사용자 맞춤형 금지어 필터링을 적용한 실시간 온라인 채팅 시스템

김동환[1], 남윤형[1], 부승환[1], 장준석[1], 정해윤[1], 트란트룽하우[1]
[1]숭실대학교 소프트웨어학부 학부생
20203087@soongsil.ac.kr, erty0834@soongsil.ac.kr, bsh0109@soongsil.ac.kr,
junseok7502@soongsil.ac.kr, 20213145@soongsil.ac.kr, 20230127@soongsil.ac.kr

Real-time online chat system with customized prohibited word filtering

Dong-hwan Kim[1], Youn-hyeong Nam[1], Seung-hwan Boo[1], Jun-seok Jang[1],
Hae-yoon Jeong[1], Tran Trung Hau[1]
[1]Dept. of Software Science, Soongsil University

요 약

온라인 채팅 시스템은 소셜 네트워킹 애플리케이션, 고객 지원 채팅, 게임 채팅 등 다양한 플랫폼에서 널리 사용되며 사용자 간의 빠르고 쉬운 상호작용을 가능하게 한다. 그러나 이러한 환경에서 공격적인 언어나 부적절한 행동의 사용은 사용자 경험에 부정적인 영향을 미치고 커뮤니케이션 환경의 질을 저하시킬 수 있다. 이에 본 연구는 사용자 맞춤형 금지어 필터링 모델을 제안한다. 이 모델은 사용자가 원하는 금지어를 직접 추가할 수 있는 금지어 관리 기능을 제공하며, 사용자별로 개별 금지어 목록을 유지하여 각 사용자의 환경에 따라 다른 필터링 규칙을 적용한다. 사용자가 추가한 금지어는 상대방이 입력하거나 사용자가 입력하더라도 필터링되어 *로 변환되며, 이를 통해 부적절한 언어 사용을 효과적으로 차단한다. 서버는 모든 채팅 메시지를 필터링 없이 수신하며, 메시지를 금지어 목록과 비교하여 처리한 뒤 클라이언트에 전달하는 방식으로 동작한다. 이 시스템은 금지어와 그 변형을 정확히 탐지하고 빠르게 처리하여 사용자 경험에 영향을 주지 않는 것을 목표로 하며, 개인화된 금지어 필터링을 통해 각 사용자의 필요에 부합하는 안전한 커뮤니케이션 환경을 조성할 수 있다.

1. 개발 배경

최근 사회에서 욕설과 비속어 사용이 심각한 문제로 부각되고 있다. 「2020년 국민의 언어 의식 조사」(국립국어원, 2020)에 따르면 응답자의 46.9%와 48.1%는 우리 국민이 각각 욕설·비속어를 사용한다는 생각을 한다고 답했다.[1] 이러한 문제는 온라인 채팅의 활성화로 인해 더욱 심화되고 있다. 방송통신위원회와 지능정보사회진흥원의 '2021년 사이버폭력 실태조사'에 따르면 사이버폭력을 경험했다는 응답은 청소년 29.2%, 성인 15.7%로 나타났다. 특히 사이버폭력 유형 중에서는 청소년과 성인 모두 '사이버 언어폭력' 경험률이 높은 것으로 나타났다. 언어폭력에 대해 청소년 12.0%는 가해, 16.4%는 피해 경험을 밝혔고,. 성인은 가해 경험률이 5.9%, 피해 경험률이 8.7%로 확인됐다.[2]

온라인상의 욕설과 비속어는 사회적 분위기를 해치고 타인에게 심리적 피해를 줄 수 있으며, 온라인 문제로 끝나지 않고 오프라인까지 확산될 수 있다. 본 연구의 목적은 실시간 채팅 시스템에서 무의식적으로 사용하는 욕설과 비속어를 탐지하고, 해당 단어들을 필터링하는 모델을 개발하는 것이다. 이를 통해 온라인 커뮤니케이션 환경을 개선하고 사용자들의 언어 사용에 대한 인식을 높이고자 한다.

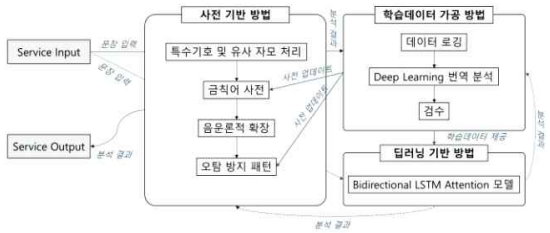
2. 관련 자료

온라인 커뮤니케이션에서 부적절한 언어 사용을 감지하고 필터링하는 방법에 대한 연구는 오랫동안 진행되어 왔다. 대표적인 방법으로는 정규 표현식(Regular Expression)을 활용한 필터링, 사전 기반 키워드 필터링을 이용한 문맥 기반 감지 등이 있다.

- 정규 표현식 기반 필터링 : 이는 미리 정의된 패턴을 기반으로 특정 단어를 탐지하는 방식이다. 구현이 간단하고 효율적이지만, 변형된 단어(예: "놈"을 "ㄴㅗㅁ"으로 변형)에는 취약하다.[3]
- 사전 기반 키워드 필터링 : 이 방법은 금지어

목록을 설정하고 메시지에서 해당 단어가 포함되는지를 검사한다. 관리가 용이하지만, 새로운 변형어나 신조어에 대한 대응이 어렵다.

• 딥러닝 기반 메시지 필터링 : 딥러닝 기반 메시지 필터링은 학습집합을 반자동적으로 확보하여 딥러닝 모델을 지속적으로 학습 가능한 메시지 필터링 구조를 제안하여 사전 기반 방법의 필터링의 단점인 변형어나 신조어에 대한 대응이 가능한 필터링 구조이다.



(그림 1) 사전 기반 방법과 딥러닝 기반 메시지 필터링 구조[4]

3. 사용자 맞춤형 금지어 필터링 모델

3.1 필터링 모델 설계

본 연구에서는 실시간으로 부적절한 단어를 탐지하고 이를 *로 대체하는 사전 기반 필터링 방식을 사용하여 간단하고 효율적인 필터링 시스템을 사용한다. 사용된 시스템은 사용자의 금지어 목록을 기반으로 부적절한 단어를 탐지하며, 사용자의 요구에 따라 목록을 확장할 수 있는 유연성을 제공한다.

더불어, 실시간 처리를 최우선 목표로 시스템을 설계하였다. 사용자가 메시지를 입력하는 즉시 필터링 과정을 거치며, 이를 통해 사용자 경험에 부정적인 영향을 미치지 않고 빠르게 메시지를 처리한다.

3.2 필터링 과정

3.2.1 개인화된 금지어 목록 관리

사용자는 개인화된 금지어 목록을 관리할 수 있으며, 필요에 따라 원하는 금지어를 추가할 수 있다. 사용자별로 금지어 목록이 독립적으로 관리되며, 각 사용자의 환경과 요구에 따라 상이한 금지어 목록이 적용된다.

3.2.2 메시지 입력 및 수신

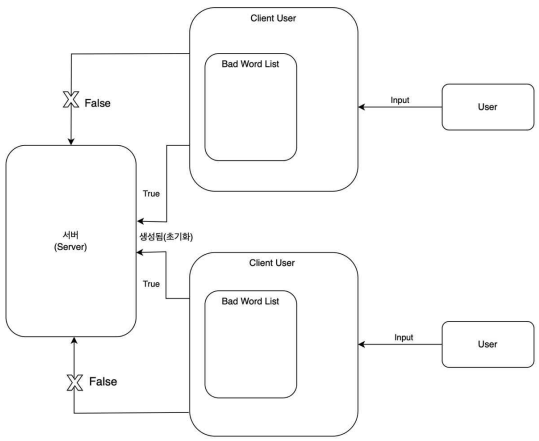
사용자가 채팅창에 메시지를 입력하면, 시스템은 이를 실시간으로 수신한다. 메시지 수신 과정에서 서버는 모든 메시지를 필터링 없이 처리하며, 메시지는 클라이언트 측에서 추가적인 필터링을 위해 전달된다.

3.2.3 단어 탐지 및 대체

시스템은 수신된 메시지 내의 단어를 검사하여 금지어 목록에 포함된 단어를 탐지한다. 탐지된 금지어는 사용자가 입력했든 상대방이 입력했든 모두 필터링되어 별표(*로 변환)로 대체된다. 이 과정은 사용자의 금지어 목록에 따라 개별적으로 수행된다.

3,2,4 메시지 출력 및 전송

필터링이 완료된 메시지는 상대방에게 전송된다. 이를 통해 금지어가 포함된 메시지가 전달되지 않도록 보장하며, 사용자 간의 불쾌한 경험을 최소화한다.



(그림 2) 사용자 맞춤형 금지어 필터링 모델 설계도

4. 실험

4.1 실험 설계

본 연구에서는 제안된 온라인 채팅 필터링을 바탕으로 시스템을 평가하기 위해 다음과 같은 실험을 진행하였다.

4.1.1 준비한 데이터 셋

• 채팅 데이터 셋: 소셜 미디어 채팅 로그와 커뮤니티 게시판에서 수집된 실제 메시지를 사용하였다.

• 금지어 데이터 셋 : 사람들이 사용하는 욕설에 대한 데이터셋을 준비하여 변수에 넣어 사용하였다.

4.1.2 실험 진행 과정

① 채팅 서버를 작동한다.

② 2개의 클라이언트를 작동하여 채팅 서버에 연결한다.

③ 필터링 확인을 위해 금지어 목록 추가 전 채팅 데이터 셋으로 채팅을 진행한다.

④ 각 클라이언트의 금지어 목록에 금지어 데이터 셋을 추가한다.(각 금지어 데이터 셋은 클라이언트에 따라 다른 데이터를 추가한다.)

⑤ 3번에서 사용한 채팅 데이터 셋으로 채팅을 진행한다.

⑥ 각 언어가 원활하게 필터링되는지 확인한다.

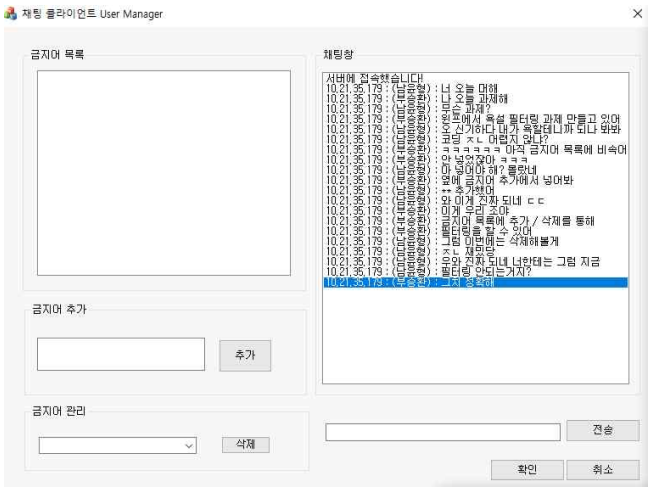
⑦ 동일한 단어가 각 클라이언트의 금지어 목록에 따라 서로 다른 필터링 현상을 띠는지 확인한다.

4.2 결과 분석

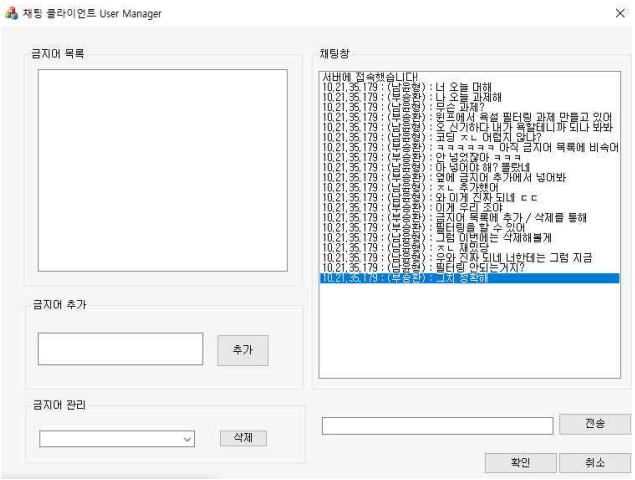
실험 결과, 각 클라이언트는 독립적으로 관리되는 금지어 목록을 기반으로 채팅 데이터를 필터링하는 데 성공하였다. 각 클라이언트의 금지어 목록에 포함된 단어는 사전 기반 필터링 방식을 통해 정확히 탐지되었으며, 금지어가 포함된 메시지는 모두 지정된 방식(*로 대체)으로 처리되었다.

특히, 동일한 단어가 클라이언트별로 상이한 금지어 목록에 따라 다르게 필터링되는 현상을 확인할 수 있었다. 이는 사용자 맞춤형 금지어 관리 기능이 효과적으로 작동하며, 다중 사용자 환경에서 개인화된 필터링이 가능함을 보여준다.

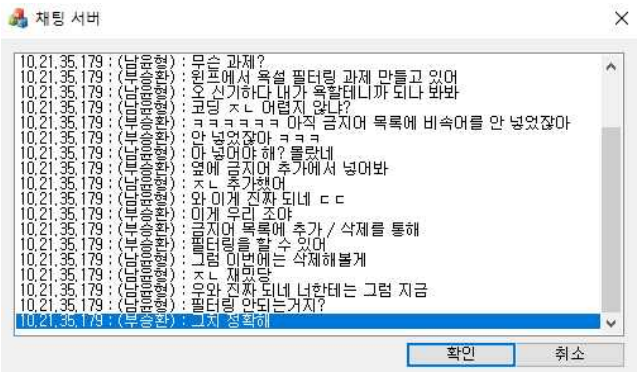
이러한 결과는 제안된 시스템이 실시간 필터링과 개인화된 금지어 관리라는 두 가지 핵심 목표를 성공적으로 달성하였음을 입증한다.



(그림 3) 사용자1



(그림 4) 사용자2



(그림 5) 채팅 서버

5. 결론

본 연구에서는 실시간 온라인 채팅 시스템에서 부적절한 단어를 효과적으로 감지하고 필터링하는 모델을 제작하였다. 제작한 모델은 사전 기반 필터링 기반 방식을 활용하여 금지어를 감지할 수 있도록 설계되었고, 실시간 처리 속도를 최적화함으로써 필터링 과정이 사용자 경험에 부정적인 영향을 미치지 않도록 구현하였다.

본 연구의 결과는 실시간 채팅 시스템에서 부적절한 언어 사용으로부터 사용자를 보호하고, 건전하고 긍정적인 온라인 커뮤니케이션 환경을 조성하는 데 기여할 수 있을 것으로 기대된다. 이를 통해, 온라인 플랫폼 운영자는 사용자 경험을 향상시키고, 윤리적이고 신뢰할 수 있는 서비스를 제공하는 데 도움을 받을 것으로 기대된다.

향후에는 딥러닝 기반 필터링 시스템을 활용한 더 복잡한 문맥 처리, 그리고 인공신경망을 활용한 동적 필터링 기술을 추가적으로 탐구하고자 한다.

참고문헌

[1] 소강춘, “2020년 국민의 언어 의식 조사 결과보고서”, 국립국어원, 2020-01-35, December,2020.

[2] 한상혁, “2021년 사이버폭력 실태조사 보고서”, 방송통신위원회, 제164003호, December,2021.

[3] 장현석, 김형우, 오병석, and 안종석, "정규표현식 Chatbot과 RPA 기반 자동 예약 시스템," 2019, pp. 1324-1326.

[4] 김선우, 권나현, 김민정, 장진후, and 김정길, "학습집합 확장을 위한 딥러닝 기반 메시지 필터링 구조 연구," 2021, pp. 310-312.