컴퓨터교육학회 논문지 2024년 제27권 제3호
https://doi.org/10.32431/kace.2024.27.3.011

메타버스 시스템 내에서 반시민적 언어 탐지 및 시각화 기반 교정 메커니즘 구현*

Implementation of the Correction Mechanism Based on Detection and Visualization of Uncivil Language Expressions in the Metaverse

박윤정† · 이세영†† · 금희조†††
Younjung Park† · Seyoung Lee†† · heejo Keum†††

요 약

메타버스의 음성 채팅 환경에서 나타나는 반시민적 언어 표현은 사이버불링의 주요 원인 중 하나로 간주되어, 사용자의 안전과 건전한 커뮤니케이션 환경의 유지를 위해 적극적인 대응 전략이 필요하다. 본 연구는 이러한 상황의 중요성을 인지하고, 메타버스 내에서 반시민적 언어 표현을 실시간으로 감지하고, 사용자들이 반시민적인 행동을 자제하도록 독려하는 메커니즘 구현에 주안점을 둔다. 특히, 본 연구에서는 어텐션 알고리즘에 기반한 반시민적 언어 탐지 모듈과 사용자에게 직접적인 시각적 피드백을 제공하는 모듈을 메타버스 환경에 통합하였다. 이와 같은 기술적 개입을 통해 사용자는 직접적인 피드백을 통해 음성 채팅의 언어 패턴에 대한 성찰의 기회를 갖게 되며, 이를 통해 개인의 커뮤니케이션 습관 개선에 이바지하게 될 것으로 예상된다. 장기적으로, 이러한 긍정적 변화는 메타버스 내의 건전한 커뮤니티 형성을 촉진하고, 사이버불링과 같은 부정적 현상의 감소에 기여, 사회 전반의 디지털 커뮤니케이션 문화의 질을 향상시킬 것으로 기대된다.

주제어: 메타버스, 반시민적 언어 표현, 사이버불링, 어텐션 알고리즘, 시각적 피드백

ABSTRACT

Within the metaverse's voice chat, uncivil language expressions are considered one of the primary causes of cyberbullying. Ensuring user safety and maintaining a wholesome communication atmosphere necessitates proactive intervention strategies. Recognizing the significance of this issue, this study focuses on the real-time detection of uncivil language expressions within the metaverse and the implementation of mechanisms that encourage users to refrain from uncivil behaviors. Specifically, this research integrates an attention algorithm-based module for detecting uncivil language and a module providing users with direct visual feedback within the metaverse environment. Through this technical intervention, users are afforded an opportunity for introspection regarding their voice chat language patterns, leading to anticipated improvements in individual communication habits. In the long run, such positive shifts are expected to promote the formation of healthy communities within the metaverse, contribute to the reduction of negative phenomena like cyberbullying, and elevate the overall quality of digital communication culture in society.

Keywords: Metaverse, Uncivil language expressions, Cyberbullying, Attention algorithm, Visual feedback

†정 회 원: 성균관대학교 글로벌융복합콘텐츠연구소 선임연구원
††정 회 원: 성균관대학교 미디어커뮤니케이션학과 부교수
†††정 회 원: 성균관대학교 미디어커뮤니케이션학과 정교수(교신저자)

논문투고: 2023년 10월 17일, 심사완료: 2024년 05월 14일, 게재확정: 2024년 05월 15일

* 본 논문은 2022년 한국정보통신학회학술지에서 "어텐션 임베딩과 다채널 CNN기반 반시민성 검출 알고리즘" 의 제목으로 발표된 논문을 확장한 것임.
본 논문은 2021년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-No.2021S1A5C2A02088387)

1. 서론

디지털 시대의 도래와 함께, 우리의 커뮤니케이션 방식과 환경도 크게 변화하고 있다. 특히 메타버스는 이 변화의 선두에 서 있다. 메타버스는 사용자들이 아바타를 통해 가상세계에서 활동하고, 다른 사용자들과 상호작용을 하면서 의사소통의 새로운 차원을 탐험하는 공간이다. 이러한 메타버스의 성장은 단순히 디지털 엔터테인먼트의 영역을 넘어, 사회와 교육, 비즈니스 등 여러 분야에서의 중요한 커뮤니케이션 플랫폼으로 자리 잡아가고 있다. 그러나 이러한 발전과 함께, 메타버스 내에서의 사이버불링 문제가 부각되며, 특히 음성 채팅을 중심으로 한 반시민적 언어 표현이 심각한 문제로 떠오르고 있다. 이에 따라, 메타버스 환경에서의 건전한 커뮤니케이션 문화 조성이 절실한 사회적 요구로 제기되고 있다.

온라인에서의 반시민적 행동은 개인이나 집단에 민감하게 반응하도록 만드는 댓글과 표현에 해당한다. 반시민적 행동이란 혐오 표현, 악성 댓글, 가짜 정보, 사이버불링 등의 다양한 문제적 표현과 행동을 포괄하는 개념으로, ‘무례 차원’과 ‘혐오 차원’으로 구분될 수 있다[1]. ‘무례 차원’은 특정 개인에 대해 사용되는 욕설이나 저속한 표현을, ‘혐오 차원’은 특정 집단의 고유한 특징(예: 성별, 국적, 인종)을 근거로 한 공격적이며 악의적인 표현을 중심으로 하게 된다[2-4]. 이러한 반시민적 댓글들은 상대방의 불쾌감을 유발하거나 온라인 뿐만 아니라 오프라인에서도 특정 집단과 그 구성원에 대한 차별 및 혐오를 야기하며, 결국 민주주의의 다원성을 위협할 수 있다[5, 6].

기존 온라인 플랫폼이나 기존 연구에서는 온라인 커뮤니티에서 발생하는 텍스트로부터 반시민적인 표현을 검출해내기 위해 여러 노력을 기울여 왔다. 특히 텍스트 분석을 위한 딥러닝 모델들이 발전함에 따라, 온라인 플랫폼에서의 반시민적 발언이나 혐오 표현을 자동으로 탐지하고 필터링하는 연구가 활발히 진행되고 있다. 소셜 미디어 사이트들은 이러한 연구의 성과를 활용하여 사용자들의 게시글을 실시간으로 모니터링하고 있으며, 위반될 가능성이 있는 콘텐츠는 자동으로 검열되거나 사용자에게 경고가 가도록 설정되어 있다. 예를 들어, 트위터의 혐오 발언을 딥러닝을 이용하여 분류하는 시스템을 규현한 연구에서는 트위터의 각 트윗을 인종 차별, 성 차별, 혹은 인종과 성 모두 차별, 그리고 혐오 발언이 아닌 카테고리로 분류하였다[7]. 이 연구에서는 word2vec 기반의 임베딩을 활용한 모델이 뛰어난 성능을 보였으며, 이런 연구 결과를 통해 플랫폼들은 혐오 발언을 효과적으로 탐지하고 관리할 수 있게 되었다. 또한 CNN(Convolutional Neural Network) 기반의 모델을 활용하여 혐오 표현을 자동으로 감지하는 연구에서는 텍스트 데이터의 시각적 패턴을 인식하여 트윗 내의 혐오적인 언어 사용을 감지하기도 하였다[8]. 이러한 연구와 기술의 발전은 메타버스와 같은 가상 공간에서의 커뮤니케이션 문화를 건전하게 유지하는 데에 큰 도움을 주고 있다. 그리고 이는 단순한 기술적 발전을 넘어 사회적 측면에서도 중요한 의미를 지닌다. 메타버스와 같은 디지털 플랫폼이 사회의 중심적 역할을 점점 더 차지하게 됨에 따라, 그 안에서의 커뮤니케이션 문화는 전체 사회의 문화와 윤리에 큰 영향을 미칠 것이다.

커뮤니케이션 문화의 변화와 그에 따른 영향은 온라인 공간의 한계를 넘어선다. 이 문제의 심각성을 충분히 인지하고, 그 해결을 위한 방안을 모색하는 것은 단순한 온라인 문제를 넘어 사회 전반의 건전한 의사소통 문화를 위한 필수적인 노력이라고 할 수 있다[2]. 이러한 노력의 일환으로서, 우리는 반시민적인 언어를 검출하였던 기존의 연구결과를 바탕으로 이를 실제적으로 활용할 수 있는 디지털 커뮤니케이션 공간을 구현하고자 하였다. 즉, 반시민적인 언어의 검출 후 사용자에게 반시민성에 대해 직접적인 피드백을 줄 수 있는 디지털 공간의 구현을 목적으로 하였으며 이를 위하여 반시민적 표현에 대해 시각적 피드백이 진행되는 메타버스 플랫폼의 구현을 시도하였다.

2. 연구 배경

2.1 연구 목적 및 필요성

메타버스는 디지털화된 현대 사회에서 빠르게 성장하고 있는 가상 공간으로, 사용자들의 일상, 엔터테인먼트, 비즈니스 활동 등 다양한 활동이 진행되고 있다. 최근 몇 년 동안, 특히 2021-2022년도에는 메타버스에 대한 관심이 급증하였는데, 이는 부분적으로 페이스북이 “Meta“로 자신의 브랜드 명칭을 변경한 것에 기인한다[9]. 현대적인 의미에서의 메타버스는 확장현실(XR) 영역에 속하며, 이는 증강현실, 혼합현실 및 가상현실을 포괄한다[10].

이러한 메타버스의 성장은 2020년대 초기의 기술적

진보와 코로나19 팬데믹에 따른 사회적 변화로 인해 가속화되었다. 이 변화는 사용자들에게 일상의 많은 활동을 가상 세계에서도 경험할 수 있게 해 주었으며, 이는 교육, 건강관리, 게임 및 엔터테인먼트, 예술, 사회 및 시민 생활 등 사회의 모든 부분에서의 혜택을 의미한다[9]. 이런 성장과 발전은 메타버스를 단순한 가상 환경을 넘어서 사회적, 문화적, 경제적 중요성을 지닌 공간으로 만들었다.

메타버스의 성장과 발전은 사용자들에게 새로운 가상의 경험과 활동의 공간을 제공하는 반면, 동시에 다양한 문제점을 야기하고 있다. 메타버스에서 사용되는 장비인 VR 및 AR 헤드셋은 해킹과 같은 사이버 공격을 허용할 수 있으며, 특히 일부 메타버스 기술은 새로운 형태의 악의적인 사이버 활동을 유발할 가능성이 있다[11].

게임 뿐만 아니라 교육 분야에서도 메타버스가 활용되면서 사이버 불링과 같은 문제가 더욱 도드라지게 되었다. 특히, Roblox와 같은 교육용 플랫폼에서의 학습 활동과 관련하여 연구된 결과에 따르면, 사이버 불링은 교육에 메타버스를 적용할 때의 주요 도전 과제 중 하나로 지적되고 있다[12]. 더욱이, 사이버 불링은 초등학교 학생들 사이에서도 나타나는 현상이다. 기존 연구에서, 초등학교 학생들 중 상당수가 사이버 불링 행위를 목격하였으며, 이들 대부분은 비디오 게임을 하거나 동영상 공유 웹사이트를 이용할 때 이러한 행위를 목격하였다는 것을 보여주었다[13]. 사이버 불링의 피해자나 가해자로서의 경험이 성별에 따라 차이를 보이지는 않았지만, 학년과 나이에 따라 불링 피해와 가해 행위의 비율에다는 차이가 있었다. 이러한 연구 결과는 초등학교 초기부터 사이버 불링이 나타나며, 학년이 올라갈수록 이러한 행위가 증가하는 경향이 있음을 시사한다.

이처럼, 메타버스의 활용과 성장은 다양한 잠재적 문제점과 도전 과제를 야기하며, 이에 대한 적절한 대응 방안과 해결책을 모색하는 것이 중요하다. 우리는 기존의 연구에서 반시민적인 언어를 탐지하는 언어 모델을 구현하였다. 본 연구에서는 기존의 연구 결과를 메타버스 환경에 적용하여 메타버스 내의 음성 채팅 중에 발생하는 반시민적 언어 사용을 실시간으로 감시하고, 이를 감지할 경우 사용자에게 즉각적인 시각적 피드백을 제공하는 시스템을 구현하는 것을 목적으로 하였다. 메타버스 환경에서 음성 채팅은 주요한 커뮤니케이션 수단 중 하나이며, 이를 통해 많은 사용자들이 정보를 공유하고 서로 소통한다. 하지만 이러한 음성 채팅을 통해 다양한 혐오 표현이나 부적절한 발언이 빈번히 발생하고 있다. 이는 메타버스 환경 내에서의 안전한 커뮤니케이션을 방해하며, 사용자들의 메타버스 경험을 부정적으로 만들 수 있다.

본 연구에서 구현된 시스템은 메타버스 내에서 음성 채팅이 진행되는 동안, 혐오 표현이나 무례 표현 같은 반시민적 언어를 실시간으로 감지하게 된다. 감지된 경우, 해당 사용자에게는 즉시 시각적 피드백이 제공된다. 이 피드백은 가해자와 피해자 모두에게 전달되어 사용자들에게 반시민적인 발언이 부적절하다는 것을 알리고, 그러한 언어 사용의 자제를 권고하는 역할을 한다.

이 시스템의 적용으로, 메타버스 내에서의 음성 채팅 환경이 더욱 건전하고 존중 받는 공간으로 변화할 것이다. 사용자들은 부적절한 발언에 대한 즉각적인 피드백을 통해 자신의 언어 사용을 스스로 조절하게 되며, 이를 통해 메타버스 커뮤니티 전체의 소통 품질이 향상될 것으로 기대된다.

2.2 이론적 배경

2.2.1 반시민적 언어의 정의

시민성은 민주주의 사회에서 가장 중요한 핵심 가치 중 하나로 간주된다. 시민성은 공동체의 바람직한 구성원으로서 지녀야 할 품성, 태도, 권리 등을 포괄적으로 포함한다[14]. 그들은 이러한 시민성이 공동체 내에서의 유대감과 소속감을 높이며, 사회적 통합과 조화를 도모한다고 설명한다.

이러한 전통적인 시민성 개념은 현대 디지털 시대에 맞게 발전해 ‘디지털 시민성’이라는 새로운 개념으로 확장되었다. 디지털 시민성은 ‘다양한 디지털 환경에서의 바람직한 행동과 자질’로 정의될 수 있으며, 이는 디지털 기술의 활용 능력뿐만 아니라 온라인에서의 윤리적 행동과 디지털 정보의 적절한 사용 등을 포괄한다[15].

온라인 반시민성은 이러한 시민성 개념의 반대되는 현상으로, 온라인 미디어와 플랫폼 상에서 나타나는 반규범적 언행을 포함한다. 구체적으로, 특정 개인이나 집단에 해를 끼치는 문제적인 표현, 태도, 행동 등을 포괄하는 개념이다[1]. 여기에는 개인 차원의 문제적 표현인 ‘무례 차원’과 공공(집단) 차원의 문제적 표현인 ‘혐오 차원’이 모두 포함된다[16, 17].

‘무례 차원’에서의 반시민성은 특정 개인의 가치

나 인격을 저하시키는 표현들로 구성되어 있다. 예를 들면 '돼지', '멍청이'와 같은 인신공격성 표현(name-calling), '죽어버려', '꺼져'와 같은 공격적 어조, '어디서 사기를 쳐'와 같이 다른 사람의 정직성을 무조건적으로 비난하는 표현 등이 해당된다[3, 9, 17]. '혐오 차원'에서의 반시민성은 집단 간의 커뮤니케이션에서 정중함의 규범을 위반하는 것을 포함한다. 이는 특정 집단의 성별, 연령, 정치적 성향, 성정체성, 종교, 장애 여부 등의 고유한 특성을 기반으로 한 적대적 또는 편견적 태도, 증오 및 차별 표현을 포함한다[5, 18].

온라인 커뮤니티에서의 반시민적 언어 사용은 단순히 개인의 가치를 저하시키는 것 이상으로 커뮤니티 전체의 유대감과 소속감을 약화시킬 수 있다[17]. 즉, 디지털 시민성이 온라인 공동체의 조화와 통합을 위해 필수적인 것처럼, 그 반대인 반시민적 언어는 그러한 조화와 통합을 해칠 위험성을 내포하고 있는 것이다. 그렇기 때문에 이러한 반시민적 언어를 검출하고 반시민적인 표현을 억제하기 위해서는 시스템적인 방법이 요구된다.

2.2.2 반시민적 언어 표현의 검출을 위한 머신러닝 알고리즘 연구

기존의 연구들에서는 주로 욕설이나 강한 혐오 표현의 검출에 중점을 둔 반면, 최근의 연구는 더욱 미묘한 반시민적 언어 표현의 검출에 초점을 맞추고 있다. 이러한 연구의 필요성은 플랫폼들이 고도화된 혐오어 필터링 기능을 도입함에 따라 사용자들이 이를 우회하기 위한 다양한 언어 표현을 사용하게 되면서 나타났다. 예를 들어, 네이버는 2020년 이후 인공지능형 클린봇을 운영하여 높은 자연어 처리 능력과 혐오 표현 검출율을 보였으며, 이로 인해 현재 댓글에서는 노골적인 욕설이나 비난은 크게 감소하였다[19].

그러나, 디지털 공간의 댓글 플랫폼 사용자들은 노골적인 표현보다는 인신공격적인 발언이나 상대의 의견을 거짓으로 치부하는 언어, 공격적 어조 등의 더 미묘한 방식을 택하고 있다[3]. 온라인 공간에서 이루어지는 댓글의 무례함은 대화의 품질과 지속성에 부정적 영향을 미칠 수 있기 때문에, CNN을 사용하여 신문 댓글 섹션에서의 인신공격성 표현(name-calling) 및 욕설이나 음란한 표현(vulgarity)을 탐지하는 연구가 이루어지기도 하였다[20]. 또한 댓글 섹션 뿐 아니라 마이크로블로그, 온라인 뉴스 댓글 등의 여러 도메인에서의 반시민적 표현들 즉 인신공격성 표현(name-calling), 욕설이나 음란한 표현(vulgarity), 위협적 표현 등을 검출할 수 있는 알고리즘이 제시되기도 하였다[21]. 여기에서는 BERT 기반 모델에 개별 이는진 분류기를 도입한 알고리즘의 성능이 다중 라벨을 적용하는 것보다 더 나은 성능을 보인다는 것을 보여주었다.

국내에서는 뉴스 댓글 데이터 중 반시민적인 표현을 검출하기 위해 어텐션 기반 특징 추출 알고리즘과 멀티채널 CNN을 활용한 연구가 수행된 바 있다[22]. 이 연구에서는 대통령 선거와 관련된 기사의 댓글 데이터를 취합하여 인신공격성 표현(name-calling), 공격적 어조, 욕설(vulgarity)과 같은 '무례 표현'과 지역, 성별, 연령 차별 등의 '혐오 표현'을 포함하여 총 13가지 항목으로 댓글의 반시민성에 대한 라벨링을 실시하였으며, 어텐션 알고리즘을 이중으로 적용하여 임베딩 벡터를 추출한 뒤 2-d CNN으로 분류하는 방식을 제안하였다. 그림 1은 13가지의 라벨링을 나타내고 있다. 그림 1에서와 같이, 라벨링은 무례 및 공격적 표현의 4가지와 혐오 및 차별적 표현 9가지로 이루어져 있으며 요약 데이터(Summarized Data)에는 해당 라벨링이 어떤 단어에 근거하고 있는지가 기입되어 있다.

Figure 1. Data labeling process to implement 'incivility detection algorithm'

이 알고리즘은 원문 데이터와 요약 데이터라는 두 개의 데이터가 각각 어텐션 알고리즘을 통과하여 임베딩 벡터화 된 후 다시 2D CNN의 입력으로 사용된다는 특징이 있다. 두 개의 입력을 받은 2D CNN은 출력으로서 13개의 라벨 즉, 무례, 폄훼, 공격적 어조, 불필요한 욕설, 지역차별, 장애차별, 성차별, 인종차별, 종교차별, 소수자차별, 직업차별, 연령차별, 정치차별 등의 총 13가지 라벨 중 하나의 라벨로 각각의

댓글을 분류하게 된다. 이는 그림 2에서 자세히 나타나 있다.

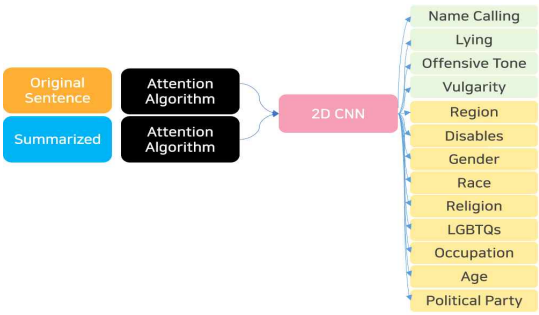

Figure 2. Rude and hate speech detection module based on Attention algorithm

2.2.3 시각적 피드백과 행동 조절 효과

반시민적 언어 표현에 관한 시각적 피드백에 대한 직접적인 연구는 현재까지의 문헌에서 확인되지 않았다. 그러나 시각적 피드백을 통한 셀프 모니터링의 효과에 대해서는 일부 연구된 바가 있다. 예를 들어, [23]은 가상현실을 이용한 코어 운동 자세 트레이닝에서 시각적 및 촉각 피드백이 사용자의 운동 자세 정확도를 향상시키는 데 어떤 영향을 미치는지 연구하였다. 이 연구에서 제공된 시각적 피드백은 사용자가 자신의 운동 자세를 실시간으로 모니터링하고 조정할 수 있게 도와줌으로써, 전반적인 운동 효과를 증진시켰다.

이러한 연구들은 시각적 피드백이 사용자의 자기 모니터링 능력을 강화시키고, 결과적으로 행동 개선을 도모할 수 있음을 보여준다. 반시민적 언어 표현에 대한 피드백 연구는 아직 초기 단계에 있지만, 다른 영역에서의 성공적인 시각적 피드백 연구를 바탕으로, 이 분야에서도 효과적인 피드백 방식을 개발하는 것이 가능할 것으로 예상된다.

3. 반시민적 언어 습관 개선을 위한 메타버스 시스템 개발

본 연구에서는 메타버스 내에서 이루어지는 음성 채팅을 실시간으로 분석하여 사이버불링의 근간이 되는 반시민적 언어 표현을 감지하고 이를 시각적으로 피드백해주는 메타버스 시스템의 구현을 목적으로 하고 있다. 본 시스템에서 감지하려고 하는 반시민적 언어의 유형은 무례함과 차별적 표현이며, 해당 언어가 감지되면 같은 공간에 있는 피해자와 가해자 모두에게 시각적 피드백을 제시함으로써 반시민적 언어 사용에 대해 셀프 모니터링이 가능하도록 하였다.

이를 위해 시스템은 그림3에 나타난 바와 같은 구현 단계를 갖추었다. 먼저, 메타버스 내에서의 아바타 사이의 가까울 때에만 음성 채팅에서만 반시민적 언어 표현에 대한 검출이 진행되도록 하기 위해 아바타 사이의 거리가 가상 공간 거리 기준 10m이내일 때에만 음성 채팅이 수집되도록 하였다. 수집된 음성은 구글의 Speech to Text(STT) API를 사용하여 텍스트로 변환되었으며 변환된 텍스트는 무례 및 혐오 표현 검출 모듈에 입력, 무례인지, 혐오인지, 둘 다 사용되었는지, 혹은 반시민적 표현이 아닌지에 따라 다른 색의 피드백으로 나타나게 된다.

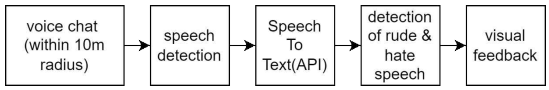

Figure 3. Metaverse system diagram for improving uncivil language habits

3.1 무례 및 혐오 표현 검출 장치

우리는 기존 연구에서 개발한 어텐션 기반 특징 추출 알고리즘과 멀티 채널 CNN을 사용하여 '무례' 및 '혐오 표현' 검출 장치로서 사용하였다. 다만, 기존의 연구에서 13가지 종류로 나오는 라벨을 모두 사용한 것이 아니라 '무례'와 '혐오' 두 가지로만 판별하도록 하였다. 즉, 무례, 폄훼, 공격적 어조, 불필요한 욕설 중 하나가 나오면 '무례'로, 지역차별, 장애차별, 성차별, 인종차별, 종교차별, 소수자차별, 직업차별, 연령차별, 정치차별 등이 나오면 '혐오'로 판별하였다.

3.2 메타버스에서의 시각적 피드백

구현된 메타버스에서는 두 가지의 방향으로 시각적 피드백이 이루어진다. 먼저, 사용자의 아바타 주변 반경 10m(가상현실 단위 기준)이내에서 반시민적인 언어 표현이 감지될 경우, 반시민적인 언어 표현을 쓴 아바타로부터 사용자의 아바타까지 특정 색으로 '길'이 나타나게 된다. 우리는 그 길을 '빛의 길'

이라고 명명하였다. 이 '빛의 길' 은 어떤 종류의 반시민적 언어를 사용하였느냐에 따라 다른 색으로 나타나게 되며, 길 뿐만이 아니라 같은 색의 이모티콘도 함께 나타나게 되었다.

그림 4에서 (a) 영역은 해당 현상을 보여주고 있다. 또한 이와 동시에 반시민적인 표현을 사용한 사용자의 화면에는 그림 4의 (b) 영역과 같은 빗살 무늬의 경고형 시각적 피드백이 발생하게 된다. 두 현상은 동시에 일어나게 되어 반시민적인 표현을 쓴 당사들자와 주변 사용자들 모두에게 반시민적인 표현 사용에 대한 경각심을 일으키게 된다.

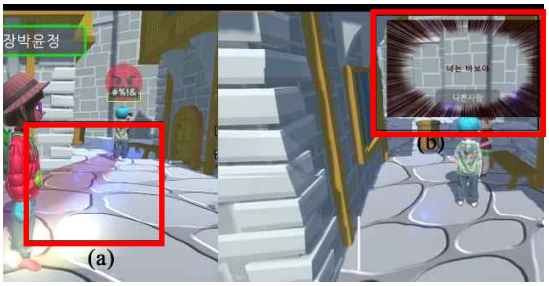

Figure 4. System provides visual feedback when it senses uncivil expression in voice chat

본 연구에서는 반시민적인 유형을 '무례', '차별' 두 가지로 하였기 때문에 이에 따라 나타내는 시각적 피드백도 이를 나타낼 수 있도록 여러 색으로 표현하였다. '무례'의 경우, 붉은 계열, '혐오'의 경우 남색 계열, '무례'와 '혐오'가 같이 발생할 경우에는 보라색 계열의 시각적 피드백을 사용하였고 반시민적인 표현이 없는 경우에는 옅은 파란색 계열의 시각적 피드백을 사용하였다. 표1은 반시민성 유형에 따라 시각적 피드백이 어떠한 형태로 나타나는지를 상세하게 나타내고 있다. 또한 아바타끼리의 거리가 3m이내로 더 가까워지면 어떤 언어를 사용하였는지도 말풍선의 형태로 확인할 수 있다.

3.3 전문가 피드백

메타버스 내에서 음성 채팅을 통해 이루어지는 반시민적 언어 표현을 실시간으로 분석, 반시민성 사용에 대한 피드백을 통해 의식 변화를 유도하는 메타버스 시스템을 평가하기 위하여 커뮤니케이션 분야의 전문가 패널을 구성하였다. 박사급 이상의 전문가 2인으로 구성된 패널들은 웹에 배포된 메타버스 플랫폼에 독립적으로 입장하여 실험자와 음성으로 소통하였다. 전문가에게는 실험 전 구글 문서를 통해 시나리오가 제공되었으며 여기에는 전문가 패널이 언급해야 할 반시민적 언어가 나타나 있어 실험 중 실험자에게 음성으로 언급하도록 하였다. 연구자 역시 반시민적 언어를 전문가 패널에게 사용하여 반시민적 언어를 직접 언어로서 경험하도록 하였다. 이 과정에서 사용된 반시민적 언어는 사전에 반시민성 척도에 다라 반시민성으로 분류된 온라인 댓글 데이터를 출처로 하고 있다.

패널들은 메타버스에 진입한 후 마이크를 활성화시키고 소통하였다. 패널들과 전문가들의 상호작용 후, 전문가 참가자들은 플랫폼의 시각적 피드백을 받은 경험에 대한 감정, 반시민적 행위에 대한 자각도 등을 설문조사를 통해 얻었다.

Table 1. Visual feedback according to the type of incivility

Type of incivility	road of light	emoticons
Rude		
Hate		
Rude + Hate		
not uncivil		

4. 결과 및 토의

본 연구는 메타버스 환경 내에서 반시민적인 언어의 사용을 실시간으로 감지하고 이에 대해 시각적 피드백을 제공하는 시스템의 효과를 평가하엿다. 이를 위해 커뮤니케이션 전문가 2인을 대상으로 실험을 진행하였으며, 다음과 같은 결과를 얻었다.

먼저, 모든 전문가 패널이 자신의 아바타 또는 다른 아바타가 반시민적인 언어를 사용하는 것을 경험했으

며, 이러한 행위는 플랫폼의 시각적 피드백을 통해 처음으로 인지되었다고 응답하였다. 이와 같은 인지는 자신의 반시민적인 언어가 단순히 '지시에 따른 행위' 였다 하여도 반시민적인 언어를 사용했다는 것에 대한 부끄러움과 죄책감을 유도한 것으로 나타났다. 또한 전문가들은 실험에 참가한 이후 자신의 반시민적 언행에 대하여 더욱 의식하게 되었다고 하였으며 이러한 피드백은 구현된 시스템이 시각화 기반 메커니즘으로서 동작할 수 있다는 가능성을 보여주었다.

5. 결론

본 연구에서는 메타버스 환경 내에서 발생하는 반시민적 언어 표현에 집중하여 실시간으로 이러한 표현을 감지하고 시각적으로 피드백해주는 시스템을 개발하였다. 이를 통해 메타버스 사용자들이 자신의 언어 사용에 대해 셀프 모니터링을 할 수 있게 되었으며, 이로 인해 의도하지 않은 반시민적 언어 사용을 줄이고 사이버불링을 예방하는 방향으로 기여하고자 하였다.

이를 위하여 기존 연구에서 사용하였던 멀티 CNN 어텐션 알고리즘을 적용하여 반시민적 언어 표현을 감지하였고, 감지된 결과를 바탕으로 사용자와 그 주변의 아바타에게 즉각적인 피드백을 제공하였다[22]. 다양한 색상의 시각적 피드백은 사용자들에게 그들의 언어 선택에 대한 직접적인 피드백을 제공하는 역할을 하게 되었다.

본 시스템의 구현은 메타버스의 안전성과 상호작용의 질을 향상시키는 것을 목적으로 하고 있다. 향후 이러한 시스템이 메타버스 환경에서의 의사소통 문화를 개선하는 데에 얼마나 기여할 수 있을지, 그리고 이러한 피드백 시스템이 사용자의 언어 사용 패턴에 어떠한 영향을 미치는지에 대한 심층적인 연구가 필요하다.

그에 대한 첫번째 연구방향으로, 메타버스를 주로 활용하는 연령층의 언어 표현 패턴을 깊게 분석하여 더욱 정확하게 반시민적 언어를 감지할 수 있는 데이터를 수집하고 무례와 차별 외 여러 차원으로 확장하여 분류할 예정이다. 또한 이렇게 수집된 데이터는 반시민성 검출 알고리즘의 학습에 활용되어 감지율의 향상을 도모할 것이다.

두 번째는 전문가들의 피드백을 바탕으로 청소년 및 대학생을 대상으로 언어 데이터 베이스를 확장한 후, 실험을 통해 본 시스템이 실제로 청소년들의 언어 사용 패턴의 변화를 가져오는지, 그리고 사용자들이 시스템에 얼마나 만족하는지를 평가하는 연구도 진행될 예정이다.

메타버스 플랫폼이 사회적으로 급속히 주목받는 가운데, 이러한 가상 공간에서의 사이버불링 현상이 현실의 심리적 피해와 동등한 수준으로 피해자에게 영향을 미칠 수 있다는 우려가 동반되어 왔다. 실제와 가상의 경계가 흐려지는 메타버스 환경에서, 사이버불링의 피해는 더욱 현실감 있게 다가오게 되며, 이로 인해 사용자들의 심리적 안정을 위협할 수 있다. 따라서 메타버스 내에서의 상호작용과 의사소통 품질을 향상시키는 연구의 중요성이 강조되고 있다. 본 연구는 이러한 문제의식을 바탕으로, 메타버스 환경에서의 시민성 및 공동체 의식 강화를 목표로 하여, 건전한 커뮤니케이션 문화 조성을 위한 기여를 하고자 한다.

참고문헌

[1] Lim, I., Lee, S., & Keum, H.. (2022). A Study on the Development of Scale for Online Incivility. *Korean Journal of Communication & Information, 116*, 215-249. DOI: 10.46407/kjci.2022.12.116.215

[2] Duggan, M.. (2014). Online harassment. *Washington, DC: Pew Research Center*. Retrieved from http://www.pewinternet.org/2014/10/22/online-harassment/

[3] Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*(4), 658-679. DOI : https://doi.org/10.1111/jcom.12104

[4] Kim, K., Cho, Y., & Bae, J. (2020) Exploratory Study on Countering Internet Hate Speech : Focusing on Case Study of Exposure to Internet Hate Speech and Experts' in-depth Interview. *JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 20*(2), 499-510.

[5] Chen, G. M. (2017). *Online incivility and public debate: Nasty talks*. New York, NY: Springer.

[6] Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research, 49*(3), 399-425. DOI : https://doi.org/10.1177/0093650220921314

[7] Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In Proceedings of the first *workshop on abusive language online*. 85-90. DOI : http://dx.doi.org/10.18653/v1/W17-301

3

[8] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-gru based deep neural network. *In The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings 15* 745-760. Springer International Publishing. DOI : https://doi.org/10.1007/978-3-319-93417-4_48

[9] Anderson, J., & Rainie, L. (2022). The metaverse in 2040. *Pew Research Centre*, 30.

[10] Pyun, K. R., Rogers, J. A., & Ko, S. H. (2022). Materials and devices for immersive virtual reality. *Nature Reviews Materials*, *7*(11), 841-843. DOI : https://doi.org/10.1038/s41578-022-00501-5

[11] Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M., ... & Wamba, S. F. (2022). Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 66*, 102542. DOI : https://doi.org/10.1016/j.ijinfomgt.2022.102542

[12] Han, J., Liu, G., & Gao, Y. (2023). Learners in the Metaverse: A systematic review on the use of Roblox in learning. *Education Sciences, 13*(3), 296. DOI : https://doi.org/10.3390/educsci13030296

[13] Lewis, T. M.. (2021). *Cyberbullying and Bystander Behavior Among Elementary School Aged Children*. Ph.D. Dissertation, Xavier University.

[14] Kim, E., & Yang, S. (2013). The New Citizenship of Digital Natives and the Influence of Network Media. *Korean Journal of Journalism & Communication Studies, 57*(1), 305-334.

[15] Searson, M., Hancock, M., Soheil, N., & Shepherd, G. (2015). Digital citizenship within global contexts. *Education and Information Technologies, 20*(4), 729-741. DOI : https://doi.org/10.1007/s10639-015-9426-0

[16] Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication, 11*, 3182-3202.

[17] Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*(2), 259-283. DOI : https://doi.org/10.1177/1461444804041444

[18] Quandt, T. (2018). Dark participation. *Media & Communication, 6*(4), 36-48. DOI : https://doi.org/10.17645/mac.v6i4.1519

[19] Lee. S. H. (2021). Biased Artificial Intelligence: Analyzing the Types of Hate Speech Classified by ‘Cleanbot’, NAVER AI for Detecting Malicious Comments. *Journal of Cybercommunication Academic Society, 38*(4). 33-75.

[20] Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2019). Incivility detection in online comments. *In Proceedings of the eighth joint conference on lexical and computational semantics*. 283-291. DOI : http://dx.doi.org/10.18653/v1/S19-1031

[21] Ozler, K. B., Kenski, K., Rains, S., Shmargad, Y., Coe, K., & Bethard, S. (2020). Fine-tuning for multi-domain and multi-label uncivil language detection. *In Proceedings of the Fourth Workshop on Online Abuse and Harms, Online*. 28-33. DOI : http://dx.doi.org/10.18653/v1/2020.alw-1.4

[22] Park, Y.,.Lee, S., Keum, H. (2022). , Detection of Incivility based on Attention-embedding and multi-channel CNN. *Journal of the Korea Institute of Information andCommunication Engineering, 26*(12), 1880-1889.

[23] Park, W., Kim, J., & Lee, J. (2020). A study on the design and effect of feedback for virtual reality exercise posture training. Journal of the Korea Computer Graphics Society, 26(3), 79-86. DOI : https://doi.org/10.15701/kcgs.2020.26.3.79

박 윤 정

2005년 연세대학교
전기전자공학과(공학사)
2010년 연세대학교
전기전자공학과(공학석사)
인지공학협동과정(공학박사)
2020년 연세대학교
인지공학협동과정(공학박사)
2020년~현재 성균관대학교 글로벌융복합콘텐츠연구소 선임연구원
관심분야: 머신러닝, HCI, 메타버스, 미디어 에이전트
E-Mail: shydeng@skku.edu

이 세 영

2006년 성균관대학교
신문방송학과(언론학사)
2008년 성균관대학교
신문방송학과(언론학석사)
2016년 State university of New York at Buffalo(커뮤니케이션학박사)

2019년~현재 성균관대학교 미디어커뮤니케이션학과 부교수
관심분야: Human Communication Behavior, Human-AI Agent Interaction, Social Influence, Compliance Gaining
E-Mail: gethemane@skku.edu

금 희 조

2004년 위스콘신대학교
미디어커뮤니케이션학박사

2006년~현재 성균관대학교 미디어커뮤니케이션학과 정교수
관심분야: 디지털 반시민성, 사회적 확산, 미디어 효과
E-Mail: hkeum@skku.edu