# 메타버스 환경에서 음성 혐오 발언 탐지를 위한 딥러닝 모델 설계

송진수[1], 딜노자[1], 손승우[2], 신용태[2]
[1]숭실대학교 컴퓨터학과
[2]숭실대학교 컴퓨터학부
iko153@soongsil.ac.kr, dkarabaeva2747@gmail.com, sons69@naver.com, shin@ssu.ac.kr

# Deep Learning Model for Metaverse Environment to Detect Metaphor

Jin-Su Song[1], Dilnoza Karabaeva[1], Seung-Woo Son[2], Young-Tea Shin[2]
[1]Dept. of Computer Science, Soong-Sil University
[2]Dept. of Computer Science and Engineering, Soong-Sil University

요 약

최근 코로나19로 인해 비대면으로 소통할 수 있는 플랫폼에 대한 관심이 증가하고 있으며, 가상 세계의 개념을 도입한 메타버스 플랫폼이 MZ세대의 새로운 SNS로 떠오르고 있다. 아바타를 통해 상호 교류가 가능한 메타버스는 텍스트 기반의 소통뿐만 아니라 음성과 동작 시선 등을 활용하여 변화된 의사소통 방식을 사용한다. 음성을 활용한 소통이 증가함에 따라 다른 이용자에게 불쾌감을 주는 혐오 발언에 대한 신고가 증가하고 있다. 그러나 기존 혐오 발언 탐지 시스템은 텍스트를 기반으로 하여 사전에 정의된 혐오 키워드만 특수문자로 대체하는 방식을 사용하기 때문에 음성 혐오 발언에 대해서는 탐지하지 못한다. 이에 본 논문에서는 인공지능을 활용한 음성 혐오 표현 탐지 시스템을 제안한다. 제안하는 시스템은 음성 데이터의 파형을 통해 은유적 혐오 표현과 혐오 발언에 대한 감정적 특징을 추출하고 음성 데이터를 텍스트 데이터로 변환하여 혐오 문장을 탐지한 결과와 결합한다. 향후, 제안하는 시스템의 현실적인 검증을 위해 시스템 구축을 통한 성능평가가 필요하다.

## 1. 서론

최근 코로나19로 인해 비대면으로 소통할 수 있는 플랫폼에 대한 관심이 높아지고 있다. 이에 따라 기존에 활용되던 인스타그램, 페이스북, 트위터와 같은 사회관계망서비스(SNS)와 더불어 가상세계의 개념을 도입한 메타버스(Metaverse) 플랫폼에 관한 연구가 활발하게 진행되고 있다. 메타버스 플래폼은 현실 세계 정보를 기반으로 가상 세계에서 다수의 사용자가 아바타를 통해 관계를 형성하고 현실과 유사한 다양한 경험을 할 수 있도록 구축된 환경이다[1,2]. 메타버스는 아바타 간 상호 교류하는 과정에서 SNS 같은 텍스트 기반의 소통도 가능하나 VR스틱을 활용해 음성과 동작, 시선 등으로 변화된 상호작용이 가능하다[3]. 음성을 활용한 소통이 증가함에 따라 다른 이용자의 아바타에게 욕설 또는 저속한 표현, 상대방 아바타의 신체, 외모, 취향 등에 대해 비하하고 모욕하는 음성 언어 폭력에 노출되고 있다. 현재 메타버스에서 발생하는 언어 폭력에 대한 대응 방식은 텍스트를 기반으로 하는 시스템이 있다. 기존 시스템은 다른 사용자에게 불쾌감을 줄 수 있는 표현은 특수문자로 대체하는 방식을 사용하여 혐오 표현을 제한한다.

그러나 기존에 사용되는 필터링 시스템은 텍스트에 대해 혐오 단어 탐지를 진행하기 때문에 음성 언어 폭력에 대해서는 제한이 없다. 또한, 사전에 정의된 특정 혐오 표현의 키워드를 사용하지 않고 은유적 으로 표현하는 음성 언어 폭력에 대해서는 분류하지 못하고 사용자에게 전달된다.

따라서 음성 언어 폭력을 제한하기 위해서는 기존 텍스트 기반의 혐오 표현 탐지 기법을 음성에 적용할 수 있어야하며 은유적인 혐오 표현을 탐지 할 수 있는 시스템이 필요하다.

이에 본 논문에서는 인공지능 기법을 활용해 음성 데이터의 감정을 추출하여 은유적 표현의 혐오 발언을 탐지하고 텍스트 변환 과정을 통해 혐오 및 외설적 표현을 분류 할 수 있는 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 혐오 표현 탐지 기법에 대해 살펴보고 제안하는 모델에 필요한 요구사항을 도출한다. 3장에서는 본 논문에

서 제안하는 모델을 설계하며 마지막 4장에서는 결론 및 향후 연구 과제를 제시한다.

## 2. 관련 연구

본 장에서는 기존 혐오 표현 탐지 기법에 대해 살펴보고, 그 결과를 기반으로 제안하는 모델의 요구사항을 도출한다.

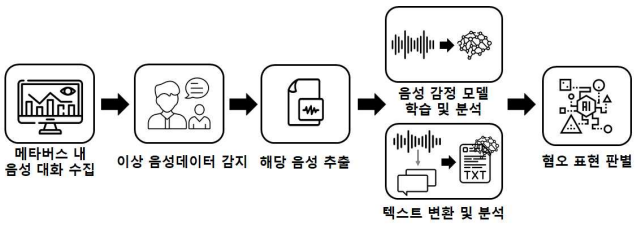
기존 혐오 표현 탐지 관련 연구는 인스타그램, 페이스북, 트위터와 같은 SNS에서 사용되는 문장을 수집해서 사용한다. 수집된 문장에서 혐오 단어를 추출하여 혐오 단어 사전을 생성하고 문장과 혐오 단어 사전을 비교하는 방식을 사용했다.

그러나 혐오 단어 사전을 활용한 방식은 문장의 문맥을 이해하지 못하고 단어의 존재유무로만 판단을 하기 때문에 단어 사전에 존재하지 않는 혐오 단어나 신조어로 생성된 혐오 단어가 존재하는 문장에 대해서는 정확도가 떨어지는 문제점이 있다[4]. 최근에는 문장을 이해도를 높이기 위해 딥러닝을 활용한 혐오 표현 탐지 연구가 진행되고 있다. 단어 사전에 의지하는 기존 연구 방식이 아닌 문장의 맥락을 고려하는 방식을 사용하면 문장에 사용되는 단어가 변하거나 순서가 달라도 탐지할 수 있는 방안에 대한 연구가 등장하고 있다[5].

## 3. 제안하는 음성 혐오 표현 탐지 시스템

본 장에서는 앞서 살펴본 모델을 기반으로 음성 혐오 표현 탐지 시스템을 제안한다.

[그림 1]은 제안하는 시스템의 구조와 기능을 나타낸다.



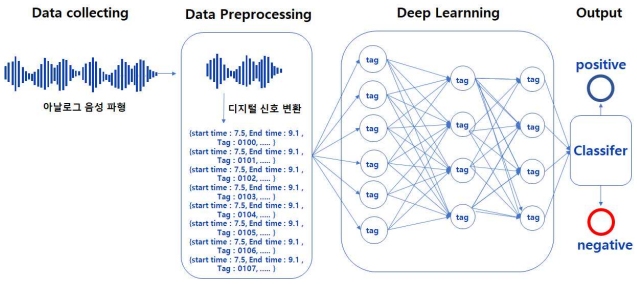

[그림 1] 제안하는 시스템 구조와 기능

제안하는 시스템은 메타버스 내 음성 대화를 수집하고 이상치를 초과하는 음성 데이터를 감지할 경우 해당 음성 데이터를 감정 분석 단계과 혐오 발언을 탐지하는 단계를 통해 혐오 표현 판별을 진행한다.

### 3.1 음성 데이터의 감정 분석 단계

아날로그 파형의 음성 데이터를 수집하고 분석하여, 감정적 특징을 추출하는 단계이다. 음성 분석은 수집된 음성 데이터에서 피치값을 추출하여 감정 변화를 특정한다. 음성 분석은 감정 변화의 피치값을 통해 진행되기 때문에, 우선적으로 평면에서 발생하는 데이터를 중점적으로 분석한다.

[그림 2]는 음성 신호의 감정 분석을 위한 인공지능 모델 설계도이다.

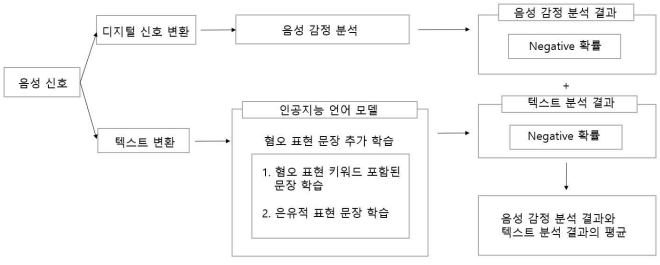



[그림 2] 음성 감정 분석 인공지능 모델 설계도

혐오 표현 탐지에 활용될 음성을 추출하고 수집된 음성 데이터를 가공하여 해당 음성 데이터와 감정 라벨을 인공지능 모델의 학습 데이터로 전달한다. 학습된 모델은 각 음성 파형마다 가중치를 부여하여 연관도를 분석하고 감정을 특정한다. 혐오 키워드를 포함하지 않는 은유적 혐오 표현은 음성 데이터에 감정적인 특징을 포함하고 있기 때문에 음성 데이터에서 분석되는 감정적 특징을 추출하여 분류에 활용한다.

### 3.2 혐오 발언 탐지 단계

혐오 발언 탐지 단계는 자연어 처리 기법을 활용하여 데이터를 처리한다. [그림 3]은 자연어 처리 기법을 적용하기 위해 텍스트 변환 및 분석 수행 단계이다.




[그림 3] 인공지능 혐오 발언 탐지 단계 순서도

디지털 음성 신호에 Speech to Text 기법을 적용하여 텍스트로 변환한다. 변환된 텍스트는 자연어 처리 모델의 입력으로 전달하여 분석을 수행한다.

자연어 처리 모델은 사전학습과 파인튜닝 단계로 이루어져있으며, 사전학습을 통해 블로그 게시물, SNS 댓글에 존재하는 기본 문장을 학습 시키고 튜닝 단계에서 혐오 문장 및 키워드를 추가 학습 시켜 기본 문장과 혐오 문장 판별 모델을 생성한다. 생성된 모델은 텍스트로 변환된 음성 문장에 대해 혐오 표현일 확률을 반환하고 음성 감정 수치와 결합한다.

## 4. 결론

본 논문에서는 인공지능 기법을 활용해 음성 데이터의 감정을 추출하여 은유적 표현의 혐오 발언을 탐지하고 텍스트 변환 과정을 통해 혐오 및 외설적 표현을 분류 할 수 있는 시스템을 제안한다. 제안하는 시스템은 인공지능 모델을 활용한 음성 데이터의 감정 분석 단계와 혐오 표현 탐지 단계로 구성한다. 음성 감정 분석 단계는 음성 데이터의 파형 분석을 통해 은유적 혐오 표현에 대한 감정적 특징을 확보한다. 혐오 표현 탐지 단계는 입력으로 전달받은 음성 데이터를 텍스트로 변환하여 혐오 키워드를 기반으로 혐오 문장을 파악하고 음성에서 추출된 감정적 특징과 결합하여 혐오 발언을 판별한다.

향후, 제안하는 시스템의 현실적인 검증을 위해 설계를 기반한 시스템 구축이 필요하다.

## ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2018000209)

## 참고문헌

[1] Soo J. Sohn.(2021).Intellectual Property(IP) Related Issues to be Addressed for Promoting Metaverse-based Co-creation.STEPI Insight, (), 1-46.

[2] Cho, Hee Kyung (2022). A Case Study on the Trans-branding Strategies in a Metaverse Environments. Journal of the Korean Society of Design Culture, 28(1), 451-465.

[3] Lee Hyewon.(2021).Art, Culture, and Human-centered Design in the Metaverse Era. Extra Archive: 디자인사연구,2(2),182-194.

[4] Jinju Hong, Sehan Kim, Jeawon Park, Jaehyun Choi.(2016).A Malicious Comments Detection Technique on the Internet using Sentiment Analysis and SVM.Journal of the Korea Institute of Information and Communication Engineering, 20(2),260-267.

[5] Wonseok Lee, Hyunsang Lee.(2020).Bias & Hate Speech Detection Using Deep Learning: Multi-channel CNN Modeling with Attention. Journal of the Korea Institute of Information and Communication Engineering,24(12),1595-1603.