- 장점 : 한 개의 단어를 가지고 유사한 패턴의 단어를 필터링 가능
- 단점 : "놈" -> "ㄴㅗㅁ"과 같이 변형 시 필터링 불가

관련 연구

2. 사전 기반 키워드 필터링

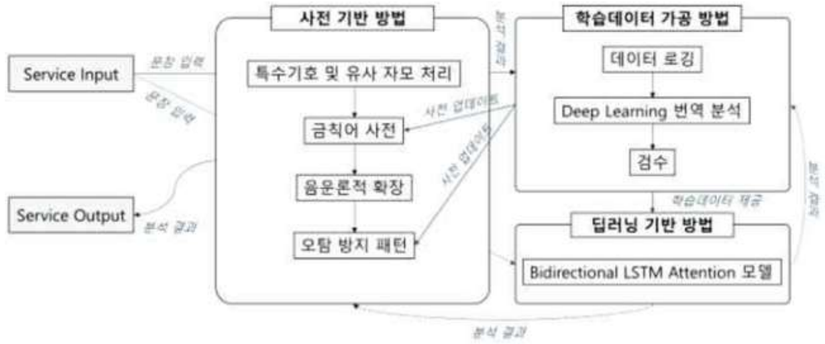
- 금지어 목록을 설정, 메시지에서 해당 단어가 포함되는지 검사
- 장점 : 저장된 금지어에 대한 관리가 용이함
- 단점 : 추가되지 않은 단어에 대한 필터링이 불가함



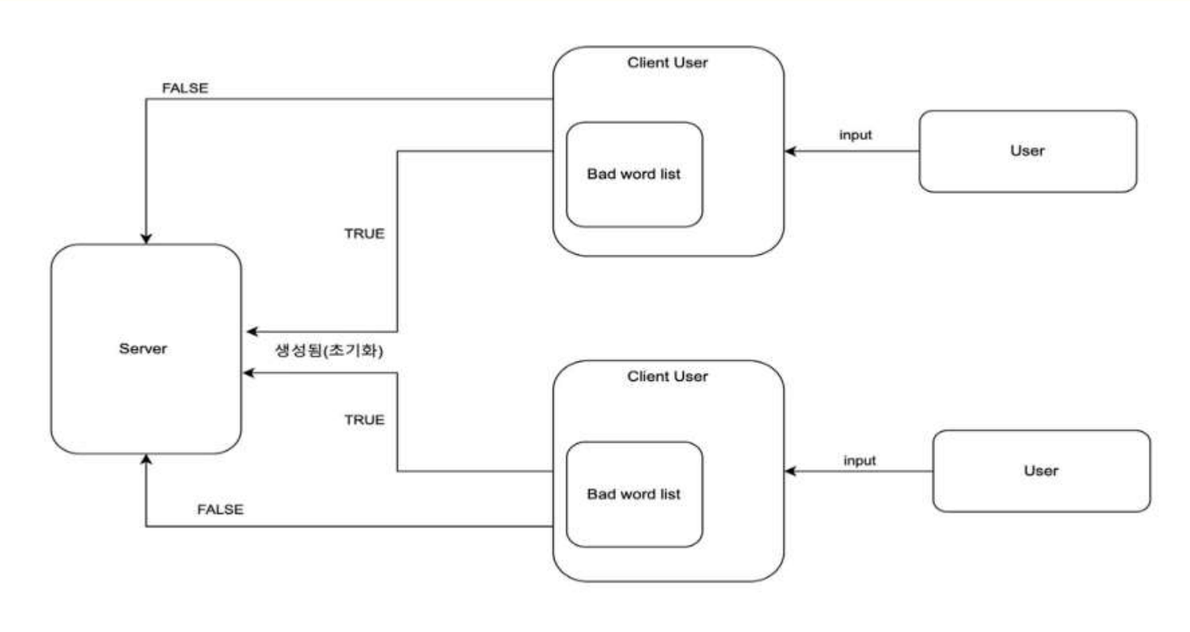
3. 딥러닝 기반 메시지 필터링

- 학습집합을 반자동적으로 확보, 딥러닝 모델이 지속적으로 학습 가능한 메시지 필터링 구조를 제안하여 해당 메시지 필터를 통해 필터링

- 장점 : 앞선 방식의 단점을 보완 가능



사용자 맞춤형 금지어 필터링 모델



사용한 필터링 모델 : 사전 기반 키워드 필터링

필터링 과정 :

1. 개인화된 금지어 목록 관리

2. 메시지 입력 및 수신

3. 단어 탐지 및 대체

4. 메시지 출력 및 전송

실험 절차

채팅 서버 작동 및 클라이언트 서버 연결

채팅 서버를 실행하고 클라이언트를 서버에 접촉하여 실시간 채팅 환경 구축

채팅 테스트

실시간으로 채팅이 잘 보내지는지 확인

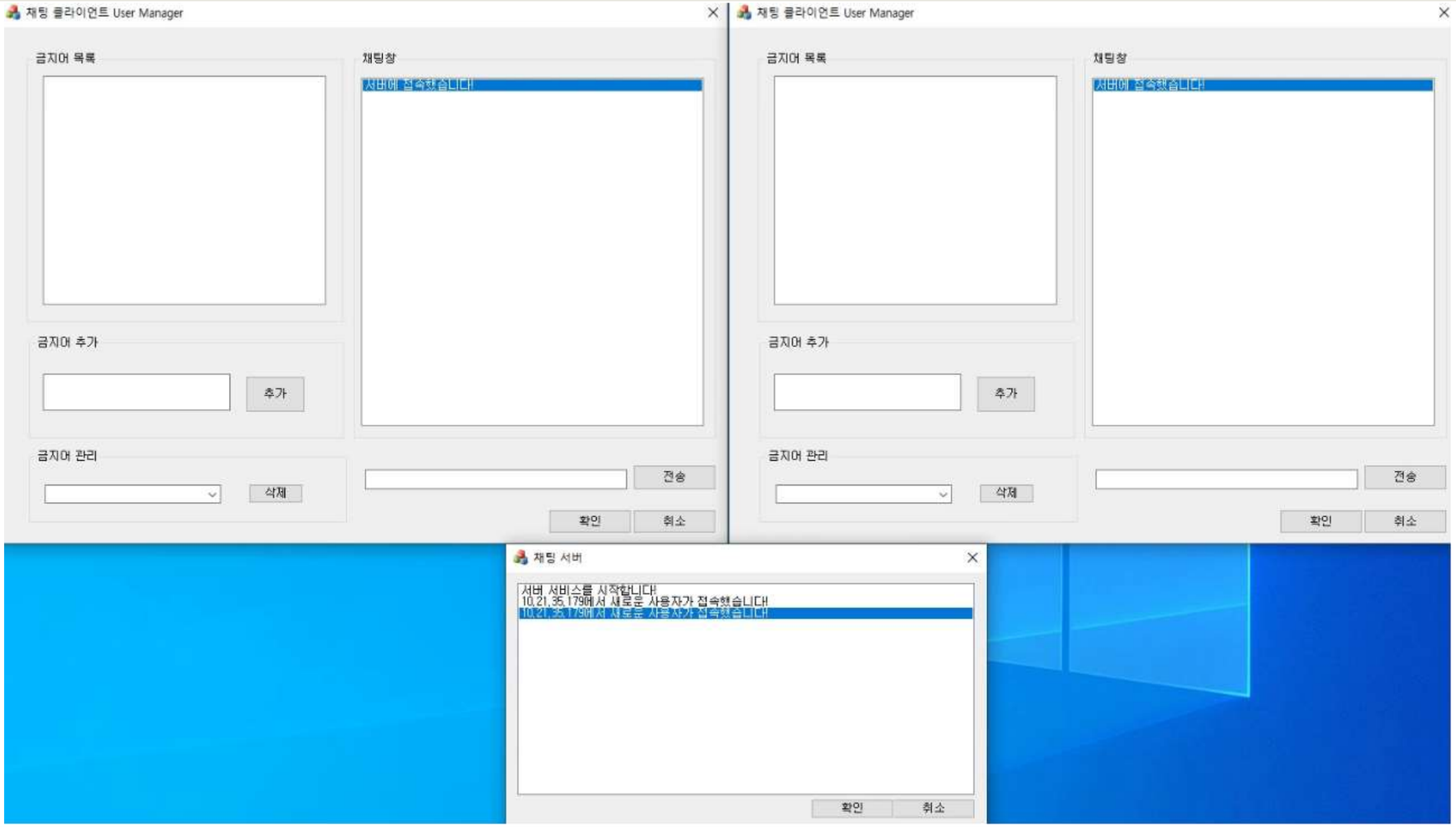
금지어 데이터셋 추가

필터링을 하고자하는 언어를 금지어 목록에 추가

필터링 여부 확인

금지어 목록에 추가된 문자가 제대로 필터링되는지 확인

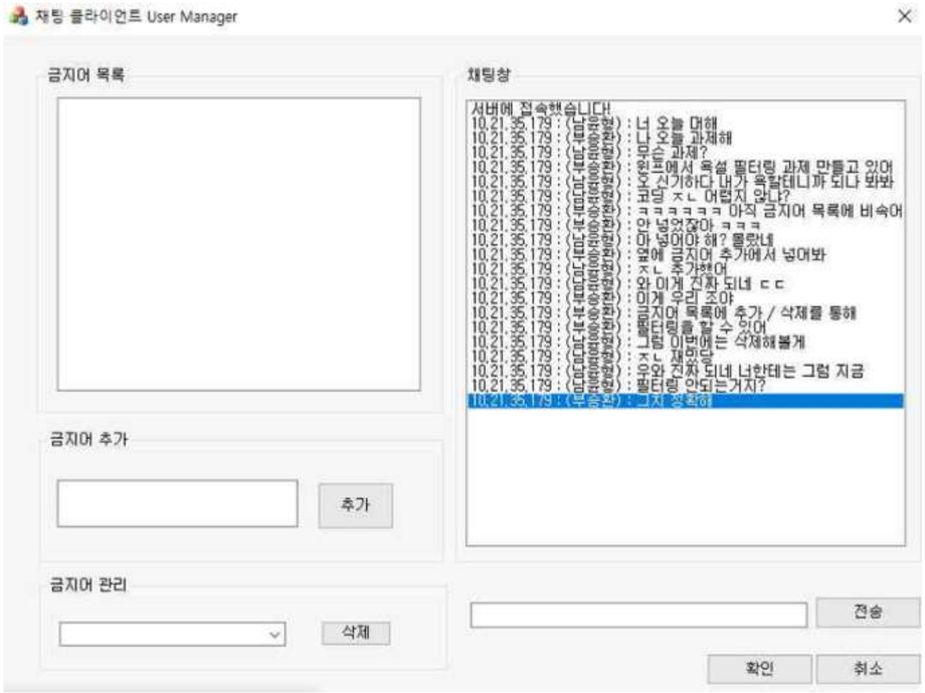
실험 및 실험 결과


채팅 클라이언트 User Manager
금지어 목록
채팅창
서버에 접속했습니다!
금지어 추가
추가
금지어 관리
삭제
전송
확인
취소
채팅 클라이언트 User Manager
금지어 목록
채팅창
서버에 접속했습니다!
금지어 추가
추가
금지어 관리
삭제
전송
확인
취소
채팅 서버
서버 서비스를 시작합니다!
10.21.35.179에서 새로운 사용자가 접속했습니다!
10.21.35.179에서 새로운 사용자가 접속했습니다!
확인
취소

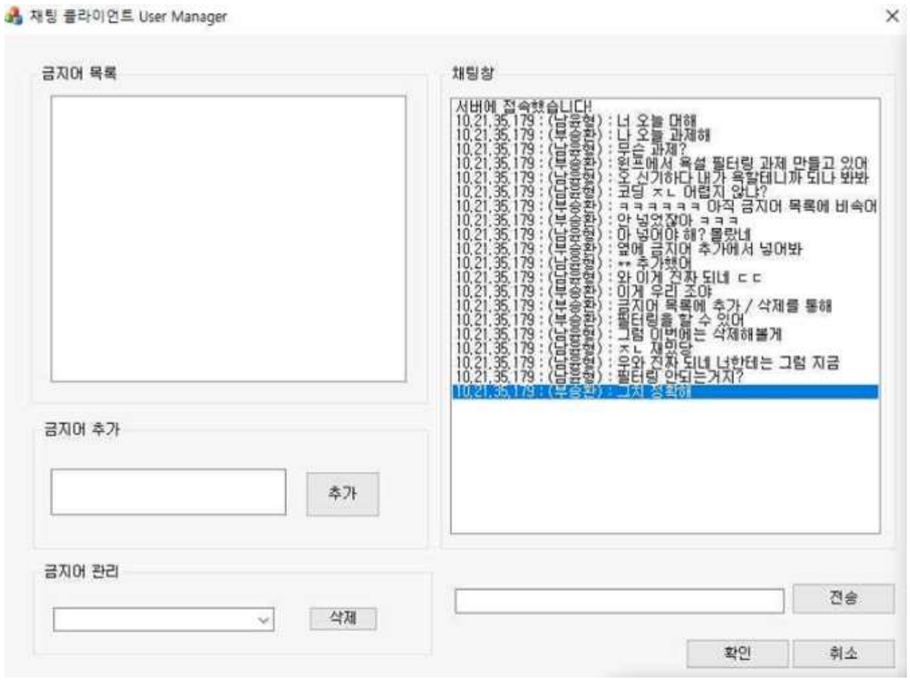
실험 결과 분석

금지어 목록에 추가한 경우 제대로 필터링이 된 모습을 확인할 수 있으며 금지어 목록에서 제외할 경우 원래는 필터링되면 단어가 다시 필터링이 되지 않는 모습을 확인할 수 있다.

분석 결과 - 1



분석 결과 - 2



결론

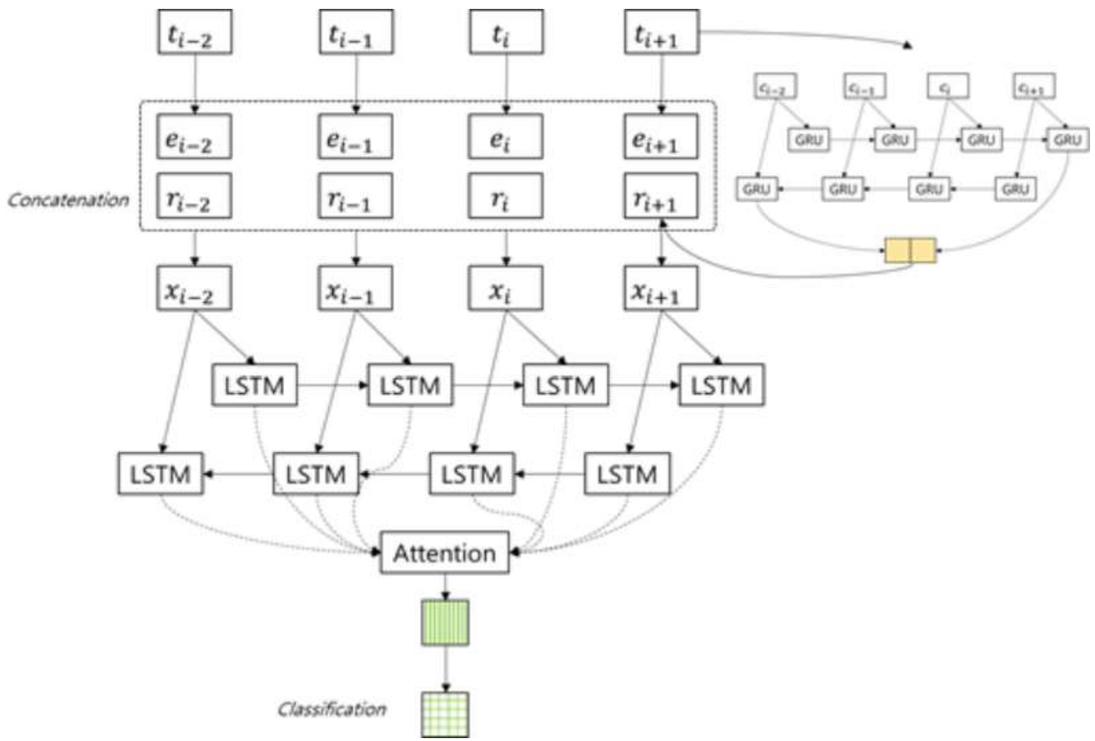
결론 - 1

실시간 채팅 시스템에서 부적절한
언어 사용으로부터 사용자를
보호하고, 건전하고 긍정적인
온라인 커뮤니케이션 환경을
조성하는 데 기여할 수 있을
것으로 기대됨

기대효과

이를 통해 온라인 플랫폼
운영자는 윤리적이고 신뢰할 수
있는 서비스를 제공 하는데
도움을 받을 수 있을것으로 보임

향후 연구 방향



- 딥러닝 기반 필터링 시스템을 활용한 복잡한 문맥 인식 시스템
- 인공신경망을 활용한 동적 필터링 기술을 추가적으로 탐구

Q&A

Thank you!