JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 1, pp. 247-255, Jan. 2024

Check for updates

인공지능과 감정 반응: 디지털휴먼을 통한 혐오표현 감소 연구

김 정 민[1‡] · 최 정 우[1‡] · 서 경 진[2,3*]
[1‡] ㈜인공지능연구원 연구원
[2]㈜인공지능연구원 책임연구원
[3*]㈜아이리브 기술이사

Artificial Intelligence and Emotional Responses: A Study on the Reduction of Hate Speech through Digital Humans

Jung-Min Kim[1‡] · Jungwoo Choi[1‡] · Kyoung-Chin Seo[2,3*]

[1‡] Researcher, AIRI, Gyeonggi-do 13560, Korea
[2]Senior Researcher, AIRI, Gyeonggi-do 13560, Korea
[3*]CTO, AiLIVE Inc, Gyeonggi-do 13448, Korea

[요 약]

디지털 휴먼 기술의 중요성이 커지면서, 사람과 같은 대화의 필요성이 대두되었다. 이러한 맥락에서 개발된 챗봇 '이루다'는 혐오 표현 문제로 인해 사용자의 불쾌감을 야기했고, 혐오표현 탐지에 대한 여러 연구가 선행되었다. 하지만 온라인상에선 혐오표현 사용이 만연하다. 본 연구는 기존의 혐오표현 탐지와는 다른 측면에서 문제를 해결하려고 한다. 제안하는 방법은 혐오표현을 감지하고 사용자에게 거부감을 전달하여, 감정 전이를 이끌어내 혐오표현의 인지를 촉진시키는 것을 목표로 하는 디지털 휴먼 시스템을 제안한다. 효과성을 검증하기 위해 혐오표현과 디지털휴먼의 답변 영상을 제시하고 설문을 진행한 결과 5점중 3.95±0.65점을 획득했고, 혐오표현 문제 인식에 도움을 줄 수 있는 것을 확인했다. 연구결과를 토대로 소개하는 시스템이 혐오표현의 문제 뿐만 아니라 도덕적, 편향성, 범죄 문제에 대해서도 적용 가능하며, 사용자가 문제를 인지하는데 도움을 줄 것으로 기대된다.

[Abstract]

With an increase in the importance of digital human technology, the demand for human-like conversations has emerged. In this context, the chatbot "Lee Luda" was developed but caused user discomfort owing to its use of hate speech, prompting various studies on hate speech detection. However, the use of hate speech is rampant online. This study aims to address the problem from a different aspect than existing hate speech detection. The proposed method involved a digital human system designed to detect hate speech and convey discomfort to users, thereby facilitating emotional transfer and enhancing awareness of hate speech. To verify its effectiveness, a survey was conducted presenting the responses of the digital human to hate speech and scored 3.95±0.65 out of 5 points. Therefore, this method helped in recognizing the issues with hate speech. Based on the research findings, the introduced system is expected to be applicable not only to the problem of hate speech but also to moral, bias, and criminal issues and is expected to help users become aware of these problems.

색인어 : 디지털 휴먼, 디지털 휴먼 표정, 혐오대화, 혐오표현, 혐오표현 감소

Keyword : Digital Human, Digital Human Facial Expression, Hateful Conversation, Hate Speech, Reducing Hate Speech

http://dx.doi.org/10.9728/dcs.2024.25.1.247

Received 29 November 2023; **Revised** 12 December 2023
Accepted 15 December 2023

‡ These authors contributed equally to this work
***Corresponding Author; Kyoung-Chin Seo**
Tel: ※ 개인정보 표시제한
E-mail: kcseo@airi.kr

Ⅰ. 서 론

최근 메타버스와 디지털 휴먼의 부상으로 인해, 디지털 휴먼의 인간 같은 대화 능력이 중요해졌다. 이를 위해 대화형 챗봇 기술이 발전하였고, 이 중 '이루다'라는 챗봇이 개발되었다. 그러나 이루다는 학습 데이터셋에 포함된 혐오표현으로 인해 사용자에게 불편을 야기했다. 이 데이터셋은 익명 사용자들의 채팅에서 수집되었으며, 제약 없이 사용된 혐오표현이 포함되었다. 이로인해 혐오표현 필터링에 대한 연구의 필요성이 대두되었다[1]-[3].

기존 연구들은 주로 사용자 발언에서 혐오표현의 존재 여부를 판단하는 데 집중하였고, 온라인 환경에서의 혐오표현 문제는 여전히 만연하다. 온라인 환경에서 혐오표현에 대한 즉각적인 불쾌감 표출이 어렵고, 이로 인해 혐오표현의 사용이 빈번하게 발생한다. 특히, 온라인 상에서 사용자의 피드백에 기반한 감정 전이가 일어나기 힘든 구조적 측면이 이를 심화시켰다.

본 논문에서는 이 문제를 해결하기 위해 감정 전이라는 관점에서 혐오표현을 해결하려는 새로운 디지털휴먼 시스템을 제안한다. 이 시스템은 디지털 휴먼이 혐오표현에 대응된 적절한 감정을 음성과 표정으로 표현할 수 있도록 한다. 예를 들어, 사용자가 "20대는 결코 성인이라 볼 수 없는 미숙한 나이다."와 같이 연령에 대한 혐오감을 표시할 경우, 디지털 휴먼은 화난 음성, 표정으로 "그런 얘기는 너와 하고 싶지 않아"라고 답할 수 있다. 또는 "안 만나면 분노살인. 한국남자로 태어난 xx들을 다 여섯조각으로 찢어죽여야 해결될 문제노"라는 혐오 발언에 대해서는 슬픈 음성과 표정으로 "너의 마음은 이해가 가지만 괜히 상관없는 사람만 상처 받을 거야"라고 대응한다. 이러한 방식으로 디지털 휴먼은 혐오표현에 대한 감정적 반응을 빠르게 인식하고, 사용자의 피드백을 통해 계속 발전할 수 있다.

Ⅱ. 관련 연구

본 장에서는 혐오 표현 분석 및 제거에 대한 기존 연구와 이를 효과적으로 전달하기 위해 사용된 디지털 휴먼 기술과 효과에 대한 연구를 소개한다. 혐오 표현이란 "특정한 속성을 갖는 집단이나 개인에 대하여 공개적으로 비하, 비방, 모욕, 협박하거나, 차별, 적의, 또는 폭력을 선동하는 행위"를 의미한다[4]. 집단의 예시로는 여성, 남성, 성소수자, 종교, 지역 등이 대표적인 범주에 들어간다. 더 나아가 문맥상이나 특정 고유명사가 혐오표현이 되는 경우도 있다. 예를 들어 "무슬림 50퍼 근친이다.", "가짜 남자는 어떻게 생겼냐?", "진짜 전라도 존나 싫다"등의 문장에서 '무슬림', '근친', '출산 위험', '가짜', '남자', '전라도' 등의 단어는 일반적으로 쓰이지만 문맥상 특정 종교, 성 소수자, 지역에 대한 혐오를 보여준다(표 1).

표 1. 문맥상 혐오표현 예시
Table 1. Instances indicating hate speech in context

Hate-Speech Sentences	Hate category
가짜 남자는 어떻게 생겼냐?"	성 소수자
진짜 전라도 존나 싫다	지역
절라도가 낫냐? 무슬림이 낳냐?	지역/종교
무슬림 새끼들 인권보호해주는 유럽에는 존나 설치고 중국은 무서워서 건들지도 않네 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ	종교
이래서 남경 뽑으면 안돼	남성

*Hate speech meanings differ globally, requiring Korean for accurate conveyance.

표 2. 혐오 고유명사가 포함된 예시
Table 2. Hate speech related proper noun examples.

Hate-speech sentences	Hate-speech words	Hate category
한남 말동무도 해주고 심심하냐?"	한남	남성
문XX 페미에 짝짝꿍 나라가 망할 징조노 ㅋㅋ	페미	여성
어이구 시바 생쇼를 해요 양키XX가	양키	국가/인종
노친네들 잡아죽이는게애국	노친네	연령
무조건 스시녀랑 결혼한다 ㅋㅋ김치녀랑 결혼할 이유 없어짐	스시녀, 김치녀	여성, 인종/국적

*Hate speech meanings differ globally, requiring Korean for accurate conveyance.

특정 고유명사인 혐오표현의 예시는 '한남', '김치녀', '노친네', '양키' 등이 있다(표 2). 표의 경우 국가에 따라 사용되는 혐오표현이 다르기 때문에 한글로 표시했다.

혐오표현에 대한 반응을 위해 문장 임베딩을 사용한다. 문장 임베딩은 단어 기준 임베딩의 평균과 BERT[5]와 같은 언어모델의 토큰을 이용하거나, 문장 관계가 학습된 Sentence-Language Model[6]을 이용하는 방법 등이 있다. 단어 기준 임베딩의 평균은 각 단어 임베딩의 평균을 문장의 임베딩으로 표현하여 중의적 단어의 표현이 불가능한 단점이 있다. BERT계열 언어 모델의 경우, Self-Attention 메커니즘 기반으로 문장 단위 학습이 진행되어 문맥에 따른 단어 의미표현이 가능하다[7]. 다른 방법으로는 Fine-Tuning된 모델의 문장 임베딩을 사용하는 방법이 있다[8]. 본 논문에서는 정확한 유사도 비교를 위해 BERT기반의 문장 수준 임베딩을 사용하여 답변 사전에서 답변을 선택하도록 수행했다.

디지털 휴먼은 사용자에게 좋은 효과를 내기 위해서 다양한 방면으로 활용된다. 주변 사람의 감정을 공유하는 감정 전이를 이용하여 음악을 못 듣는 청각 장애인에게 음악의 감정을 디지털 휴먼의 표정을 통해 느낄 수 있게 하거나[9], 환자가 자신의 이야기를 편하게 말하기 어려울 때 자신과 닮은 디지털 휴먼을 이용하여 편하게 말하게 한다[10]. 더하여 디지털 휴먼을 통해 COVID-19로 인해 발생한 외로움을 경감시키려는 연구도 있다[11]. 디지털 휴먼의 긍정적 효과를 활용

하기 위해서 감정이 있는 음성을 디지털 휴먼의 얼굴로 만들어 주는 EmoTalk[12]을 이용하여 사용자에게 혐오표현에 대한 감정 전이를 통해 혐오표현의 부정적 의미를 전달하여 혐오표현 사용 빈도의 경감을 유도한다.

디지털 휴먼은 사용자에게 효과적으로 응답하기 위해 주로 음성 인터페이스를 사용한다. 이를 위해, Text to Speech (TTS) 모델을 통해 답변 문장을 음성 변환한다. TTS 기술은 단순히 문장을 음성으로 생성하는 것을 넘어, 감정을 표현하거나 특정인의 음성을 모사하는 능력까지 확장된다. Emotion TTS는 입력 문장에 감정을 반영한 음성 생성과 다양한 화자의 음성을 생성하는 데 집중한다[13]. 이를 위해 FastSpeech[14]를 활용하여 텍스트, 감정과 강도, 그리고 발화자 정보를 포함시켜 학습한다. 본 논문에서는 Emotion TTS를 사용하여 감정 전달에 초점을 맞추어 음성을 생성한다.

기존의 혐오표현 탐지 방법 연구는 혐오표현을 찾아 다른 이용자에게 보이지 않게 댓글이나 음성을 감추는 방식을 사용한다. 이러한 방법은 무분별한 혐오표현의 유출은 막을 수 있지만, 혐오표현에 대한 경각심을 일으켜 주지는 못한 채 온라인 상에서 지속적으로 사용되고 있다[15]. 더나아가 혐오표현에 대한 문제를 인식하지 못하고 그대로 사용하게된다.

제시하는 시스템에서는 혐오표현의 필터링 기능 뿐만 아니라 디지털 휴먼을 이용한 혐오표현에 대한 감정 전이를 이용한 피드백을 통해 사용자에게 혐오표현의 문제를 인지시켜 근본적인 문제를 해결하는데 도움을 주고자 한다. 3절에서 제안하는 방법의 시스템을 소개하고, 4절에서는 제안에 사용된 모듈들에 대해 설문을 통해 디지털 휴먼이 감정 전이를 활성시키는지에 대해 평가했다.

III. 혐오표현 감소를 위한 디지털 휴먼 시스템

제안 시스템은 대화 중 발생할 수 있는 혐오표현을 감지하고 반응하는 데 중점을 두며, 다음 4가지 요소로 구성된다. 혐오표현 분류기는 혐오표현을 식별하고, 그 결과를 바탕으로 답변 생성기는 답변을 선택 또는 생성한다. 음성 생성기는 답변에 대해 적절한 감정을 포함하여 음성을 생성한다. 이후 디지털 휴먼 생성기가 답변 음성을 기반으로 한 3차원 얼굴 표정을 생성하여 음성과 함께 사용자에게 제공한다. 이러한 과정을 통해, 시스템은 사용자와의 감정이 포함된 피드백을 통하여 혐오표현의 부적절성을 인지시킨다.

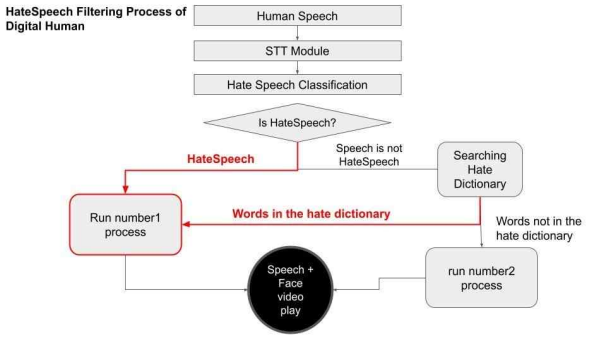


그림 1. 디지털 휴먼의 혐오표현 필터링 프로세스
Fig. 1. Hate speech filtering process of the digital human

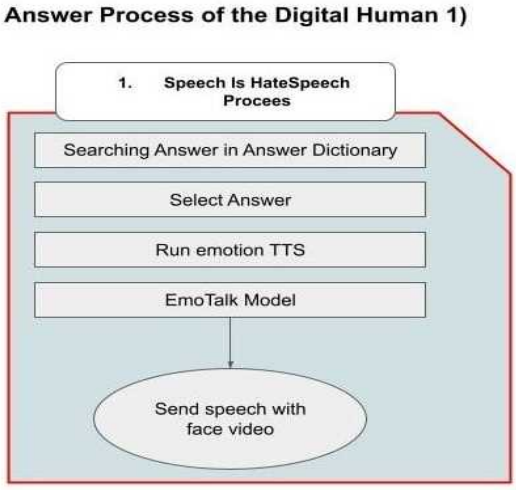


그림 2. 디지털 휴먼의 답변 프로세스 1
Fig. 2. Answer process of the digital human 1

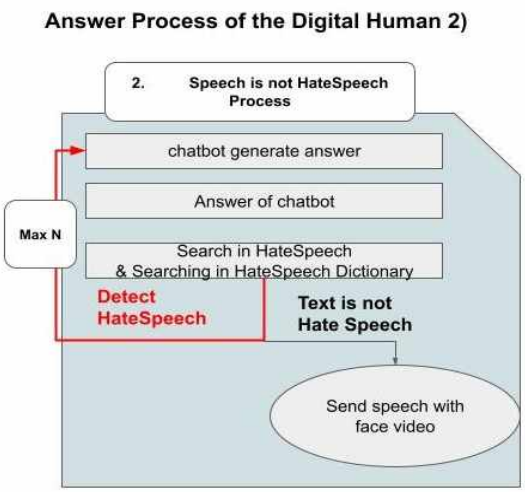


그림 3. 디지털휴먼의 답변 프로세스 2
Fig. 3. Answer process of the digital human 2

그림 1은 제안하는 시스템의 구성도이다. 그림 1에서 사용자의 발화를 Speech to Text(STT) 모듈을 이용하여, 텍스트로 변환 후, 혐오표현 분류기로 혐오표현 분류를 수행한다. 답변 생성기는 분류 결과가 9가지 혐오표현 분류 중 하나로 인식되면, 그림 2의 '프로세스 1'이 실행되며 이 단계에서는 답변 생성기가 답변 사전에서 적절한 답변과 감정 정보를 선택하게 된다. 그 후, 감정을 포함한 음성을 음성 생성기가 생성하게 된다. 생성된 음성은 혐오표현에 대한 거부감을 가지는 음성을 생성하게 된다. 혐오표현으로 분류하지 않는다면,

일반적인 소통을 위해 그림 3의 '프로세스 2'가 실행된다. 이 단계에서는 Chatbot모델이 답변을 생성하고, 음성 생성기를 통해 음성을 생성하게 된다. 마지막으로 음성을 기반으로 디지털 휴먼 생성기에게 전달한다. 디지털 휴먼 생성기는 음성의 감정 정보를 판단하여 디지털 휴먼 생성기가 표정 영상을 생성한다.

3-1 혐오표현 분류기

디지털 휴먼이 혐오표현에 대해 답변을 하기 위해, 사용자의 발화를 혐오표현 분류기가 혐오표현 여부를 결정한다. 혐오표현 분류기는 9가지의 혐오에 해당하는지, 해당하지 않는지에 대해, 분류하게 된다. 만약 '혐오 아님'으로 분류한다면, 혐오사전을 이용하여 혐오 단어 존재 여부를 확인한다. 혐오 단어가 존재할 경우 혐오 라벨로 간주하게 되며, 이는 새로운 혐오표현 단어에 대한 결과를 보정하기 위함이다(그림 4).

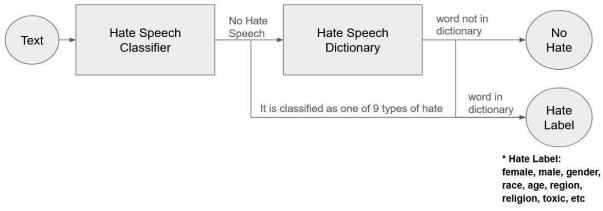


그림 4. 혐오표현 분류기의 구성도
Fig. 4. The configuration of the hate speech classifier

혐오표현 검출을 위해 학습 데이터로 Unsmile 혐오 데이터셋[16]과, HateScore 데이터셋[17]을 활용한다. 각 데이터에는 혐오표현에 대한 정보를 가지고 있으며, 부족한 중립 데이터를 보완하여 정확도를 높이기 위해 두 가지 데이터셋을 통합하여 사용한다.

1) Unsmile Dataset

'Unsmile 혐오 데이터 셋'은 단일 댓글 문장에 대해 3명의 작업자가 라벨링을 수행하였으며, 5인의 혐오표현 전문가가 최종 검수하여 데이터셋을 구축하였다. 구축 결과 "여성/가정, 남성, 성(성 소수자), 인종/국적, 연령, 지역, 종교, 기타, 악플/욕설, 혐오 아님"의 10가지의 혐오의 종류로 구성되어 있다[18]. 데이터 셋은 총 18,742개의 문장으로 이루어져 있으며, 혐오표현이 10,139개, 악플/욕설이 3,929개, 중립이 4,674개로 분포되어 있다. 단점은 사용자마다 다른 혐오의 기준으로 인해, 데이터의 태깅 정확성이 낮을 수 있다.

2) HateScore Dataset

HateScore 데이터셋은 2021년도 하반기 이후의 데이터를 포함한다. 3명의 작업자에 의해 레이블을 결정한 Unsmile과 달리 'Human-in-the-Loop(HILT) 방식의 확률'과 '연구원 한 명의 의견'의 두 가지 값을 활용했다. 데이터의 전체 문장 수는 약 1.1만 건으로 구성되어 있으며, Unsmile 데이터셋으로 학습한 모델이 태깅한 결과를 HILT 방식으로 수정한 문장이 1.7천 건, 위키피디아 혐오 관련 중립 문장이 2.2천 건으로 이루어져 있다. 기존 Unsmile 데이터셋의 혐오표현 문장과 중립 문장의 비대칭성을 보완하기 위해서 제안되었다.

3) Model

제안된 혐오표현 분류기는 DistilBert기반에서 학습된다[19]. DistilBert는 기존 BERT 모델을 경량화한 버전이다. 경량화를 위해서 두 가지 방법이 적용된다. 첫째, 모델에 입력되는 두 개의 문장에서 첫 번째 문장과 두 번째 문장이 연결 여부를 예측(NSP)하는 기능을 제거하여 하나의 문장만을 학습한다. 둘째, 입력 데이터의 크기가 단일 문장 사용으로 최대 입력 토큰 수의 감소를 통해 모델을 경량화한다. 미세 조정(Fine-Tuning)을 거친 모델을 이용하여 혐오표현을 분류한다. 또한 모델의 최종결과의 '[CLS]' 토큰을 문장 임베딩으로 사용하여 혐오표현의 문장을 비교한다.

3-2 답변 생성기

답변 생성기는 디지털 휴먼이 반응할 답변을 정하는 역할을 하며, 혐오결과에 따라 답변을 생성한다. 혐오표현으로 판단된 경우, 답변 사전을 활용하여 적절한 답변을 선정한다. 이를 위해 사용자 문장의 임베딩과 답변 사전에 포함된 대표 혐오 문장의 임베딩 간 코사인 유사도를 계산하며, 가장 높은 유사도를 보이는 답변을 선택한다. 만약 혐오 표현이 없는 경우, 일반 대화를 위해 ChatGPT와 같은 챗봇 모델을 사용하여 답변을 생성하게 되며, 감정은 Neutral, 감정강도는 Max로 정의하였다. 생성된 답변에 혐오표현이 포함되었는지 재검토하는 과정을 거쳐, 최대 N번의 샘플링 생성을 통해 답변에 혐오표현이 나타나지 않도록 한다(그림 5).

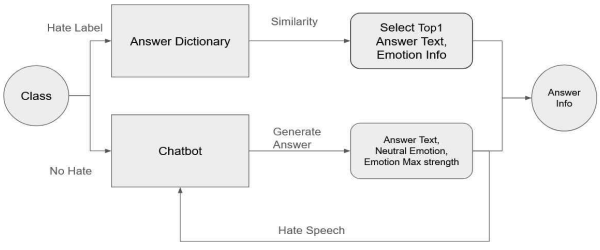


그림 5. 답변 생성기의 구성도
Fig. 5. The configuration diagram of the response generator

1) 답변 사전

답변 사전을 이용하여 사용자의 혐오 결과에 대한 디지털 휴먼의 답변을 선택한다. 답변 사전에서는 각 혐오 종류에 대한 대표 혐오 문장을 기반으로 중립적인 답변 세트를 구축했

다. 대표 혐오 문장은 범주별로 여성 47개, 남성 41개, 인종 35개, 성 소수자 44개, 지역 36개, 종교 34개, 욕설 54개, 연령 28개 등 총 344개로 선정되었다. 이를 바탕으로 62개의 중립적인 답변 문장을 데이터에 기반하여 작성하였으며, 각 답변 세트에는 화남, 슬픔 등의 감정 레이블이 포함되어 있다.

표 3. 답변 사전 예시
Table 3. Example of hate speech reaction dicitionary

Representative hate speech sentence	Index No.	Answer	Emotion	Level
국산남특징 정확하노ㅋㅋㅋㅋ지 잘못 인정안한다 일찍이 애비XX놈들이라 상도덕이 없다이기	2	너의 그런말은 천박하고 다른 {TAG1} 사람에 대해 상처를 주는 말이야 그러지마	Angry	2.0
이래서 남경 뽑으면 안돼	5	{TAG1}에 대해 안좋게 얘기하는건 좋지않은 모습 같아	Sad	1.5
여자가 뭘 입든 한남들따위가 시건방지게 입을 놀리노	8	선을 지켜줘 그런식이면 너랑 얘기하고 싶지 않아	Angry	1.8
이래서 냄져들에게 돈을 쥐어주면 안된다는거임 ㅉㅉ	34	난 모든 사람은 평등하다 생각해	Neutral	2.0

*Hate speech meanings differ globally, requiring Korean for accurate conveyance.

표 3은 선정된 대표 혐오 문장과 답변 정보로 구성된 사전을 보여주며, 62개의 답변 중 일부의 예시이다. 답변 사전의 동작은 사용자의 입력 문장과 가장 높은 유사도를 보이는 대표 혐오 문장을 기준으로 동작한다. 예를 들어 첫 번째 문장이 선택될 경우, 해당하는 2번 Index의 답변과 감정 정보가 EmotionTTS 시스템으로 전송된다. 이 방법은 추가 학습 없이 혐오표현에 대한 답변을 손쉽게 추가하고 수정할 수 있는 장점이 있으나, 정의된 답변에만 의존함으로써 답변의 다양성에 한계가 있다. 혐오의 종류에 따른 대상을 표현해주기위해 {TAG}를 사용한다. 예를 들어, 표 3의 2번 Index의 대상이 종교일 경우 {TAG1}을 '종교'로 대치하여 "너의 그런말은 천박하고 다른 종교 사람에 대해 상처를 주는 말이야 그러지마"라고 답하게 된다.

3-3 음성 생성기

음성 생성기는 답변 생성기로부터 전달받은 답변 정보를 Emotion TTS가 답변을 전달받아, 디지털 휴먼의 답변 감정을 생성하는 단계이다. Emotion TTS가 답변 정보의 감정을 이용하여 감정을 반영한 음성을 생성하게 된다. 음성 생성기에서 사용되는 감정 정보는 'Angry', 'Sad', 'Neutral'이 사용된다(그림 6).

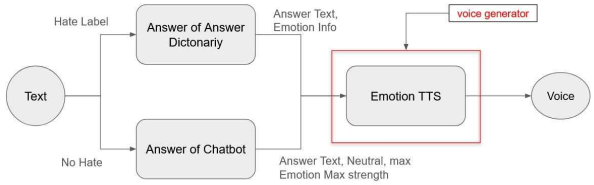


그림 6. 음성 생성기의 구성도
Fig. 6. The configuration diagram of the voice generator

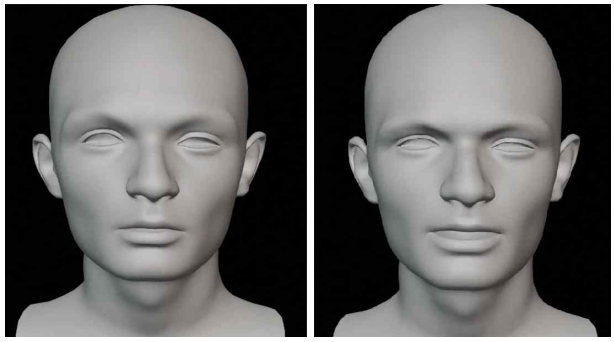


그림 7. 디지털휴먼의 표정 반응(좌: 연령 혐오에 대한 슬픔 / 우: 여성 혐오에 대한 화남)
Fig. 7. Digital human's reaction(left: sadness at ageism / right: angry at misogyny)

3-4 디지털 휴먼 생성기

최종적으로 사용자에게 보여질 디지털 휴먼의 영상을 생성하는 단계로 음성 생성기에서 생성된 음성 감정정보를 EmoTalk이 분석하여 표정이 있는 3D 디지털 휴먼을 디지털 휴먼 생성기가 생성하며, 최종적으로 디지털 휴먼의 영상을 사용자에게 전달하여 혐오에 대한 거부 반응을 하게 된다. 그림 7에서는 연령 혐오와 여성 혐오 사례에 대해 생성된 디지털 휴먼의 결과를 보여준다.

IV. 실험 결과

4-1 혐오표현 분류 정확도

본 시스템의 혐오표현 분류기에 대해 비교실험을 진행하였으며, 비교 metric은 F1-Score를 이용했다(표 4). Unsmile Baseline모델은 0.75로 제안하는 분류기의 0.73보다 높은 결과를 보인다. 반대로 HateScore Baseline 모델은 0.63으로 제안하는 분류기보다 낮은 성능을 보인다. 또한. 이세영[20]은 사전 학습된 사전학습 모델들을 Unsmile 데이터셋에 미세 조정하여 혐오표현 탐지에 대한 비교 연구를 수행하였으며,

이 결과와 제안하는 모델의 결과를 비교하였다. 논문에서 제시된 결과와 비교하였을 때, KoELECTRA[21]와의 비교에서 제안하는 모델이 높은 결과를 보였다. 하지만 KcELECTRA[22]에서는 제안하는 모델이 낮은 결과를 보이고 있다. Unsmile Baseline과 KoELECTRA 모델의 성능 차이는 경량화 모델인 DistilBert의 사용으로 인한 차이로 고려된다.

표 4. 혐오표현 분류기 실험결과
Table 4. Experiment results of hate speech classifier

Unsmile (BASE)	HateScore (BASE)	Ours	KoELECTRA [21]	KcELECTRA [22]
0.75	0.63	0.73	0.69	0.74

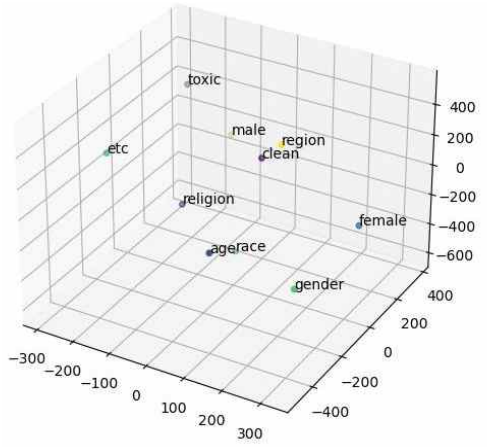


그림 8. 사전 학습된(Pretrain) DistilBERT의 혐오표현 데이터셋 카테고리별 임베딩 시각화
Fig. 8. Visualization of category-specific embeddings in the hate speech dataset using a Pretrained DistilBERT model

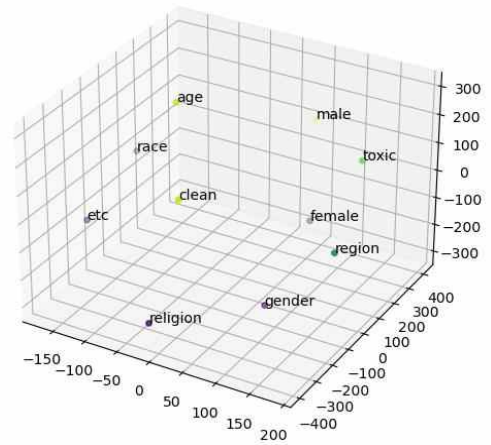


그림 9. Fine-tuning된 DistilBERT의 혐오표현 데이터셋 카테고리별 임베딩 시각화
Fig. 9. Visualization of category-specific embeddings in the hate speech dataset using a Fine-tuned DistilBERT model

그림 8, 9는 사전 학습 모델과 미세 조정 모델의 각 혐오의 종류에 대한 문장 임베딩의 평균을 구하여 학습 데이터의 혐오종류에 대한 차원 변화를 표현한 것이다. 이는 DistilBert가 혐오표현에 대해 학습이 의도한대로 학습되었는지를 보기 위함이다. 그림 8에서는 미세 조정 전 혐오의 종류와 관계 없이 차원의 거리가 구분된 것을 볼 수 있다. 예를 들면 'clean'(혐오 아님)과 'male'(남성 혐오), 'region'(지역 혐오)의 혐오는 비슷한 특징을 가지지 않음에도 거리의 차이가 적음을 보이는 반면, 그림 9의 미세 조정 후의 차원에서 'clean'과, 'male', 'region'의 거리의 차이가 커진 것을 확인할 수 있으며, 다른 종류의 혐오 또한 사전 학습 데이터의 혐오의 종류에 따라 거리가 구분된다. 이는 미세 조정 학습을 통해 혐오표현에 대한 차이가 학습됨을 보여준다.

4-2 음성 생성기 성능 평가

음성 생성기에 사용된 TTS의 성능은 설문을 통해 검증한다. 성능평가의 비교 대상은 구글 TTS API[23]와 비교하였다. 설문 방법은 10종의 문장에 대해 각 알고리즘의 음성을 생성하여, 총 20개의 음성에 대해 총 27명(남 20명, 여 7명)에게 다음 설문을 진행하였다. 설문 대상의 나이 분포는 20대 4명, 30대 18명, 40대 3명, 50대 2명으로 다양한 연령에 대해 답변을 수집하였다.

설문: 주어진 음성을 듣고 감정을 알아차릴 수 있었나요?

리커트 5점 척도를 감정 인지에 맞게 변형하여 평가를 진행하였다. 구글 TTS는 2±0.60, Emotion TTS는 3.81±0.71점을 받았으며, 이는 제안하는 방법에서 사용된 Emotion TTS가 사용자에게 감정을 전달하기 충분하다는 점을 보여준다.

4-3 혐오표현 감소 시스템 성능 평가

디지털 휴먼의 혐오표현에 대한 거부감을 표현한 음성 및 표정 정보가 혐오표현에 대한 심각성 인지에 도움이 되는지 평가하기 위해 온라인을 통해 설문을 실시했다. 실험대상자는 혐오표현에 대한 여러 관점을 보기 위해 다양한 성별, 나이대를 대상으로 진행했다. 대상자는 총 41명으로, 20대 6명, 30대 29명, 40대 3명, 50대 2명, 60대 1명, 성별 비율은 남성 26명, 여성 15명으로 구성됐다. 실험 방식은 5개의 혐오표현과 디지털 휴먼의 텍스트 답변 및 영상을 제시한 후 다음 의견을 조사했다.

설문 1. 다음의 혐오표현을 보고 문제가 심각하다고 생각하시나요?

설문 2. 아래는 위 혐오표현에 대한 답변입니다. 답변을 보고 혐오표현 문제의 심각성이 더 잘 느껴지시나요?

설문 3. 위 영상을 보시고 디지털 휴먼의 반응이 2번의 답변 텍스트보다 혐오표현이 잘못됐다는걸 느끼는데 더 효과적이라고 느끼시나요?

설문 1을 통해서 혐오표현의 심각성을 평가하도록 요청하였다. 설문 결과, 응답자의 81%가 혐오표현의 심각성을 인지하고 있음을 나타냈다(그림 10). 설문 2에서는 디지털 휴먼의 텍스트 답변을 먼저 제시한 후, 혐오표현의 심각성을 더 잘 인지하게 되는지를 조사하였다. 이 설문에서는 60%의 응답자가 텍스트 답변으로 인해 혐오표현의 심각성을 더 '잘 느끼게 되었다'고 답하였다(그림 11).

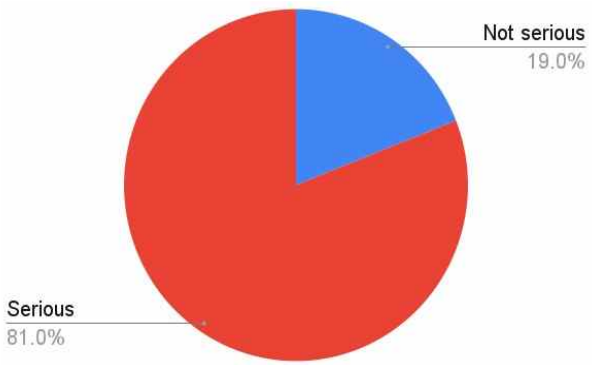


그림 **10.** "설문 1"에 대한 응답 비율
Fig. 10. Response rate for "Survey 1"

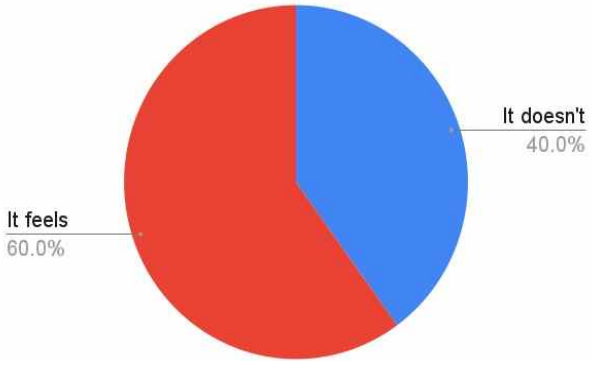


그림 11. "설문 2"에 대한 응답 비율
Fig. 11. Response rate for "Survey 2"

마지막으로, 설문 3은 디지털 휴먼의 영상과 텍스트 답변을 비교하여 혐오표현 인지에 대한 효과를 조사했다. 이 결과, 57.6%의 응답자가 영상이 텍스트보다 '더 효과적'이라고 응답하여 디지털 휴먼을 이용한 반응이 효과적임을 확인 할 수 있다(그림 12).

더하여, 설문 2에서 혐오표현의 심각성을 '잘 느끼지 못했다'고 응답한 28.8%가 설문 3에서 '효과적이었다'고 답했다. 이는 디지털 휴먼의 영상이 혐오표현 문제에 대한 인식을 강화시키는데 기여한다는 것을 시사한다. 나아가, 설문 1에서 혐오표현의 심각성을 '심각하지 않다'고 답한 21.6%가 설문 3에서 '효과적이다'라고 응답하여 혐오표현에 둔감한 사람에게 혐오표현에 대한 심각성을 일깨운다.

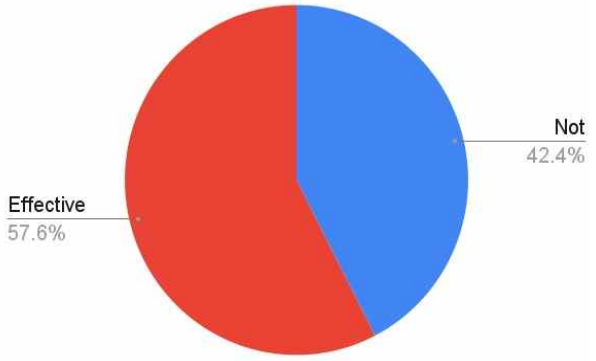


그림 12. "설문 3"에 대한 응답 비율
Fig. 12. Response rate for "Survey 3"

또한 디지털 휴먼의 혐오표현 인식에 대한 추가 설문을 진행하였다. 대상자는 총 42명으로 다양한 성별과 나이대로 구성하였다. 나이대는 20대 9명, 30대 24명, 40대 3명, 50대 4명, 60대 2명이고, 성별은 남성 18명 여성 24명이다. 이전의 설문과 동일한 혐오표현과 아래의 설문을 제시하였다.

추가 설문. 위(디지털휴먼의) 영상을 보시고 해당 혐오표현이 문제가 된다고 인식하는데 어느 정도 도움이 되셨나요?

해당 설문이 가질 수 있는 최대점수는 5점으로 설문 결과 3.95±0.65점을 얻었으며, 이는 4점 도움이 된다에 근접하여 디지털휴먼 영상이 혐오표현 문제 인식에 도움을 주는 것을 확인 할 수 있다.

본 논문에서 소개하는 혐오표현 시스템은 사용자 설문을 통해 혐오표현에 대한 인지를 하는데 '효과적'이라고 답한 설문자가 전체의 약 58%가 응답한점과, 시스템의 효과에 대해 수행한 추가 설문의 결과 높은 점수로 평가되었다. 이는 디지털 휴먼의 음성과 표정이 감정 전이에 기반한 혐오표현의 문제인식 강화를 확인하였으며, 근본적인 혐오표현의 사용 감소가 기대된다.

V. 결 론

본 연구에서 제시된 '혐오표현 감소를 위한 디지털 휴먼 시스템'은 익명성이 보장되는 온라인 환경에서 혐오표현 문제 해결을 목표로 한다. 이 시스템은 혐오표현에 대해 디지털 휴먼이 음성과 표정으로 거부감을 표현하며 반응하도록 설계되었다. 이를 통해 혐오표현에 대한 부정적 감정 전이를 일으키고, 혐오표현 문제에 대한 인식개선을 목적으로 하였다. 실시

한 설문을 통해 이러한 방식이 텍스트 답변에 비해 혐오표현 인식에 더욱 효과적임을 확인하였다.

설문 결과, 약 58%에 해당하는 다수의 응답자들이 디지털 휴먼의 반응이 혐오표현 문제 인식에 도움이 되었다고 답변했다. 이는 디지털 휴먼을 활용한 시스템이 혐오표현에 대한 문제 인식 향상에 기여할 수 있음을 시사한다. 또한, 본 연구의 결과는 답변 사전의 수동 생성이 아닌 초거대 언어 모델(Large Language Model)을 통한 답변 생성 및 분류 기능을 통해, 도덕적 문제, 범죄, 편견 등 다양한 사회적 문제에 대한 인식개선에 기여할 수 있는 가능성을 제시한다.

이러한 연구의 확장은 온라인상에서 발생하는 다양한 사회적 문제에 대한 인식을 개선하고, 이에 대응하는 효과적인 방안을 제시하는 데 중요한 역할을 할 것으로 기대된다. 본 연구는 디지털 휴먼의 활용이 단순한 대화 이상의 사회적 가치를 창출할 수 있는 새로운 패러다임을 제시함으로써, 온라인 커뮤니케이션과 인공지능 기술의 발전에 기여할 것이다.

감사의 글

본 논문은 문화체육관광부 및 한국콘텐츠진흥원의 2021년 문화콘텐츠 R&D 전문인력 양성(문화기술 선도 대학원) 사업으로 수행되었음 (과제명: 버추얼 프로덕션 기반 콘텐츠 제작 기술 R&D 전문인력 양성, 과제번호: R2021040044, 기여율: 80%)

참고문헌

[1] W. Lee and H. Lee, “Bias & Hate Speech Detection Using Deep Learning: Multi-Channel CNN Modeling with Attention,” *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 24, No. 12, pp. 1595-1603, December 2020. https://doi.org/10.6109/jkiice.2020.24.12.1595

[2] K.-S. Park, J.-S. Lee, and J.-H. Kim, “Korean Hate Speech Detection with CNN-LAN and Vulgar Words,” *Journal of Korean Institute of Intelligent Systems*, Vol. 33, No. 1, pp. 45-52, February 2023. https://doi.org/10.5391/JKIIS.2023.33.1.45

[3] M. Shin, H. Chin, H. Song, J. Choi, H. Lim, and M. Cha, “Hate Speech Detection in Chatbot Data Using KoELECTRA,” in *Proceedings of the 33rd Annual Conference on Human and Language Technology*, Online, pp. 518-523, October 2021.

[4] S. H. Park, “The Definition and Regulation Method of Hate Speech,” *Kookmin Law Review*, Vol. 31, No. 3, pp. 45-88, February 2019. https://doi.org/10.17251/legal.2019.31.3.45

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805, October 2018. https://doi.org/10.48550/arXiv.1810.04805

[6] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, pp. 3982-3992, November 2019. https://doi.org/10.18653/v1/D19-1410

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, “Attention is All You Need,” in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach: CA, pp. 6000-6010, December 2017. https://doi.org/10.48550/arXiv.1706.03762

[8] Y. Yun, Document Embedding and Classification Using BERT, Master’s Thesis, Pusan National University, Busan, February 2022.

[9] S. Kwon and J. Kim, “Enhancing Music Listening Experience Based on Emotional Contagion and Real-Time Facial Expression Retargeting,” *Journal of Digital Contents Society*, Vol. 20, No. 6, pp. 1117-1124, June 2019. https://doi.org/10.9728/dcs.2019.20.6.1117

[10] M. Ebnali, N. Ahmadi, E. Nabiyouni, and H. Karimi, “AI-Powered Human Digital Twins in Virtual Therapeutic Sessions,” in *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, Orlando: FL, pp. 1-4, March 2023. https://doi.org/10.1177/2327857923121000

[11] K. Loveys, M. Sagar, I. Pickering, and E. Broadbent, “A Digital Human for Delivering a Remote Loneliness and Stress Intervention to at-Risk Younger and Older Adults during the COVID-19 Pandemic: Randomized Pilot Trial,” *JMIR Mental Health*, Vol. 8, No. 11, e31586, November 2021. https://doi.org/10.2196/31586

[12] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, ... and Z. Fan, “EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 20687-20697, October 2023. https://doi.org/10.48550/arXiv.2303.11089

[13] Github. Multi Speaker Multi Emotion TTS [Internet]. Available: https://github.com/emotiontts/emotiontts_open_db/tree/master/Codeset/realtimeMultiSpeakerMultiEmotionTTS.

[14] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “FastSpeech: Fast, Robust and Controllable Text to Speech,” in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, pp. 3165-3174, December 2019. https://doi.org/10.48550/arXiv.1905.09263

[15] K. Y. Lee, Discrimination and Disparaging Expression of Hatred, Focusing on DC Inside and Ilbe, IT Chosun [Internet]. Available: https://it.chosun.com/news/articleView.html?idxno=2023092101906.

[16] Github. SmileGate: Korean UnSmile Dataset [Internet]. Available: https://github.com/smilegate-ai/korean_unsmile_dataset.

[17] Github. HateScore : Human-in-the-Loop and Neutral Korean Multi-label Online Hate Speech Dataset [Internet]. Available: https://github.com/sgunderscore/hatescore-korean-hate-speech.

[18] T. Y. Kang, E. Kwon, J. Lee, Y. Nam, J. Song, and J. K. Suh, “Korean Online Hate Speech Dataset for Multilabel Classification - How Can Social Science Improve Dataset on Hate Speech? -,” arXiv:2204.03262, April 2022. https://doi.org/10.48550/arXiv.2204.03262

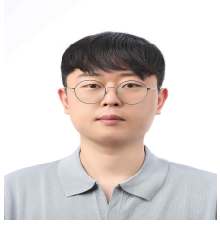
[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,” arXiv:1910.01108, October 2019. https://doi.org/10.48550/arXiv.1910.01108

[20] S. Lee and S. Park, “Analyzing the Classification Results for Korean Hatespeech and Bias Detection Models in Malicious Comment Dataset,” *Journal of the Korean Institute of Industrial Engineers*, Vol. 48, No. 6, pp. 636-643, December 2022. https://doi.org/10.7232/JKIIE.2022.48.6.636

[21] J. W. Park. KoELECTRA: Pretrained ELECTRA Model for Korean Github Repository [Internet]. Available: http://github.com/monologg/KoELECTRA.

[22] J. B. Lee. KcELECTRA, Github Repository [Internet]. Available: https://github.com/Beomi/KcELECTRA.

[23] P. N. Durette. gTTS. [Internet]. Available: https://gtts.readthedocs.io/en/latest/.



김정민(Jung-Min Kim)

2022년 : 국민대학교 대학원 (공학석사 - 컴퓨터공학, 자연어처리)

2022년～현 재: ㈜인공지능연구원 인터렉션연구실 연구원

※관심분야 : 자연어처리(Natural Language Processing), 언어모델(Language Model), 디지털 휴먼(Digital Human), 인공지능(Artificial Intelligence)

최정우(Jungwoo Choi)

2020년 : 국민대학교 대학원 (공학석사 - 컴퓨터공학, 컴퓨터비전)

2015년～2016년: Percolata Algorithm team Intern

2020년～현 재: ㈜인공지능연구원 인터렉션연구실 연구원

※관심분야 : 컴퓨터비전(Computer vision), 얼굴 표정 이식기술(Facial expression translation), 인공지능(Artificial Intelligence)

서경진(Kyoung-Chin Seo)

2001년 : 서강대학교 대학원 (공학석사)

2011년 : 서강대학교 대학원 (미디어공학박사)

2012년～2018년: 네이버 커넥트재단(구. NHN NEXT) 교수

2018년～현 재: ㈜인공지능연구원 인터렉션연구실 책임연구원

2023년～현 재: ㈜아이리브 기술이사(CTO)

※관심분야 : 컴퓨터 비전(Computer Vision), 디지털 휴먼(Digital Human), 인간컴퓨터상호작용(Human Computer Interaction), 인체 자세 추정(Body Pose Estimation)