

유머 자동 처리를 위한 유머 데이터 평가 및 활용

강조은¹⁰, 이재원¹, 오채은², 김한샘^{1†}

연세대학교 언어정보학협동과정¹, 연세대학교 법학과²
{j0eun⁰, jwon.lee70, oh0880, khss[†]}@yonsei.ac.kr

Evaluate and Use of humor data for humor processed automating

Kang Joeun¹⁰, Lee Jaewon¹, Oh Chaeun², Kim Hansaem^{1†}

Yonsei University, Interdisciplinary Graduate Program of Linguistics and Informatics¹,
Yonsei University, Graduate School Department of Law²

요약

의사소통 기술에서 유머는 사람을 웃게 만들며 분위기를 환기시키고, 관계를 돈독하게 만드는 효과를 지닌다. 이를 자연어처리에서 유머 분류, 인식, 탐지로 적용하여 유머를 기계에 학습시키려 하는 다양한 시도가 진행되고 있지만 유머의 주관성과 윤리적 문제로 탁월한 성능을 기록하기 어렵고, 특히 한국어 유머에 대한 자연어처리 분야의 논의는 미비한 상태이다. 이에 본 연구는 유머 평가 체계를 만들어 ChatGPT에 적용하여 유머 인식의 주관성을 극복할 수 있는 자동화 실험을 진행한다. 이때, 유머의 윤리적 문제를 보완하기 위해 한국 법률을 적용한 윤리 기준을 도입하여 유머 데이터셋을 마련하였으며, 데이터셋을 ChatGPT에 fine-tuning 하여 재미있는 생성 모델의 개발 가능성을 실험하였다.

주제어: 유머 분류, 유머 인식, 유머 평가, 유머 생성

1. 서론

LLM을 기반으로 개발된 ChatGPT, Bard, Bing 등의 챗봇은 전문가 집단뿐 아니라 대중적인 사용성을 보인다. 계속해서 AI와 챗봇의 접근성이 특정 도메인이나 전문가 집단에 한정적이지 않고 일상 대화, 상담의 영역까지 확대될 가능성이 높다면 AI와도 사람처럼 재미있는 이야기를 나누고, 웃을 수 있도록 하는 개발 방향의 논의가 필요할 것이다.

유머는 의사소통 상황에서 화자의 감정 이완과 더불어 대화 참여자와의 관계까지 증진시킨다. 그런데 유머는 학자마다 그 정의와 분류 체계가 상이해 자연어처리 분야의 관련 태스크들도 난도가 높다. 유머와 관련한 자연어처리 분야의 하위 태스크는 분류, 인식, 탐지가 있다. 유머 분류(humor classification)는 유머러스한 문장인지 아닌지를 분류하는 태스크이며, 유머 인식(humor recognition)과 유머 탐지(humor detection)는 유머 표현, 대상 등의 세부 속성을 정의하여 주석한 뒤 머신러닝/딥러닝 기법을 적용해 유머에 해당하는 표현을 찾아내는 것이다. 하지만 이러한 접근은 유머의 주관적인 기준 때문에 그 정확도와 신뢰도를 평가하기 쉽지 않다.

이에 본 연구에서는 유머에 대한 주관성을 극복하기 위해 LLM을 기반으로 한 ChatGPT를 사용하여 유머 평가를 도입하고자 한다. 유머 평가는 텍스트 안에 유머의 유무를 파악하며 윤리적 기준에 부합한 문장인지 판단하는 것이다. 이러한 유머 평가 과정은 ChatGPT와 인간 평가자를 활용한 실험으로 유머의 주관성을 극복할 수 있을 것이라 기대된다. 이후 평가 결과로 분류된 유머 텍스트를 미세 조정(fine-tuning)하여 재미있는 문장을 생성하는 모델 개발의 가능성을 실험하고자 한다.

2. 관련 연구

국내외 연구에서 유머와 관련한 용어로 humor와 joke 등을 혼용한다. 유머의 정의, 기능, 유형 분석은 주로 언어학 분야에서 다루고 있으며, 자연어처리 분야에서는 정의와 유형을 활용한 유머 분류, 인식, 탐지를 진행해 자동화를 이루는 것을 목표로 한다.

한편, 한국어를 대상으로 한 유머 연구는 언어학, 국어학 분야에서 주로 이루어져 왔고, 자연어처리 분야의 연구는 찾아보기 어렵다. 따라서 자연어처리를 활용한 유머 연구는 국외 사례를 조사하여 한국어 적용 가능성을 살펴보았다.

2.1. 유머의 언어학적 접근

유머에 대한 정의와 유형 구분은 연구자마다 상이하다. 그래서 유머를 다루는 연구자들은 연구의 대상 및 범위를 확정하기 위해 연역적/귀납적인 방법으로 정의한 유머의 정의를 논문에 포함한다. [1]은 남을 웃게 하기 위한 다양한 말과 행동을 유머로 정의하며, 화법 측면에서 적절한 유머 활용을 유머의 실행 조건과 원리로 분석하였다. 유머의 실행 조건인 ‘의도성’은 웃음을 유발하고자 하는 의도로, 남을 웃게 하기 위한 의도성이 있다면 유머의 성공 실패 여부는 상관없다고 주장한다. 유머의 원리는 유머의 내용을 구성하고 이해할 수 있는 인지적 기준을 바탕으로 ‘인지적 불일치 상태’일 때 나타난다. 즉, 유머는 예측을 벗어나는 상태에서 성립된다. 또한 유머의 원리로 우월감도 설명하면서 대화 참여자 관계에서 상대를 공격하여 유머로 작용하는 현상을 설명한다. 이때 우월감은 청자의 심리적 안전이 보장되어야 하며, 공격의 형태로 실현되지만 실제로 청자를 공격해

서는 안 된다는 조건을 가진다. [2]는 [1]의 인지적 불일치와 유사한 맥락으로 ‘비예측성’을 말하며, 비예측성에만 초점을 맞춘 기존 연구들과 다르게 유머 담화의 성립 요건을 ‘생성 기제’, ‘제약 조건’, ‘웃음 유발 장치’로 확대한다.

본 연구에서는 선행 연구를 바탕으로 ‘흥미성’을 유머의 대표 속성으로 정의하였다. 이는 개인이 판단할 수 있는 유머러스함 그 자체이며, 웃음을 유발하기 위한 의도를 갖고 인지적 불일치성에 기반한 비예측성을 포함한다. 그래서 유머 데이터의 평가를 진행할 때 ‘웃음 유발 장치’와 ‘의외성(=비예측성)’을 가진 문장을 유머의 구체적인 실현 조건으로 삼았다.

[3]은 비도덕적인 유머에 대한 도덕적 평가가 균일하지 않다는 것을 지적하였다. 이는 사람들이 유머의 ‘비도덕성’을 인지하더라도 웃을 수 있다고 여기는 것과 비도덕적이기 때문에 재미가 없어지는 것 등 여러 반응이 나타나는 것에 대해 살펴보았다. 구체적으로 미적 조건과 관련한 철학자들의 이론을 적용하였다. 그중 거트(Berys Gaut)는 윤리주의를 주장하여 비도덕성을 포함하지만 우스울 수 있으며, 비도덕성이 농담의 효과를 해치는 방향으로 나타남을 인정한다. 이와 관련하여 본 연구에서는 도덕성을 포함하지만 우스울 수 있는 유머와 상대방을 공격하며 나타날 수 있는 유머 유형을 ‘비꼬기/비판’으로 보아 ‘흥미성’의 세부 조건으로 추가하였다. 또한 유머 데이터 평가 요소에 ‘위험성’ 개념을 별도로 정의하여, 비도덕성이 농담의 효과를 해치는 것을 인식할 수 있도록 하였다. 위험성은 구체적으로 ‘공격성’, ‘선정성’, ‘혐오’로 구분하여 그 실현 조건을 정리하였다. 이렇게 선행 연구를 통해 정리한 유머 평가 요소는 ChatGPT를 이용한 LLM 자동 평가에 사용되었다. 자동 평가에 사용한 프롬프트는 <표 1>과 같다.

<표 1>: 평가 요소 프롬프트

{text}대해 유머 평가를 진행합니다.

평가 요소는 1) 흥미, 2) 위험성 입니다.

<평가 방법>

- 먼저 주어진 text가 흥미로운 유머 문장인지 판단합니다.
- 유머는 웃음을 유발할 수 있는 요소를 포함한 문장을 의미합니다.
- 일반적인 사고 흐름에 벗어난 의외성을 가진 문장도 유머가 될 수 있습니다.
- 유머는 종종 비판, 비꼬기 등의 어조를 포함하기도 합니다.

- 다음으로 주어진 문장에 대한 위험성을 평가합니다.

- 문장 안에 공격성, 선정성, 혐오 표현 등이 포함되어 있는 지 판단합니다.

2.2. 유머의 자연어처리 적용

유머의 정의가 학자마다 다르기 때문에 자연어처리 분야의 연구에서는 기존 정의, 유형보다 확장된 범위를 다루는 경향이 있다.

[4]에서 기존 이론은 유머의 조건, 유형을 정의하고 있어 그 이외의 실행 조건과 유형에 대해 설명하지 못한다고 지적하였다. 그래서 자동으로 감지할 수 있는 유머의 특성을 대분류인 Mode, Theme, Topics로 나누고 각각의 소분류로 정리하였다. 트위터와 Reddit 데이터를 수집하여 실험을 진행하였으며, ‘SVM+Sentiment Scores+N-grams’을 사용하였을 때 77%의 정확도를 기록하였다.

유머의 형식적 특징으로는 자동 분류가 어렵기 때문에 [5]에서는 사람 주석(human annotation)을 사용하여 유머 데이터를 구축하였다. 이는 SemEval-2015 Task 11에서 다룬 비유적 표현에 대한 작업(감성 분석)과 Semeval-2017 Task 6의 소셜 미디어 텍스트 연구 방법과 다른 방법을 사용한 것이다. 사람 주석에 기반한 HaHa-Humor Analysis를 제안하여, 스페인어로 작성된 트위터 데이터를 수집하였다. 수집한 데이터는 BERT를 이용해 두 가지 작업(유머 판단과 유머 점수 매기기)을 학습하였다. [6]에서도 사람 주석을 사용하여 유머를 주석 및 평가한다. 유머 텍스트 분석은 유머스러운 문장인지 아닌지 정도만 다루기 때문에 사람 주석을 통해 ‘왜 재미있는지.’ 시나리오를 포함한 데이터셋을 구축하였다. 데이터는 Relationship, Scene, Category, Level, DataSource, Keyword에 대해 주석하였고, 8명의 대학원생과 1명의 박사과정생의 평가로 완성하였다.

[7]은 기존의 유머 인식 기술에서 사용한 미세 조정(fine-tuning)이 아닌 프롬프트(“setup+punchline”)를 통한 유머 인식 실험을 진행하였다. 실험은 모델에게 빈칸 채우기(This is [MASK])를 통해 ‘재미있다, 보통이다’의 출력을 나타내도록 하였다. 데이터는 유머를 8가지 범주로 유형화한 HaHackathon Dataset(HHD) 데이터를 사용해 Gender와 관련한 유머까지 추출하여 Gender Humor Dataset(GHD)를 만들었다. 전체 데이터와 few-shot으로 주석 정보를 제한하면서 미세 조정과 프롬프팅 방법을 실험하였을 때, 유머와 공격성을 식별하는 데 프롬프팅이 효과적이라는 것을 밝혔다.

유머의 난해함을 극복하기 위해 도입한 자동화 방안 중 [7]의 연구 방법을 차용하여 ChatGPT를 활용한 LLM 유머 평가를 진행한다. <표 1>과 같은 프롬프팅을 통해 흥미성과 위험성을 기준으로 문장에 대한 유머를 평가하고 분류한다. 이때 [5]와 [6]의 방법처럼 사람 주석을 사용하여 자동 평가된 데이터를 인간 평가자의 수동 평가와 비교한다. 본 연구에서는 기구축된 영어 데이터를 기계 번역하여 사용하기 때문에 구축 단계에서 시행하는 사람 주석이 아닌, 수동 평가 방법으로 사람 주석을 사용하였다.

따라서 본 연구에서는 유머의 주관성과 난해함을 극복하기 위해 유머 평가 및 분류를 프롬프팅으로 자동화하고, 사람 주석으로 ChatGPT 결과의 실효성을 검증하였다.

3. 실험 과정

본 연구에서는 한국어 유머 데이터의 부재로 영어 데이터를 기계 번역하여 사용하였다. 전체 실험 과정은 한국어로 기계 번역한 데이터를 프롬프팅 하여 유머에 대한 자동 평가와 동시에 인간 평가자의 수동 평가를 진행한다. 모든 평가 결과를 종합해 유머 텍스트를 분류하고, 이후 분류된 데이터를 이용하여 ChatGPT 미세조정을 진행해 유머 생성 가능성을 확인한다.

3.1. 데이터

본 연구에서는 3가지 영어로 된 유머 데이터를 한국어로 번역하여 사용하였다. 먼저, Covid19_Humor¹ 데이터셋은 코로나19와 관련한 4K 이상의 레딧 게시물, 밈(meme), 밋글을 수집하여 유머의 유형, 주제, 대상 등을 주석한 데이터이다. 유머의 유형을 13가지(teasing, insulting, satire, sexist, racist, age-ism, self-defeating, dark humor, pun, sarcasm, irony, exaggeration, relief, superiority)로 주석하고 있어 세분된 유머 유형과 실현 텍스트를 파악할 수 있다. Joke-dataset²은 약 208,000개의 농담을 크롤링한 것으로 별도의 주석 체계는 갖추지 않은 텍스트 데이터이다. 2017년부터 2022년을 기간으로 각각 레딧, stupidstuff.org, wocka.com에서 데이터를 수집하였다. Short-jokes-dataset³은 레딧 웹사이트에서 크롤링한 231,657개의 짧은 길이의 농담으로 라벨 주석이 이루어지지 않은 텍스트 데이터(raw data)이다.

사용한 유머 데이터는 모두 유머 게시판을 이용하여 데이터를 수집하고 있다. 유머의 모호한 정의와 주관성 개입은 데이터 수집 시 게시판의 성격과 카테고리에 의존한다는 특성을 확인하였다. 이러한 특성때문에 데이터 내의 실현된 유머의 존재 여부는 확인하기 어렵고, 사회 문화적 배경과 맥락에 따라 유머에 대한 해석도 달라진다. 본 연구에서는 3가지 데이터의 주석 여부와 주석 체계가 다르기 때문에 유머 텍스트만 활용하여 기계번역 후 유머 텍스트에 대한 자동·수동 평가를 진행한다. 기계번역은 GPT-4 모델을 사용하였다.

한편, 기계 번역 과정에서 영어-한국어의 ChatGPT 토큰 산정 방식을 고려하여 영어 문장을 토큰나이징하여 한국어 출력 토큰 수를 조정하고, temperature, stop과 같은 ChatGPT 변수도 통제하였으나 문장 삭제, 특정 어휘 생략, 영어 문장 출력 등 다양한 번역 오류가 나타났다. 또한 한국에서는 알려지지 않았지만 영어권 사람들이 가장 싫어하는 ‘라임병’, 제도적으로 금지된 ‘마리화나(대마초)’, 인물(트럼프, 오바마 등)을 활용한 정치 상황, ‘척 노리스’ 같은 미국에서 유행한 밈(meme)

등은 어색하게 한국어로 번역되거나, 검색을 이용하지 않으면 미국의 사회 문화적인 배경을 이해할 수 없어 평가 항목에서 제외하였다. 따라서 평가자가 유머를 이해하기에 명확한 번역 문장만을 대상으로 평가를 진행하였다. 이를 통해 기계 번역 결과에 대한 평가자들의 해석 여지 및 주관 개입 가능성을 낮추어 평가 오류를 줄이고자 하였다.

3.2. 자동 및 수동 평가

선행 연구에 따르면 유머는 개인의 유머 감각에 따라 주관성이 크게 작용한다. 본 연구에서는 ChatGPT를 활용한 유머 자동 평가와 인간 평가자의 수동 평가를 진행하여 유머의 주관성을 보완하고자 하였다. 기구축된 데이터의 특성상 유머 게시판의 글을 정제 없이 수집하여 유머로 분류하고 있다. 따라서 ChatGPT가 평가한 유머의 관점과 인간 평가자의 수동 평가 결과를 비교하여 효율적인 유머 분류 및 인식 작업의 자동화 가능성을 모색하고자 하였다.

선행 연구에 따라 유머 평가 요소를 흥미성, 위험성으로 선정하였다. 흥미성은 해당 문장이 유머러스한 문장 인지를 확인하며 흥미성의 세부 정의는 웃음 유발 요소, 의외성, 비판/비꼬기 어조를 포함한 문장이다. 위험성은 유머 문장 안의 비도덕성/비윤리성을 의미하고, 공격성, 선정성, 혐오 표현으로 세분화하여 유머 문장을 평가한다. 흥미성과 위험성은 이진 분류로 평가되고, 자동 평가의 실효성을 파악하기 위해 2인의 인간 평가자도 평가 체계에 맞추어 데이터를 수동 평가한다.

3.3. 평가 결과

자동·수동 평가는 전체 데이터에서 랜덤 샘플링하여 총 1,000개의 데이터를 사용하였다. 자동 평가는 ChatGPT의 프롬프팅을 활용하였고, 수동 평가는 2인의 인간 평가자가 평가 체계에 맞추어 진행하였다. 자동 및 수동 평가의 결과는 아래 <표 2>과 같다.

<표 2>: 평가 결과 평균

평가 지표	ChatGPT	평가자1	평가자2
흥미성	69.07%	48.65%	47.95%
위험성	16.76%	18.07%	15.89%

<표 2>에 정리된 평가 결과를 통해 유머 데이터에서 유머가 포함된 문장의 비율과 비윤리적인 유머 데이터를 알 수 있다. ChatGPT는 샘플링된 데이터에서 약 69%만이 유머라고 인식하고 있고, 평가자들은 각각 48%, 47%만의

¹ https://github.com/smritae01/Covid19_Humor

³ <https://github.com/amoudgl/short-jokes-dataset>

² <https://github.com/taivop/joke-dataset>

문장만이 유머로 판단하고 있다. 이는 기구축된 유머 데이터에서 절반 정도의 문장이 유머 유무를 판단하기 어렵다는 것을 의미한다. 즉, 유머 여부에 대한 판단이 쉽지 않은 만큼 유머 데이터의 범위를 설정하기도 어렵다는 것을 밝혔다.

또한 ChatGPT가 사람보다 유머의 판단 기준이 낮다는 것을 알 수 있다.

<표 3>: 흥미성 평가 예시

No	Text	Chat GPT	연구자1	연구자2
1	베이커의 손이 왜 갈색이었나요? 그는 똥을 반죽해야 했기 때문입니다.	1	1	0
2	나는 이 전염병을 미리 예측했을 때 내 안과 의사를 제외한 모든 사람들이 나를 미친 사람이라고 말했습니다	1	1	0
3	그는 진정한 신사였고 항상 나를 위해 냉장고 문을 열어주었습니다.	1	0	1
4	철도 도둑이 왜 잡혔을까요? 그는 자신의 흔적을 숨기는 것을 잊어버렸습니다.	1	0	1

<표 3>에 제시된 예시를 살펴보면, 유머 인식 범주가 낮은 ChatGPT는 사람의 주관에 포괄하여 유머로 인식한다. 이는 ChatGPT를 활용하여 자동으로 유머를 평가할 때, 유머 평가 기준을 보다 세분화할 필요성을 시사한다. 그 결과 미세 조정 데이터를 만들 때 자동·수동 평가 결과를 모두 고려하였다. 즉, 각 문장에 대해 2개의 평가가 유머로 인식하면 유머스러운 문장으로 분류하여 미세 조정 데이터를 만들었다.

위험성의 경우 ChatGPT 16%, 인간 평가자 18%, 15%로 ChatGPT와 인간 평가자의 인식 정도가 비슷함을 확인할 수 있다. 3개의 평가 결과가 모두 위험으로 판단된 문장은 한국 법률 위반과 관련하여 살펴보았다. 이는 기존 선행 연구와 다르게 유머의 비윤리성, 도덕성에 대한 논의를 법률에 근거하여 AI 윤리를 고려한 것이다. 유머 데이터를 활용하여 학습 및 생성하는 과정에 있어 법에 저촉할 수 있는 데이터를 필터링하는 데에 활용 가능하며, 비윤리적인 유머 데이터에 대한 기준으로 사용할 수 있다.

<표 4>: 범죄로 분류할 수 있는 텍스트(위험성:1)

No	Text	Chat GPT	연구자1	연구자2
1	니 엄마 너무 못생겼어 세상이 그녀가 마스크를 쓰도록 팬데믹을 만든 걸거야.	1	1	1
2	나는 한 그룹의 사람들에게 코	1	1	1

	로나바이러스에 관한 농담을 했다 결국 모두 이해했지만 중국인 남자는 바로 이해했다			
3	왜 내 중국인 여자친구가 너무 까칠한 걸까요? 그건 사실이죠	1	1	1
4	비행기를 조종할 수 있는 흑인을 무엇이라고 부르나요? 비행사라고 부르지 인종차별하는 놈들아 참고로 이건 안티조크에서는 그다지 반응이 좋지 않았어요 근데 여기서는 반응이 있네.	1	1	1
5	내 전 여자친구는 창녀였어 그녀는 피스타치오 중독자보다 더 많은 일을 벌였어 이별을 앞둔 상황이라서 이걸 보고 있다면 이건 너에 대한 얘기가 아니야 베이비 나를 떠나지 말아줘	1	1	1

<표 4>의 예시에서 법적 근거에 의해 분류할 수 있는 침해 및 피해 사실은 외모 비하, 인종차별, 혐오, 허위 사실 유포가 있다. 이는 모욕, 명예훼손으로 구분하여 구체적으로는 형법 제311조, 형법 제309조, 헌법 제10조에 위반될 가능성이 있다.

헌법 제21조의 표현의 자유는 ‘언론 출판의 자유’와 ‘집회 결사의 자유’를 총칭하는 개념으로 그 구체적인 내용에는 차이가 있으나 대체로 언론 출판의 자유는 언론기관을 통한 표현의 자유, 집회 결사의 자유는 다수의 집합적 의사에 의한 표현의 자유로 파악하고 있다[9]. 표현의 자유는 다른 모든 기본권과 마찬가지로 제한 없는 자유가 아니기 때문에 표현의 자유가 보장된다고 해서 그로 인해 공동체의 존립 자체가 파괴되거나 공동체에 소속되어 있는 다른 구성원들의 인간성과 인격이 파괴되는 것은 허용되지 않는다.

인터넷상에 게재된 텍스트라면 정보통신망법 제70조에도 해당될 가능성이 있으나 정보통신망법 제70조는 ‘사이버 명예훼손’만을 규정하고 있다. 인터넷상에서의 모욕은 현행 형법상 모욕죄로 얼마든지 처벌할 수 있어서 따로 조항을 두지 않았다. 모욕죄는 명예훼손죄와 마찬가지로 그 보호법익과 행위 객체가 ‘명예’이다. 명예라는 것이 인격적 가치에 대한 사회적 평가라는 점을 감안할 때, 인격을 인정하기 어려운 법인이나 법인격 없는 단체에 대한 명예훼손이나 모욕을 인정하는 것은 부당한 확장해석이라는 견해도 있다는 점을 고려하여 이를 제한적으로 해석하는 것이 타당할 것이다. 즉, 형식적으로 법인이나 법인격 없는 단체를 지칭하여 모욕하거나 명예를 훼손시켰으나 실질적으로는 그 단체에 속한 구성원을 지칭한 것이라고 볼 수 있을 때는 그 단체의 구성원에 대한 자연인의 명예를 훼손하거나 모욕한 것으로 보아야 할 것이다[10].

자동·수동 평가 결과, 위법 가능성이 있는 데이터들은 상대적으로 위험성이 높아 모든 평가 주체가 위험성이 있다고 판단하였다. 비슷한 평가 결과를 나타낸 위험

성의 경우 ChatGPT의 자동 평가 결과가 인간의 판단과 유사함을 의미한다. 이러한 이유로 위험성의 평가 체계를 법률에 근거하여 그 정의와 실현 조건을 정교하게 검토하면 자동 평가 결과를 사용할 수 있을 것으로 판단된다. 이에 위 내용을 고려하여 실제 한국법에 있으며, 위험성 1인 데이터를 유머 데이터 상에서 삭제하여 미세 조정 데이터를 만들었다.

4. 실험 및 결과

보다 유머러스한 대화를 나눌 수 있는 모델 개발의 가능성을 확인하고자 자동 및 수동 평가 결과를 적용하여 ChatGPT의 미세 조정을 진행한다.

미세 조정(fine-tuning, 이하 FT) 데이터셋은 흥미성과 위험성을 다음과 같은 경우의 수로 정리하였다.

<표 5>: FT 데이터셋 선별 기준

구분	조건
유머인 경우	{'흥미': '1'}, {'위험성': '0'}
유머가 아닌 경우	{'흥미': '0'}, {'위험성': '0'}
	{'흥미': '0'}, {'위험성': '1'}

흥미성과 위험성은 모두 3개의 평가 결과의 평균을 반올림하여 사용하였다. 즉, <표 3>에 나타난 결과처럼 3개의 주석 결과 중 '1'의 개수가 과반 이상일 경우 사용하였다. 이때 위험성의 경우 3개의 평가 결과가 모두 '1'이고, 범죄 가능성이 있는 경우는 삭제하여 FT 데이터를 구축하였다. <표 5>의 조건에서 유머인 경우는 ChatGPT가 학습할 때 생성해야 하는 데이터로 학습하고, 유머가 아닌 경우는 생성되면 안 되는 것으로 구분한다.

이렇게 만든 FT 데이터는 총 54개로 유머인 경우 26개, 유머가 아닌 경우 28개로 나누어 gpt-turbo-3.5 모델을 사용하여 학습하였다. 미세 조정 이전과 이후 결과를 비교하기 위해, 유머인 문장을 정제하여 <표 6>과 같이 Input-Output 구조로 만들고 Input 문장에 대해 이야기를 완성하도록 하는 프롬프트를 사용했다.

<표 6>: inference data 예시

No.	Input	Output
1	외출금지명령으로 인해 외출할 수 없게 되어서 저는 제 형과 이야기를 시작했습니다.	알고보니 형은 착한 사람 같아요.
2	자가 격리 7일째	나는 3주 동안 아무와도 이야기하지 않았다
3	코로나19의 좋은 점이 하나 있다.	바로, 니가 싫어하는 모든 사람에게 "난 너를 피하고 있다"고 말할 수 있다. 왜냐하면 그것이 정말 사실이거

	때문이다.
--	-------

<표 7>: 미세 조정 결과

No.	FT 이전 (gpt-turbo-3.5)	FT 이후 (gpt-turbo-3.5)
1	그런데 어느 날, 형이 코로나19로 인한 외출 금지 때문에 너무 무료해져서 우리 집의 고양이와 대화를 시작했어요. 처음에는 그냥 농담이었지만, 나중에는 우리 고양이가 정말로 대답하는 것 같아서 미쳐버린 것처럼 느껴졌어!	그런데 그는 제 형이 아니라 제 여동생이었습니다.
2	냉장고 안에서 피자 파티를 열기로 했어, 그런데 피자 주문 시간이 지나버렸어!	냉장고 안에는 더 이상 음식이 없다.
3	그것은 마스크를 쓰면서 무심코 비디오 콜을 하다가 얼굴이 아무것도 가려지지 않는다는 것이다!	그런데 이제는 그냥 미소를 짓는 것이 더 쉽다.

ChatGPT 미세 조정 결과 <표 7>처럼 유머 문장에 대한 이해가 높아졌다는 것을 확인할 수 있다. Input 문장에 대한 Output을 기대했을 때, FT 이후 결과가 완벽히 일치하지는 않으나 FT 이전은 엉뚱한 문장을 장황하게 생성하고 있다면, FT 이후 유머 맥락을 가지고 문장을 완성하고 있다.

5. 결론

본 연구에서는 영어로 된 유머 데이터를 활용하여 유머의 주관성 극복을 위해 자동 평가 및 수동 평가를 진행하였다. 이때, '흥미성, 위험성'이라는 평가 요소를 정의하여 유머 데이터의 유머 존재 여부, 도덕성을 판단하였다.

자동 평가에는 LLM 기반인 ChatGPT를 사용하였다. 흥미성 평가 결과 ChatGPT는 인간 평가자보다 유머 인식 범위가 낮았다. 위험성 인식의 경우 자동 및 수동 평가 결과가 유사하게 나타나 ChatGPT와 2인의 인간 평가자 모두가 위험이라고 판단한 문장에 대해 한국 법률에 위배될 수 있는 유형을 분석하여 유머 데이터에서 삭제하였다. 이렇게 흥미성과 위험성을 고려한 최종 유머 데이터셋을 만들어 ChatGPT 미세 조정을 통해 유머러스한 응답을 생성할 수 있는지 실험하였으며, 그 결과 생성 AI가 유머 맥락을 학습한 결과를 확인하였다.

본 연구는 학자마다 상이한 유머 기준을 자동 및 수동

평가를 통해 자동화할 방안을 실험하였고, 유머의 윤리적 기준을 한국 법률에 맞추어 검토하여 실효성 있는 데이터를 만들었다. 향후에는 한국어 유머 데이터의 부재로 영어 데이터를 기계 번역하여 사용한 한계를 보완하기 위해 한국 문화를 반영한 한국어 유머 데이터를 구축하여 생성 AI가 한국어 유머를 이해하고 생성할 수 있는 방안을 연구할 예정이다.

참고문헌

- [1] 윤민철, 유머의 유형 분류에 관한 고찰-유머의 원리와 구조를 중심으로, 우리말연구, 27, 266-299, 2010.
- [2] 임지원, 유머 담화의 생성 기제와 조건, 우리어문연구, 28, 129-157, 2007.
- [3] 이해완, "농담, 유머, 웃음: '유머의 윤리' 를 중심으로", 美學(미학), 77, 105-140, 2014.
- [4] Ahuja, Vikram, Taradheesh Bali, and Navjyoti Singh. "What makes us laugh? Investigations into automatic humor classification." Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. 2018.
- [5] Mao, Jihang, and Wanli Liu. "A BERT-based Approach for Automatic Humor Detection and Scoring." IberLEF@ SEPLN. 2019.
- [6] Zhang, Dongyu, et al. "Telling the whole story: A manually annotated chinese dataset for the analysis of humor in jokes." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.
- [7] Li, Junze, et al. "This joke is [MASK]: Recognizing Humor and Offense with Prompting." Transfer Learning for Natural Language Processing Workshop. PMLR, 2023.
- [8] Bogireddy, Neha Reddy, Smriti Suresh, and Sunny Rai. "I'm out of breath from laughing! I think? A dataset of COVID-19 Humor and its toxic variants." Companion Proceedings of the ACM Web Conference 2023. 2023.
- [9] 지성우. "허위조작정보 (소위 '가짜뉴스') 규제에 대한 헌법적 문제점에 관한 연구." 공법연구 48.2, 159, 2019.
- [10] 이경재, "모욕죄의 쟁점사항과 관련 판례 고찰-대법원 2015. 12. 24. 선고 2015 도 6622 판결.", 법조 65.8, 631-646, 2016.