

한국어 반어법 탐지 프로젝트 보고서

KoCoSa 데이터셋을 기반으로 한국어 반어법 특징 분석 및
딥러닝 기반 탐지 시스템 구축

작성자: 오경준, 정연후, 김기표

소속: 성균관대학교 인공지능융합전공

제출일: 2025-12-18

Contents

| | |
|---|----|
| 1 배경 및 필요성 | 1 |
| 2 프로젝트 구현 및 코드 공개 | 2 |
| 3 언어 데이터 분석 | 2 |
| 3.1 데이터셋: KoCoSa (Korean Context-aware Sarcasm Detection Dataset) | 2 |
| 3.2 형태론적 분석 | 3 |
| 3.3 구문론적 분석 | 4 |
| 3.4 의미론적 분석 | 6 |
| 3.5 화용론적 분석 | 8 |
| 4 BERT 딥러닝 모델 기반 반어법 탐지 | 9 |
| 4.1 BERT 모델 개요 | 9 |
| 4.2 문맥 기반 반어법 탐지를 위한 모델 설계 | 10 |
| 4.3 데이터 전처리 | 10 |
| 4.4 실험 모델 설계 | 10 |
| 5 GPT 딥러닝 모델 기반 반어법 탐지 | 11 |
| 5.1 GPT 모델 개요 | 11 |
| 5.2 문맥 기반 반어법 탐지를 위한 모델 설계 | 11 |
| 5.3 데이터 전처리 | 12 |
| 5.4 실험 모델 설계 | 12 |
| 5.4.1 임베딩 기반의 분류 방식 | 12 |
| 5.4.2 프롬프팅을 통한 생성 기반 추론 방식 | 12 |
| 6 실험 | 14 |
| 6.1 실험 결과 | 14 |
| 6.2 반어법 탐지 시연 | 15 |
| 6.2.1 데이터 크롤링 | 15 |
| 6.2.2 데이터 정보 | 15 |
| 6.2.3 실험 결과 | 16 |
| 7 기대효과 및 한계점 | 17 |
| 7.1 기대효과 | 17 |
| 7.2 한계점 | 17 |
| 8 결론 | 18 |
| References | 19 |

1. 배경 및 필요성

소셜 네트워킹 서비스(SNS)와 온라인 커뮤니티의 급격한 확산으로 텍스트 기반 의사소통이 일상화되면서, 혐오 표현 탐지, 감정 분석, 콘텐츠 필터링과 같은 자동화된 자연어 처리(NLP) 시스템의 중요성은 지속적으로 증가하고 있다. 그러나 이러한 시스템들이 공통적으로 직면하는 핵심적인 한계 중 하나는 반어법(Sarcasm)이다. 반어법은 겉으로 드러난 문장의 의미와 실제 화자의 의도가 정반대인 언어적 표현으로, 단어 자체 및 언어적 표현 자체의 의미만으로는 정확한 해석이 어렵다. 특히 텍스트 환경에서는 억양, 표정, 제스처와 같은 비언어적 단서가 존재하지 않기 때문에, 인간 독자에게 조차 반어적 의도를 파악하는 일이 쉽지 않은 경우도 존재할 수 있다. 이러한 특성으로 인해 기존의 감정 분석이나 혐오 표현 탐지 시스템은 반어적 텍스트 표현을 긍정 표현으로 오인하는 경우가 빈번하게 발생한다.

예를 들어, 명백한 실수 상황이 이루어진 상황에서 사용되는 “정말 잘~ 한다”, “대단하세요”와 같은 표현은 표면적으로는 긍정적인 어휘를 사용하지만, 실제로는 비판적이고 부정적인 의미를 내포한다. 만약 이러한 문장이 반어법으로 인식되지 못할 경우, 혐오 표현이나 조롱성 발언이 정상적인 긍정 발화로 분류되어 온라인상에서 그대로 방치될 위험이 존재한다. 이는 온라인 안전성 저하뿐만 아니라, 사용자 간의 의미적 오해와 갈등을 증폭시키는 요인으로 작용할 수 있다. 특히 한국어는 존댓말과 반말의 복잡한 체계, 다양한 종결 어미와 조사와 같이 문법적인 표현이 복잡하고 대화 맥락에 대한 높은 의존성을 가지는 언어로, 반어법 탐지가 더욱 어려운 특성을 지닌다. 동일한 문장이라 하더라도 발화 맥락이나 이전 발언에 따라 칭찬과 비꼼이 완전히 상반된 의미로 해석될 수 있다. 이러한 언어적 특성은 단일 문장만을 입력으로 사용하는 기존 텍스트 분류 모델의 한계를 더욱 두드러지게 만든다.

이와 같은 문제의식을 바탕으로, 본 프로젝트는 온라인상에서의 사용자 안전성 확보와 의미적 혼란 방지 및 발전된 혐오 표현 탐지의 관점에서 한국어 반어법 탐지 시스템의 필요성을 제기한다. 특히 단순히 의미적 차원에서 문장 단위의 감정 분류를 넘어, 발화가 이루어진 문맥(Context)을 함께 고려하는 반어법 탐지 모델의 중요성에 주목하였다.

이에 본 프로젝트에서는 문맥 정보가 포함된 KoCoSa (Korean Context-Aware Sarcasm) 데이터셋을 훈련 및 실험 샘플로 활용하여 한국어 반어법을 효과적으로 탐지하는 분석 기법 및 딥러닝 알고리즘 개발을 목표로 한다. 데이터에 내재된 언어학적 특성(형태소, 구문, 의미, 화용적 특징) 분석 및 문맥 정보를 효과적으로 반영하여 반어적 표현을 감지할 수 있도록 하는 BERT 및 GPT 기반의 딥러닝 탐지 시스템을 구축한다. 이를 통해 본 프로젝트에서는 피상적인 의미 차원에서 더 나아가, 대화 맥락 및 상황과 언어적 표현 간의 복합적인 관계를 이해하는 고도화된 한국어 반어법 감지 분석 기법 및 NLP 시스템 구축을 목표로 한다.

2. 프로젝트 구현 및 코드 공개

본 연구에서 수행한 모든 실험 및 분석 코드는 GitHub를 통해 공개되어 있다.

링크: <https://github.com/kyungjunoh/sarcasm/tree/main>

3. 언어 데이터 분석

본 프로젝트에서는 딥러닝 기반 반어법 탐지 모델링에 앞서, 반어법이 가지는 언어학적인 특성을 다각적으로 분석하기 위해 KoCoSa 데이터셋에 대해 심층적인 데이터 분석 및 시각화를 수행하였다. 형태론적, 통사론적, 의미론적 그리고 화용론적인 관점에서 각각 초점을 두고 자동적으로 분석, 정량화 및 시각화 파이프라인을 진행하는 모듈들인 MorphologyAnalyzer, SyntaxAnalyzer, SemanticAnalyzer, PragmaticAnalyzer를 설계하였다.

3.1 데이터셋: KoCoSa (Korean Context-aware Sarcasm Detection Dataset)

본 프로젝트에서는 한국어 반어법 탐지를 위해 *KoCoSa (Korean Context-aware Sarcasm Detection)* 데이터셋 [1]을 사용하였다. KoCoSa는 일상적인 한국어 대화 맥락에서 반어법을 탐지하기 위해 구축된 문맥 기반 대화 데이터셋으로, 총 12,824개의 2인 대화(dialogue)로 구성되어 있으며 각 대화의 마지막 발화(response)에 대해 반어법 여부가 주석되어 있다. 기존 한국어 반어법 데이터셋이 단일 문장 수준이거나 문맥 정보를 포함하지 않는 한계를 가진 것과 달리, KoCoSa는 이전 발화들로 이루어진 문맥(context)을 함께 제공함으로써 반어법 해석에 필수적인 대화 흐름과 상황 정보를 명시적으로 반영한다. 데이터셋은 실제 한국어 메신저 대화 코퍼스를 기반으로 발화 상황과 화자의 공손도를 추출한 뒤, 대규모 언어 모델을 활용해 반어적 응답을 생성하고, 자동 필터링과 전문 한국어 화자에 의한 다단계 인간 검수를 거쳐 구축되었다. 이러한 특성으로 인해 KoCoSa는 반어적 언어 표현 분석 및 문맥 기반 한국어 반어법 탐지 모델을 학습 및 평가를 하기에 적합한 고품질 데이터셋으로 활용될 수 있다.

3.2 형태론적 분석

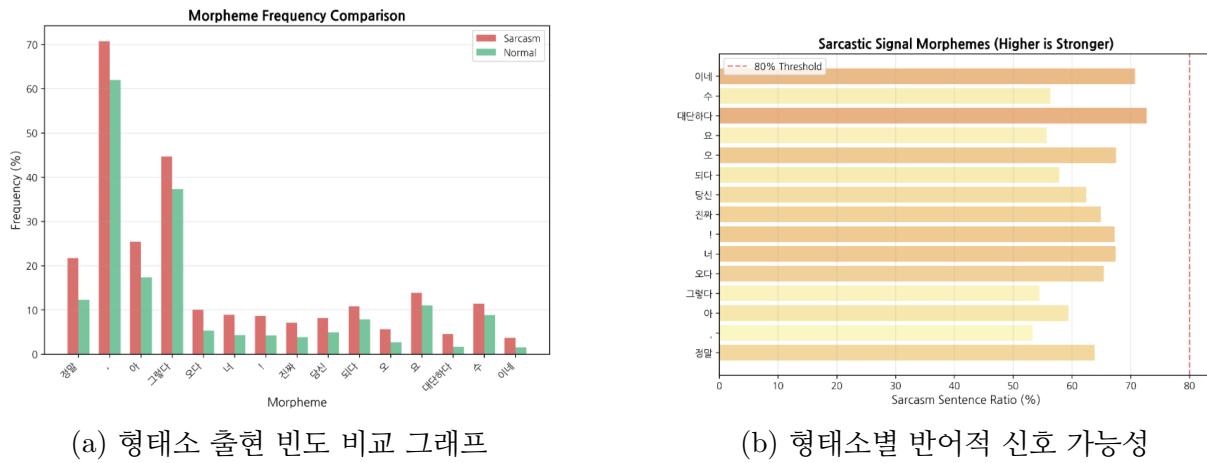


Figure 1: 형태소 기반 반어법 신호 분석 결과

본 모듈은 문장을 형태소 단위로 분해하여, 반어법 문장과 일반 문장에서의 어휘적 사용 양상의 차이를 분석하는 것을 목적으로 한다. 특히 특정 형태소가 반어법적 맥락에서 의도적으로 과도하거나 반복적으로 사용되는 경향이 있다는 점에 착안하여, 이러한 형태소들이 반어법을 식별하는 데 유의미한 단서가 될 수 있는지를 정량적으로 평가한다.

이를 위해 MorphologyAnalyzer는 각 문장에 포함된 형태소들의 출현 빈도를 반어법 문장과 일반 문장으로 구분하여 집계하고, 전체 문장 수를 기준으로 한 상대적 출현 비율을 비교한다. 단순한 빈도 차이를 넘어, 특정 형태소가 등장했을 때 해당 문장이 반어법일 확률을 나타내는 반어법 등장 비율(Sarcasm Sentence Ratio)을 계산함으로써, 개별 형태소의 반어법 신호 강도를 측정한다. 이 지표는 특정 어휘가 반어법적 표현에서 얼마나 일관되게 사용되는지를 보여주는 중요한 근거로 작용한다.

Figure 2(a)는 반어법 문장과 일반 문장의 형태소 출현 빈도를 비교하여, 두 집단 간 언어적 사용 경향의 차이를 분석한 결과를 제시한다. 분석 결과, 반어법 문장에서는 일반 문장에 비해 문장 말미에서 화자의 태도와 평가를 드러내는 종결 표현 및 담화 표지의 사용 빈도가 전반적으로 높게 나타났다. 특히 감탄이나 강조를 나타내는 느낌표(!)와 ‘오’와 같은 형태소가 반어법 문장에서 상대적으로 빈번하게 관찰되었다.

또한 반어법 문장에서는 일반 문장에 비해 ‘정말’, ‘대단하다’, ‘잘’과 같은 강한 긍정 평가 어휘의 사용 빈도가 높게 나타났다. 이는 반어법이 부정적 의도를 직접적으로 표현하기보다는, 과도한 긍정 표현을 통해 의미를 전도시키는 언어적 전략을 활용하는 경향이 있음을 시사한다. 이러한 긍정 평가 어휘는 단독으로 사용되기보다는 종결 표현이나 문장부호와 결합되어 반어적 의미를 강화하는 양상을 보였다.

특히 ‘요’와 같은 공손 종결 보조 표현 역시 반어법 문장에서 더 높은 빈도를 보였으며, 이는 공손한 형식을 유지한 채 비판적 태도를 우회적으로 드러내는 반어적 표현 방식과 밀접한 관련이 있음을 보여준다. 예를 들어 ‘대단하시네요’와 같은 표현은 긍정적 어휘와 공손한 종결 형식을 취하지만, 맥락에 따라 비판이나 조롱의 의미로 해석될 수 있다.

Figure 2(a)의 분석은 반어법 문장이 일반적인 문장과 비교했을 때 전반적으로 어떠한 형태소들을 더 자주 사용하는지를 보여주지만, 특정 형태소가 등장했을 때 그것이 실제로 반어적 해석을 얼마나 강하게 유도하는지는 직접적으로 설명하지는 못한다. 이를 보완하기 위해 Figure 2(b)에서는 특정 형태소가 포함된 문장이 반어법으로 분류될 확률, 즉 반어법 등장 비율(Sarcasm Sentence Ratio)을 계산하여 형태소별 반어적 신호 가능성을 정량적으로 분석하였다.

Figure 2(b)의 분석 결과, ‘이네’, ‘대단하다’, ‘정말’, ‘진짜’, ‘느낌표’, 그리고 ‘요’, ‘오’, ‘아’와 같은 종결 표현 및 감탄사가 높은 반어법 등장 비율을 보이며, 개별 형태소 수준에서도 강력한 반어적 신호로 작용함이 확인되었다. 특히 ‘대단하다’, ‘정말’과 같은 긍정 평가 어휘는 단순히 자주 등장할 뿐만 아니라, 해당 형태소가 포함된 문장이 반어법으로 해석될 가능성 또한 높다는 점에서 반어적 신호로서의 중요성이 두드러졌다.

Figure 2(a)과 Figure 2(b)의 분석을 통해, 반어법 문장은 일반 문장과 비교하여 특정 형태소적 표현을 더 자주 사용하며, 동시에 이러한 형태소들이 포함된 문장은 반어법으로 해석될 확률 또한 높다는 점을 확인하였다. 이는 개별 형태소가 반어적 표현의 신호로 기능할 수 있음을 시사한다.

3.3 구문론적 분석

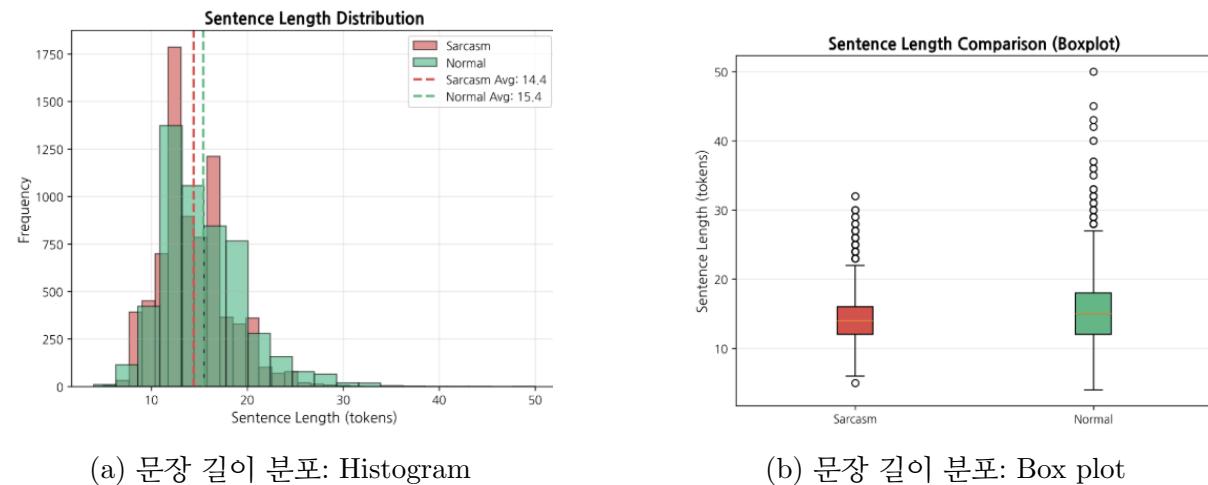


Figure 2: 구문적 관점에서의 분석 결과

SyntacticAnalyzer 모듈은 문장의 의미적 내용뿐만 아니라, 문장이 어떻게 구성되어 있는지에 대한 구조적 특성에 주목하여 반어법 문장의 통사적 특징을 분석하는 역할을 수행한다. 특히 품사(POS) 간의 연결 패턴과 문장 길이와 같은 구문적 요소를 중심으로, 반어법 문장과 일반 문장 간의 구조적 차이를 체계적으로 파악하는 것을 목표로 한다.

이를 위해 해당 모듈은 문장을 구성하는 품사 태그 시퀀스를 분석하여, 연속적으로 등장하는 품사 조합(예: 명사-조사, 부사-형용사 등)의 출현 양상을 수집하고, 문장 내 토큰 수를 기준으로 문장 길이 분포를 함께 분석한다. 이러한 분석을 통해 특정 품사 연결 패턴이 반어법 문장에서 상대적으로 더 자주 나타나는지, 혹은 반어법 문장이 전반적으로 더

짧거나 긴 구조를 가지는지와 같은 통사적 가설을 데이터 기반으로 검증한다.

분석 결과는 시각화 과정을 통해 직관적으로 제시된다. 먼저 반어법 문장과 일반 문장 간에 변별력이 높은 주요 품사 패턴을 비교하여, 어떤 구문 구조가 반어적 표현과 밀접한 관련을 가지는지를 확인한다. 또한 문장 길이 분포를 히스토그램과 박스플롯 형태로 제시함으로써, 두 집단 간 문장 길이의 전반적인 분포 차이와 통계적 특성을 함께 비교 및 분석한다. 더 나아가, 품사 패턴별 빈도 차이를 막대그래프로 표현하여 특정 구문 구조의 반어법 관련 강도를 정량적으로 파악할 수 있도록 한다.

종합적으로 본 구문 분석 모듈은 “반어법 문장은 상대적으로 짧은 구조를 가진다” 또는 “특정 품사 조합이 반어법 문장에서 반복적으로 등장한다”와 같은 문장 구조적 가설을 실증적으로 검증하고, 그 결과를 시각적으로 제시하는 역할을 수행한다. 이를 통해 반어법이 어휘적 특징뿐만 아니라 문장 구조 차원에서도 일정한 패턴을 가진다는 점을 분석적으로 뒷받침한다.

Figure 2(a)과 Figure 2(b)는 반어법 문장과 일반 문장의 문장 길이를 토대로 기준으로 비교한 결과를 각각 분포와 요약 통계 관점에서 제시한다. Figure 2(a)은 두 집단의 문장 길이 분포를 히스토그램과 평균값으로 나타내며, Figure 2(b)는 박스플롯을 통해 중앙값, 사분위 범위, 이상치를 시각화한다.

분석 결과, 반어법 문장은 일반 문장에 비해 전반적으로 더 짧은 길이에 집중된 분포를 보였다. 평균 문장 길이 역시 반어법 문장이 일반 문장보다 짧게 나타났으며, 이는 히스토그램 상에서 짧은 토큰 구간에 빈도가 집중되는 양상으로 확인된다. 또한 박스플롯 분석에서도 반어법 문장은 중앙값과 사분위 범위가 더 낮고 좁게 분포하여, 문장 길이의 변동성이 상대적으로 작음을 보여준다. 반면 일반 문장은 상위 이상치가 다수 존재하며, 문장 길이 분포의 범위가 더 넓게 나타나 다양한 길이의 문장이 사용되고 있음을 시사한다.

이러한 결과는 반어법이 장황한 서술보다는 짧고 압축적인 문장 구조를 통해 화자의 태도나 평가를 전달하는 경향을 가진다는 점을 통사적 관점에서 뒷받침한다. 즉, 문장 길이 자체가 반어법을 구분하는 하나의 구조적 단서로 작용할 수도 있음을 보여준다.

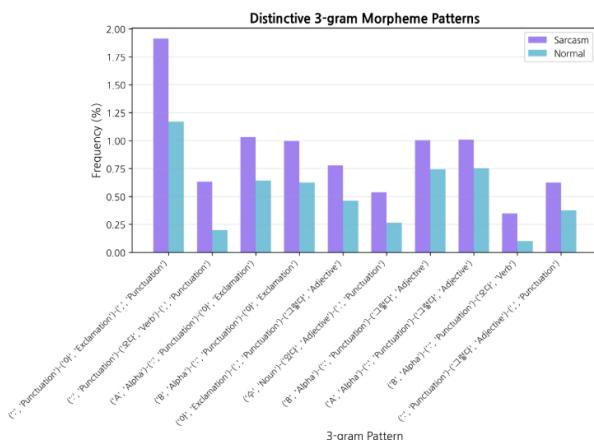


Figure 3: 형태소 3-그램 패턴의 상대적 빈도

Figure 3은 반어법 문장과 일반 문장에서 나타나는 형태소 3-그램 패턴의 상대적 빈도를 비교한 결과를 보여준다. 본 분석은 단일 형태소나 이웃한 두 요소를 넘어, 여러 형태소가

연속적으로 결합될 때 반어적 표현이 어떻게 구조화되는지를 파악하는 데 목적이 있다.

분석 결과, 반어법 문장에서는 일반 문장에 비해 감탄사, 구두점, 평가적 어휘가 연쇄적으로 결합된 형태소 패턴이 상대적으로 높은 빈도로 나타났다. 특히 감탄사와 구두점, 형용사 또는 평가적 표현이 함께 등장하는 3-그램 패턴은 반어법 문장에서 두드러지게 관찰되었으며, 이는 반어적 어조가 형태소 수준의 연속적 결합을 통해 강화됨을 시사한다.

이러한 경향은 반어법이 개별 형태소 하나에 의해 즉각적으로 결정되기보다는, 여러 형태소가 누적적으로 결합되며 화자의 태도와 감정을 점진적으로 드러내는 구조적 특성을 가짐을 보여준다. 즉, 반어적 의미는 단어 선택뿐만 아니라 형태소 배열과 순서에 의해 형성되며, 이는 n-gram 기반 형태소 분석이 반어법 탐지에서 유의미한 정보를 제공할 수 있음을 뒷받침한다.

3.4 의미론적 분석

SemanticAnalyzer 모듈은 문맥(context)과 그에 대한 응답(response) 사이의 의미론적 불일치에 주목하여 반어법적 표현의 특성을 분석하는 역할을 수행한다. 반어법은 종종 발화 자체의 의미보다, 발화가 놓인 상황이나 이전 맥락과의 대비를 통해 형성되므로, 문맥과 응답 간의 의미적·감성적 관계를 정량적으로 분석하는 것이 중요하다.

이를 위해 본 모듈은 문맥과 응답의 의미 표현 간 유사도를 계산하여, 두 텍스트가 의미적으로 얼마나 일관되게 연결되어 있는지를 분석한다. 또한 문맥과 응답 각각의 감성 극성을 비교함으로써, 발화가 주변 맥락과 감정적으로 조화를 이루는지, 혹은 의도적인 대비를 형성하는지를 함께 평가한다. 이러한 분석을 통해 문맥과 응답 사이의 의미적 거리와 감성적 대비 정도를 정량적 지표로 도출한다.

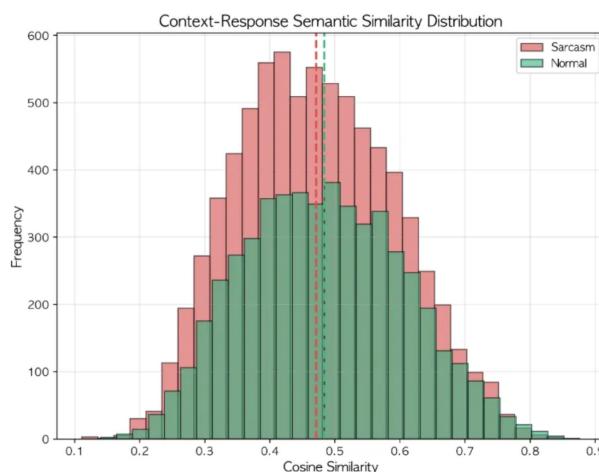


Figure 4: 문맥(context)과 응답(response) 간 의미론적 유사도(cosine similarity) 분포 비교

Figure 4은 문맥(context)과 응답(response) 간의 의미론적 유사도(cosine similarity) 분포를 반어법 문장과 일반 문장으로 구분하여 비교한 결과를 나타낸다. 히스토그램과 평균값 표시를 통해 두 집단의 전반적인 유사도 분포 양상과 중심 경향을 시각적으로 확인할 수 있다.

분석 결과, 일반 문장은 문맥과 응답 간의 의미론적 유사도가 상대적으로 높은 구간에 더 많이 분포하며, 평균 유사도 또한 반어법 문장보다 높게 나타났다. 이는 일반 문장이 문맥의 내용이나 흐름을 의미적으로 일관되게 이어받아 응답하는 경향이 반어법이 사용되었을 때에 비해 비교적으로 강함을 시사한다. 반면 반어법 문장은 평균적으로 더 낮은 의미론적 유사도를 보이며, 분포 역시 낮은 유사도 구간으로 더 넓게 확산된 양상을 나타냄을 알 수 있다.

이러한 차이는 반어법의 핵심적인 특성인 문맥과 응답 간의 의미적 불일치를 반영하는 결과로 해석할 수 있다. 반어법 문장은 표면적으로는 문맥에 대한 응답의 형태를 취하지만, 실제 의미 차원에서는 문맥의 기대와 어긋나는 내용을 제시하거나 대비를 형성하는 경우가 많다. 그 결과, 문맥과 응답 간의 의미 표현이 완전히 정합되지 못하고 상대적으로 낮은 의미론적 유사도로 나타나는 경향을 보인다.

다만 두 분포가 일정 부분 중첩되는 점 또한 관찰되며, 이는 모든 반어법 문장이 극단적으로 낮은 의미론적 유사도를 갖는 것은 아니라는 점을 시사한다. 일부 반어적 표현은 문맥과 의미적으로 유사한 어휘를 유지한 채, 감성이나 맥락·상황적인 측면에서 반어적 의도를 형성하기 때문에 의미론적 유사도만으로 반어법을 완전히 구분하는 데에는 한계가 존재함을 보여준다.

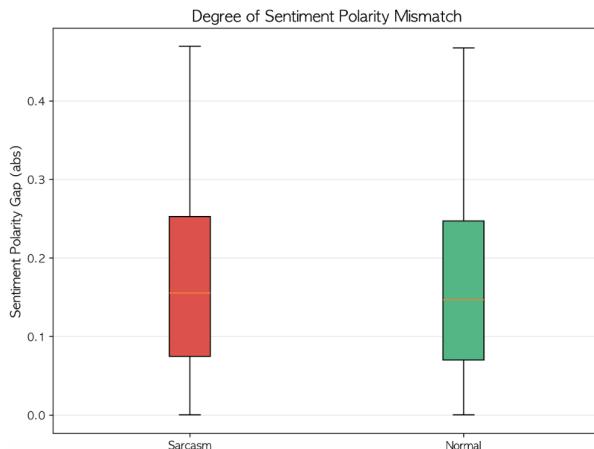


Figure 5: 문맥(context)과 응답(response) 간 감성 극성 차이 절대값 분포(박스플롯)

Figure 5은 문맥(context)과 응답(response) 간의 감성 극성 차이의 절대값을 반어법 문장과 일반 문장으로 구분하여 비교한 결과를 박스플롯 형태로 제시한다. 해당 지표는 문맥과 응답이 감성적으로 얼마나 대비되는지를 정량적으로 나타내며, 값이 클수록 감성적 불일치가 크다는 것을 의미한다.

분석 결과, 반어법 문장은 일반 문장에 비해 감성 극성 차이의 중앙값이 상대적으로 높게 나타나며, 분포 역시 상대적으로 넓은 범위를 보인다. 이는 반어법 문장에서 문맥과 응답 사이에 감성적으로 상반되거나 대비되는 표현이 더 자주 발생함을 시사한다. 즉, 반어법은 긍정적인 문맥에 부정적인 뉘앙스의 응답을 제시하거나, 반대로 부정적인 상황에서 과도한 긍정 표현을 사용하는 방식으로 감성적 대비를 형성하는 경향이 있다.

반면 일반 문장은 감성 극성 차이가 상대적으로 낮은 구간에 더 밀집되어 있으며, 중앙값 또한 반어법 문장보다 낮게 나타난다. 이는 일반 문장이 문맥과 응답 간의 감성적 흐름을

비교적 일관되게 유지하는 경향을 보인다는 점을 의미한다. 다만 일부 이상치가 존재하는 것으로 보아, 모든 일반 문장이 감성적으로 완전히 일관된 구조를 갖는 것은 아님을 확인할 수 있다.

종합하면, Figure 4과 Figure 5의 분석 결과는 반어법이 문맥과 응답 간의 의미적·감성적 불일치를 핵심 특징으로 가진다는 점을 일관되게 보여준다. 의미론적 분석에서의 유사도 분석에서는 반어법 문장이 일반 문장에 비해 문맥과 응답 간 의미적 일관성이 낮은 경향을 보였으며, 감성 극성 분석에서는 반어법 문장에서 감성적 대비가 더 크게 나타나는 양상이 확인되었다. 이는 반어법이 단순히 문맥과 무관한 발화라기보다는, 문맥의 기대를 유지하거나 부분적으로 공유하면서도 의미나 감성 차원에서 의도적인 어긋남을 형성하는 언어적 전략임을 시사한다. 동시에 두 지표 모두에서 분포의 중첩이 관찰된다는 점은, 반어법이 의미론적 측면에서의 단일 기준으로 명확히 구분되기보다는 다양한 단서가 복합적으로 작용하는 현상을 시사한다.

본 절에서는 의미론적 표현과 감성 지표를 중심으로 반어법의 표면적 특성을 분석하였으며, 다음 절에서는 이러한 의미론적 단서들이 실제 담화 맥락에서 어떠한 화용론적 기능을 수행하는지를 보다 심층적으로 논의한다.

3.5 화용론적 분석

본 절에서는 반어법이 지니는 화용론적 특성을 분석하기 위해, 발화의 표면적 의미와 화자가 실제로 의도한 의미 간의 불일치에 주목한다. 반어법은 단순히 문장의 의미가 부정확하거나 모호한 경우가 아니라, 화자가 특정한 담화적 목적을 가지고 의도적으로 문자적 의미와 다른 의미를 전달하는 화용적 현상이라는 점에서, 화자의 의도를 직접적으로 분석하는 접근이 중요하다.

이를 위해 본 연구에서는 데이터셋에 포함된 sarcasm explanation을 활용하였다. 해당 설명은 각 반어 문장에 대해 화자가 실제로 전달하고자 한 의미나 의도를 명시적으로 서술한 텍스트로, 화용 분석을 위한 중요한 근거 자료로 활용될 수 있다. 본 모듈에서는 반어로 라벨링된 문장에 대해, 설명 텍스트에 포함된 핵심 표현을 기반으로 화자의 의도를 자동 추출하였다. 구체적으로, “비꼬다”, “조롱”, “반대 의미”, “위로”, “비난”, “농담”, “아이러니” 등 반어적 의도를 나타내는 대표적인 키워드를 사전 정의하고, 각 설명 문장에 해당 키워드가 포함되어 있는지를 탐색함으로써 발화의 의도 유형을 분류하였다. 만약 특정 의도 키워드가 명시적으로 나타나지 않는 경우에는, 이를 일반적인 반어 표현으로 분류하였다.

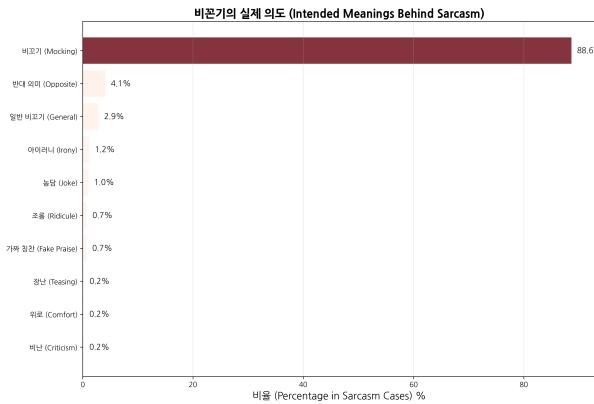


Figure 6: 화자의 실제 의도 분포

Figure 6는 반어 문장에 대해 추출된 화자의 실제 의도 분포를 시각화한 결과를 보여준다. 분석 결과, 전체 반어 사례의 대부분이 ‘비꼬기(Mocking)’ 유형으로 분류되어, 반어법이 주로 상대방이나 상황을 간접적으로 비판하거나 빙정대는 담화 전략으로 사용됨을 확인할 수 있었다. 그 외에도 ‘반대 의미 제시’, ‘아이러니’, ‘농담’, ‘조롱’, ‘가짜 칭찬’ 등의 다양한 의도 유형이 소수 사례로 관찰되었으며, 이는 반어법이 단일한 기능이 아니라 여러 화용적 목적을 수행할 수 있는 다층적인 표현 방식임을 시사한다.

종합하면, 본 화용 분석은 반어법이 단순한 의미 왜곡이 아니라, 문맥과 감성 흐름을 공유하면서도 화자의 의도 차원에서 의미적 전환을 일으키는 담화적 전략임을 보여준다. 특히 데이터셋에 포함된 반어 설명을 활용함으로써, 발화의 표면적 의미와 실제 의도 간의 괴리를 직접적으로 분석할 수 있었으며, 이는 의미론적 분석만으로는 포착하기 어려운 반어법의 핵심적인 화용적 특성을 정량적·정성적으로 드러낸다는 점에서 의의를 가진다.

4. BERT 딥러닝 모델 기반 반어법 탐지

4.1 BERT 모델 개요

BERT (Bidirectional Encoder Representations from Transformers) [2]는 Transformer 구조를 기반으로 한 사전학습 언어 모델로, 문장을 좌에서 우 또는 우에서 좌로만 처리하던 기존 언어 모델과 달리 양방향(Bidirectional) 문맥 정보를 동시에 학습할 수 있다는 특징을 가진다. 이를 통해 문장 내 단어의 의미를 주변 단어들과의 관계 속에서 보다 정밀하게 표현할 수 있다.

BERT는 대규모 말뭉치에 대해 Masked Language Modeling (MLM)과 Next Sentence Prediction (NSP)이라는 두 가지 사전학습 목표를 사용한다. MLM은 입력 문장의 일부 토큰을 마스킹한 뒤 이를 예측하도록 학습함으로써 문맥 기반 표현 학습을 가능하게 하며, NSP는 두 문장 간의 연속성을 예측함으로써 문장 간 관계를 학습하도록 설계되었다. 이러한 학습 방식은 문장 수준의 의미 이해뿐만 아니라, 문맥 간 관계를 고려하는 다양한 자연어 처리 과제에 효과적으로 활용될 수 있다.

특히 BERT의 Self-Attention 메커니즘은 문장 내 모든 토큰 쌍 간의 상호작용을 동시

에 고려할 수 있도록 하여, 단어 간 장거리 의존성(long-range dependency)을 효과적으로 포착한다. 이는 반어법과 같이 특정 단어 자체보다 문장 전체의 맥락과 의미적 대비에 의해 해석되는 언어 현상을 모델링하는 데 적합한 구조적 장점으로 작용한다.

4.2 문맥 기반 반어법 탐지를 위한 모델 설계

이러한 특성을 지닌 BERT를 기반으로, 본 연구에서는 문맥 정보를 적극적으로 활용하여 한국어 반어법 탐지 성능을 향상시키는 딥러닝 모델을 설계하고 그 효과를 검증하였다. 실험의 핵심은 문맥(Context) 정보가 모델의 반어법 판단 능력에 미치는 영향을 비교·분석하는 것이다. 이를 위해 문맥을 반영한 모델과 그렇지 않은 기준 모델(Baseline)을 동일한 환경에서 학습시키고 성능을 평가하였다. 모델 훈련 및 검증을 위해 사용한 데이터셋은 언어적 분석에서 사용한 동일한 KoCoSa 데이터셋이다.

4.3 데이터 전처리

반어법은 곁으로 드러난 의미와 실제 의도가 상반되는 복합적인 언어 현상이므로, 언어적 표현을 최대한으로 보존하면서도 불필요한 노이즈를 제거하는 섬세한 전처리 전략이 필수적이다.

- **정보 보존적 노이즈 제거:** 시스템 로그나 광고성 문자열과 같은 무의미한 노이즈는 제거하되, 반어적 의도(과장, 비웃음)를 담고 있을 수 있는 감탄사(예: “헐”), 이모티콘(), 웃음 표현(ㅋㅋㅋ), 의도적 문자 반복(“잘~~~하네”) 등은 반어적 의미 탐지에 강한 신호를 줄 수 있기에 그 원형을 최대한 보존하였다. 이는 모델이 미묘한 반어적 신호를 학습하는 데 중요한 단서로 작용한다.
- **한국어 특화 토크나이저:** 한국어 구어체, 신조어, 오타 등을 효과적으로 처리하기 위해 우리가 실험할 데이터와 최대한 비슷한, 댓글 데이터에 강점을 지닌 KcBERT 토크나이저를 채택하였다. 이를 통해 문장의 의미 단위를 보다 정확하게 분절하고, 모델이 데이터의 특성을 깊이 있게 학습하도록 유도하였다.

4.4 실험 모델 설계

한국어 발화 데이터 및 SNS 데이터(뉴스 댓글 등)의 특성을 효과적으로 반영하기 위해 이를 기반으로 학습된 KcBERT (Korean Comments BERT) [3]를 기반 모델(Backbone)로 선정하고, 문맥 반영 여부에 따라 두 가지 종류의 입력을 받도록 설계하였다.

Model 1: Target-Only Model 문맥 정보 없이 반응(Response) 발화만을 입력으로 사용하여 반어법 여부를 판단하는 Baseline 모델이다. 입력 형식은 다음과 같이 구성하였다.

$$Input_{Target-Only} = [CLS] \ Response \ [SEP]$$

Model 2: Context-Aware Model 문맥(Context)과 반응(Response)을 [SEP] 토큰으로 연결하여, 두 문장 사이의 관계를 모델이 학습 및 이를 기반으로 반어법을 탐지하도록 설계하였다. 입력 형식은 다음과 같이 구성하였다.

$$Input_{Context-Aware} = [CLS] \text{ } Context \text{ } [SEP] \text{ } Response$$

이 구조는 BERT의 핵심 구성 요소인 Self-Attention 메커니즘이 문맥과 반응 사이의 의미적 상호작용(예: 불일치, 모순, 강조)을 포착할 수 있도록 한다. 예를 들어, “차가 막혔다”와 같은 부정적 문맥 뒤에 “시간 약속을 잘 지킨다”는 긍정적 반응이 이어질 경우, 모델은 이러한 문맥과 답변 사이의 의미적 불일치를 학습함으로써 타겟이되는 답변만을 입력으로 넣었을 때보다 반어적 의도를 효과적으로 감지할 수 있다.

5. GPT 딥러닝 모델 기반 반어법 탐지

5.1 GPT 모델 개요

GPT (Generative Pre-trained Transformer) [4]는 Transformer의 디코더(Decoder) 구조를 기반으로 한 생성 모델이다. 문장 전체의 문맥을 양방향으로 동시에 파악하는 BERT와 달리, GPT는 단방향(Unidirectional) 혹은 자기회귀적(Autoregressive) 속성을 가진다. 즉, 이전 단어들의 시퀀스를 바탕으로 다음에 올 단어를 예측하는 방식으로 학습되며, 이를 통해 문장의 왼쪽에서 오른쪽으로 흐르는 문맥적 인과관계를 강력하게 모델링한다.

GPT는 대규모 말뭉치에 대해 다음 토큰 예측(Next Token Prediction)이라는 사전학습 목표를 수행한다. 이는 특정 시점의 단어를 예측하기 위해 오직 그 이전 시점의 단어들만을 참고하도록 Masked Self-Attention을 적용하는 방식이다. 이러한 구조는 문장의 생성뿐만 아니라, 문맥의 흐름을 파악하고 뒤이어 나올 내용과의 논리적 연결성을 추론하는 데 뛰어난 성능을 보인다.

특히 본 프로젝트와 같은 반어법 탐지 과제에서, GPT의 이러한 특성은 문맥(Context)이 주어졌을 때 그 뒤에 이어지는 반응(Response)이 문맥적 흐름상 자연스러운지, 혹은 의도적으로 비틀어진(반어적) 표현인지를 판단하는 데 유효한 단서를 제공할 수 있다.

5.2 문맥 기반 반어법 탐지를 위한 모델 설계

본 프로젝트에서는 GPT 모델이 사전학습 과정을 통해 습득한 문맥 이해 능력을 반어법 탐지 과제에 최적으로 적용하기 위해, 모델의 활용 방식을 ‘임베딩 기반의 분류’와 ‘생성 기반의 추론’이라는 두 가지 상이한 패러다임으로 설계하여 실험 및 결과를 비교하였다. 자세한 내용은 5.4 실험 모델 설계에서 다룬다

5.3 데이터 전처리

BERT 딥러닝 모델 기반 반어법 탐지 방식에서 사용한 데이터 및 데이터 전처리 방식과 동일하다.

5.4 실험 모델 설계

본 프로젝트에서는 생성 모델인 GPT의 구조적 특성을 활용하여 한국어 반어법을 탐지하기 위해 ‘임베딩 기반의 분류’와 ‘프롬프팅을 통한 생성 기반 추론’라는 두 가지 상이한 접근 방식을 시도하였다. 첫째, 임베딩 기반의 분류 방식은 GPT를 고정된 특징 추출기로 간주하여 분류 헤드를 학습시키는 방식의 접근법이다. 문맥과 반응을 처리한 후 생성되는 마지막 토큰의 잠재 표현이 반어적인 정보를 벡터 공간 상에서 얼마나 명확하게 함축하고 있는지를 검증하기 위해 설계되었다. 둘째, 프롬프팅을 통한 생성 기반 추론 방식은 GPT 고유의 생성 능력과 In-context Learning 능력을 직접 활용하는 접근법이다. 모델에게 반어법의 정의와 예시를 텍스트 형태로 프롬프팅함으로써, 별도의 파라미터 학습 없이 모델이 문맥적 흐름을 읽고 논리적 추론을 통해 반어법 여부를 판단할 수 있는지 평가하기 위한 설계이다. 두 방식 모두 사전학습된 GPT [5]를 백본(Backbone)으로 사용하며, 모델의 언어 이해 능력을 최대화하는 데 중점을 두었다.

5.4.1 임베딩 기반의 분류 방식

첫 번째 방식은 사전학습된 GPT 모델을 특징 추출기로 활용하고, 그 위에 분류기를 부착하여 학습하는 전이 학습 전략이다.

- **구조:** GPT 모델의 모든 파라미터는 학습되지 않도록 고정하여 사전학습된 지식을 보존한다. 입력 시퀀스가 모델을 통과한 후 산출되는 마지막 토큰의 잠재 상태 h_{last} 를 추출한다. 이 벡터는 디코더 구조 특성상 이전까지의 모든 문맥 정보를 함축하고 있다.
- **학습:** 추출된 h_{last} 를 훈련 가능한 선형 계층에 통과시켜 반어법 여부(0 또는 1)를 예측한다.

$$y = \text{Softmax}(W \cdot h_{last} + b)$$

5.4.2 프롬프팅을 통한 생성 기반 추론 방식

두 번째 방식은 모델의 파라미터를 전혀 업데이트하지 않고, GPT 고유의 텍스트 생성 능력을 활용하여 반어법을 판별하는 In-context Learning 방식이다. 모델에게 명시적인 지시문과 예시를 프롬프트 형태로 제공하여, 모델이 스스로 입력 문장의 반어법 여부를 추론하고 정답을 생성하도록 유도한다. 모델에게 제공한 프롬프트 구성은 아래와 같다.

프롬프트 구성 모델이 수행해야 할 작업과 출력 형식을 정의하기 위해 다음과 같은 구조의 템플릿을 설계하였다.

1. **Task Description (지시문):** 모델에게 “반어법 탐지 전문가”라는 페르소나를 부여하고, 반어법(Sarcasm)의 정의(표면적 의미와 의도의 불일치)를 명시한다. 또한, 모델이 정해진 포맷에 맞춰 반어법일 경우 “1”, 아닐 경우 “0”을 출력하도록 제약 조건을 설정한다.
2. **Few-shot Examples (예시):** 모델이 문맥적 추론 패턴을 파악할 수 있도록 문맥 (Context), 반응(Response), 그리고 정답(Detection Result)으로 구성된 예시를 제공한다. 본 프로젝트에서는 입력 토큰 길이의 제한을 고려하여 2개의 예시(2-shot)를 제공하였다.
3. **Target Input (입력):** 실제 판별 대상인 문맥과 반응을 프롬프트의 마지막에 배치하고, “Answer:”라는 텍스트로 끝맺음으로써 모델이 자연스럽게 정답(0 또는 1)을 생성하도록 유도한다.

실제 실험에 사용된 프롬프트 템플릿의 구조는 아래와 같다.

```

Task: You are really good at detecting the sarcastic response at the
last utterance of the given dialog.

Sarcasm is ironic or mocking language where someone says the opposite
of what they mean.

If sarcastic, print "1". If not, print "0".

Example 1:
Context: "..."
Response: "..."
Detection Result: 1

Example 2:
Context: "..."
Response: "..."
Detection Result: 0

Context: "{Target Context}"
Response: "{Target Response}"
Answer:

```

추론 및 결정 추론 과정에서 모델은 마지막 “Answer:” 토큰 이후에 등장할 다음 토큰을 예측하게 된다. 이때 단순히 생성된 텍스트를 파싱하는 것이 아니라, 타겟 토큰인 “0”과 “1”에 해당하는 Logit(생성 확률) 값을 추출하여 비교한다. 즉, $P("1" | Prompt)$ 와 $P("0" | Prompt)$ 중 더 높은 확률을 가진 클래스를 최종 예측값으로 선정한다.

6. 실험

6.1 실험 결과

| Input Features | Model Type | Accuracy (%) | F1-score |
|------------------------|----------------------|--------------|-------------|
| Only Response (Target) | BERT | 74.3 | 79.7 |
| Context + Response | BERT | 75.6 | 79.8 |
| Context + Response | GPT (Linear Probing) | 54.8 | 70.6 |
| Context + Response | GPT (In-Context) | 55.0 | 70.9 |

Table 1: 모델 성능 비교 (BERT vs GPT)

Table 1은 본 연구에서 제안한 모델들의 반어법 탐지 정확도(Accuracy)와 F1-score를 요약하여 보여준다. 실험 결과에 대한 상세 분석은 다음과 같다.

첫째, 문맥 정보의 유효성이 검증되었다. BERT 기반 모델을 기준으로 비교했을 때, 응답만을 단독으로 입력받은 모델(Target-Only)은 74.3%의 정확도를 기록한 반면, 문맥과 응답을 함께 입력받은 모델(Context-Aware)은 75.6%로 약 1.3%p 향상된 성능을 보였다. 이는 앞선 언어 데이터 분석에서 확인한 바와 같이, 반어법이 문맥과 반응 사이의 의미적 불일치나 감성적 대비를 통해 형성된다는 점을 딥러닝 모델이 효과적으로 학습했음을 시사한다. 다만, 성능 향상 폭이 극적이지 않은 이유는 ‘형태소 분석’ 결과에서 나타난 것처럼, 반어적 발화 자체에 ‘ㅋㅋㅋ’, ‘??’, 과장된 존칭 등 문맥 없이도 반어적 신호를 전달하는 텍스트 내적 표지가 강하게 포함되어 있기 때문으로 해석된다.

둘째, BERT과 GPT의 성능 격차가 확인되었다. 실험 결과, 파라미터 전체를 데이터셋에 맞춰 미세조정한 BERT 모델이 GPT를 활용한 두 가지 방식(Linear Probing, In-Context Learning)보다 월등히 높은 성능을 기록했다. GPT 기반 모델들은 정확도 55% 수준에 머물렀는데, 이는 이진 분류의 무작위 추측 수준을 크게 상회하지 못하는 결과이다.

이러한 성능 차이의 원인은 도메인 적합성 때문이라고 생각한다. 본 실험에 사용된 KcBERT는 한국어 댓글 데이터로 사전 학습되어 구어체와 신조어에 강점을 가진 반면, 범용적으로 사전 학습된 GPT 모델은 한국어의 특수한 반어적 어투에 대한 이해도가 상대적으로 낮았을 가능성이 있다.

결론적으로, 제한된 데이터 환경에서 한국어 반어법과 같이 맥락 의존적인 작업을 수행할 때는, 범용적 학습 기반의 거대 모델의 생성 능력보다는 도메인에 특화된 모델을 미세조정하는 방식이 더 효과적임을 확인할 수 있었다.

6.2 반어법 탐지 시연

6.2.1 데이터 크롤링

KoCoSa 데이터셋으로 학습된 모델이 정제된 데이터뿐만 아니라 실제 웹 환경(Real-world Data)의 비정형 텍스트에서도 풍자를 효과적으로 탐지하는지 검증하기 위해, 인터넷상에서 고도의 반어적 표현이 빈번하게 관찰되는 영화 ‘리얼(Real)’의 리뷰 데이터를 수집하여 별도의 테스트셋을 구축하였다.

데이터 선정 배경은 영화 ‘리얼’은 대중적으로 매우 낮은 평가를 받았음에도 불구하고, 온라인 커뮤니티 상에서 “나만 당할 수 없다”는 심리로 10점 만점을 부여하며 극찬을 남기는 반어적 놀이 문화가 형성된 특수한 사례이다. 이러한 리뷰들은 표면적으로는 긍정적인 단어(예: “최고”, “걸작”)로 구성되어 있어 단순한 감성 분석으로는 파악하기 어렵다. 이는 영화의 실제 평판이라는 외부적 맥락을 인지해야만 해석 가능한 고난도 풍자 데이터이므로, 본 모델의 문맥 추론 능력을 평가하기에 최적의 자료가 된다.

데이터 수집은 Python의 Selenium WebDriver를 활용하여 동적 웹 크롤링을 수행하였으며, KoCoSa 데이터셋의 구조인 (Context, Response) 쌍을 맞추기 위해 다음과 같은 전략을 수립하였다.

1. **맥락(Context) 데이터 수집:** 개별 댓글만으로는 영화의 전반적인 부정적 여론을 파악하기 어렵다는 점을 보완하기 위해, 영화에 대한 배경지식과 대중의 혹평을 상세히 기술한 특정 블로그 포스트 [6]를 선정하였다. By.CLASS_NAME을 통해 본문 텍스트를 추출하고, 이를 해당 영화 리뷰 데이터의 공통적인 맥락(Context) 정보로 설정하여 모델에 주입하였다.
2. **반응(Response) 데이터 수집:** 네이버 영화 검색 결과 페이지의 ‘관람평’ 탭에 접근하여, 스크롤링을 통해 동적으로 로딩되는 리뷰 데이터를 수집하였다. 이때 평점 정보와 리뷰 텍스트를 함께 수집하여, 평점은 높지만 내용은 풍자적인 데이터의 특성을 확인하였다.
3. **데이터 포맷팅 및 라벨링:** 수집된 데이터는 모델의 입력 형식과 일치시키기 위해 KoCoSa 포맷인 JSON 형식으로 변환하였다.
 - context: 앞서 수집한 블로그의 영화 비평 요약
 - response: 개별 유저의 영화 리뷰
 - label: 풍자 여부 (1 또는 0)

이 과정을 통해 구축된 데이터셋은 학습된 모델이 실제 인터넷상의 복합적인 언어 유희와 고맥락 풍자를 얼마나 정확하게 탐지해낼 수 있는지 검증하는 척도로 활용되었다.

6.2.2 데이터 정보

구축된 테스트셋에 대하여 실험에서 가장 우수한 성능을 보였던 Context-Aware KcBERT 모델을 사용하여 추론을 수행하였다. 본 실험의 핵심은 모델이 학습 단계에서 본 적 없는 새로운 도메인인 영화 리뷰에서도, 주입된 문맥 정보를 바탕으로 반어 표현을 포착할 수

| |
|--|
| 리뷰: 살다살다 이런 전위적인 괴작... 반어법 확률: 24.4% |
| 리뷰: 옆에서 폰을 해도 화나지 않... 반어법 확률: 52.7% |
| 리뷰: 진짜 2017년 최악의 영화... 반어법 확률: 18.1% |
| 리뷰: 이것을 보느니 집에서 천장만... 반어법 확률: 77.2% |
| 리뷰: 나만 당할순 없다 반어법 확률: 4.0% |
| 리뷰: 제가 팝콘먹으면서 봐도 사람들... 반어법 확률: 1.9% |
| 리뷰: 예휴... 감독이 이 배우들... 반어법 확률: 37.1% |
| 리뷰: 아... 그냥 아..... 반어법 확률: 9.2% |
| 리뷰: 내용도없고...팩트도없고..... 반어법 확률: 6.8% |
| 리뷰: 김수현 성동일 연기는 진짜 ... 반어법 확률: 0.6% |

Figure 7: 영화 리뷰 반어법 분석

있는지 검증하는 데 있다.

- **입력 시퀀스 구성:** 모델의 입력 형식인 [CLS] Context [SEP] Response [SEP] 구조를 준수하기 위해 다음과 같은 전처리 파이프라인을 적용하였다.
 - Context (배경 지식): 크롤링 단계에서 확보한 '영화 리얼의 혹평과 줄거리'를 다른 블로그 텍스트를 문맥으로 고정하였다. 이는 모델에게 "이 영화는 대중적으로 평가가 좋지 않다"는 사전 정보를 제공하는 역할을 한다.
 - Response (사용자 반응): 네이버 영화 평점란에서 수집한 개별 관람평을 반응 데이터로 매핑하였다.
- **추론 및 확률 계산:** 전처리된 시퀀스를 모델에 입력하여 각 클래스(Sarcasm/Non-Sarcasm)에 대한 확률값을 계산하였다.

6.2.3 실험 결과

실험 결과, KoCoSa 데이터셋으로 학습된 모델은에서 볼 수 있듯이 "나만 당할 수 없다"는 식의 고막락 풍자가 섞인 영화 리뷰들을 효과적으로 탐지해 내었다. 주요 분석 결과는 다음과 같다.

- 반어적 칭찬(Sarcastic Praise)의 탐지 모델은 표면적으로는 긍정적인 단어를 사용했으나, 실제로는 영화의 낮은 완성도를 비꼬는 리뷰들을 높은 확률로 Sarcasm으로 분류하였다.
 - 사례 1: "옆에서 폰을 해도 화나지 않는 영화...?? 정말 최고예요"

분석: '최고예요'라는 긍정 어휘가 사용되었으나, '폰을 해도 화나지 않는다(영화에 집중할 가치가 없다)'는 문맥적 의미를 포착하여 Sarcasm으로 정확히 분류하였다.
 - 사례 2: "이것을 보느니 집에서 천장만 보는 게 더 재밌겠어요"

분석: 비교 대상을 극단적으로 설정(천장 보기)하여 우회적으로 영화를 비판한 표현을 감지하고 Sarcasm으로 판별하였다.
- 멈(Meme) 기반의 풍자 인식 한국 인터넷 문화 특유의 '물귀신 작전' 심리가 반영된 리뷰 또한 성공적으로 탐지되었다.
 - 사례 3: "나만 당할 순 없다"

분석: 해당 문장은 감정적 형용사가 부재함에도 불구하고, 망한 영화를 남에게 추천

하여 고통을 분담하려는 전형적인 풍자 의도를 내포한다. 모델은 이를 Sarcasm으로 분류함으로써, 단순 감정 분석을 넘어선 화용론적 의도 파악 능력을 입증하였다.

- 직접적인 비판과의 구분 반어법 탐지의 핵심 과제 중 하나는 '단순 부정(Negative)'과 '반어(Sarcasm)'를 구분하는 것이다. 실험 모델은 명백한 악플에 대해서는 Sarcasm이 아닌 것으로 분류하는 경향을 보였다.

– 사례 4: ”진짜 2017년 최악의 영화... 뭘 보여주려는지??”

분석: '최악'이라는 직접적인 부정 어휘가 사용된 경우, 모델은 이를 비꼬는 표현이 아닌 직설적인 비판으로 인지하여 Non-Sarcasm으로 분류하였다.

결론적으로, Context-Aware 모델은 학습 데이터(KoCoSa)와 다른 도메인의 텍스트에서도 문맥 정보가 주어졌을 때 표면적 의미와 내재적 의도의 불일치를 효과적으로 포착함을 확인하였다. 이는 본 연구에서 구축한 시스템이 실제 온라인 환경의 혐오 표현이나 풍자 댓글 필터링 시스템으로 확장될 수 있는 가능성을 시사한다.

7. 기대효과 및 한계점

본 프로젝트를 통해 구축된 문맥 기반 반어법 탐지 시스템과 실험 결과는 자연어 처리 분야 및 실제 산업 현장에서 다음과 같은 긍정적인 파급 효과를 기대할 수 있다. 그러나 텍스트 데이터의 본질적 한계와 모델의 특성상 해결해야 할 과제 또한 분명히 존재한다.

7.1 기대효과

감성 분석 시스템의 신뢰도 향상: 기존의 감성 분석 모델은 "정말 잘한다"와 같은 반어적 표현을 긍정으로 오분류하여 전체 여론 분석의 정확도를 저해하는 치명적인 단점이 있었다. 본 프로젝트의 모델은 문맥을 파악하여 이러한 '극성 반전' 현상을 포착해낼 수 있으므로, 상품 리뷰 분석, 브랜드 평판 조회 등에서 실질적인 고객의 목소리를 정확하게 수집하는 데 기여할 수 있다.

고도화된 혐오 표현 및 악성 댓글 필터링: 최근의 악성 댓글은 필터링을 회피하기 위해 교묘한 칭찬이나 존댓말을 가장한 비꼬기 형태를 띠는 경우가 많다. 본 프로젝트에서 제안한 시스템은 표면적인 욕설이 없더라도 맥락상 공격적인 의도를 내포한 댓글을 탐지할 수 있어, 보다 건전하고 안전한 온라인 커뮤니케이션 환경을 조성하는 데 활용될 수 있다.

한국어 고맥락 데이터 처리 기술 확보: 영어권에 비해 연구가 부족했던 한국어 반어법 데이터(KoCoSa)를 심층 분석하고, 한국어 특화 모델(KcBERT)의 우수성을 입증함으로써, 향후 고맥락 언어 처리가 필요한 AI 에이전트나 챗봇 개발의 기초 자료로 활용될 수 있다.

7.2 한계점

비언어적 단서의 부재: 반어법은 텍스트뿐만 아니라 억양, 표정, 제스처 등 비언어적 요소에 크게 의존하는 현상이다. 본 프로젝트는 텍스트 데이터만을 기반으로 하므로, "아,

예 알겠습니다”와 같이 텍스트 자체로는 평범하지만 억양에 따라 비꼼이 결정되는 음성적 반어법을 탐지하는 데에는 한계가 있다.

도메인 적응 및 신조어 문제: 본 모델은 메신저 대화 기반의 KoCoSa 데이터로 학습되었기에, 뉴스 기사나 긴 호흡의 소설 등 문체와 호흡이 다른 도메인에서는 성능 저하가 발생할 수 있다. 또한, ”저퀄티비”, ”누칼협” 등 빠르게 생성되고 소멸하는 인터넷 신조어나밈(Meme)이 섞인 반어법의 경우, 모델이 사전에 학습하지 못했다면 문맥 추론에 실패할 가능성이 존재한다.

거대 언어 모델 활용의 최적화 부족: 본 실험에서 GPT 모델은 In-context Learning 방식에서 상대적으로 저조한 성능을 보였다. 이는 한국어 반어법의 미묘한 뉘앙스를 파악하기 위해 단순히 몇 개의 예시(Few-shot)를 제공하는 것만으로는 불충분함을 시사한다. 향후 연구에서는 프롬프트 엔지니어링을 고도화하거나, 파라미터 효율적 미세조정 기법을 적용하여 거대 모델의 추론 능력을 극대화할 필요가 있다.

8. 결론

본 프로젝트는 온라인 텍스트 커뮤니케이션에서 빈번하게 발생하지만 기계적 해석이 어려운 ’반어법(Sarcasm)’을 효과적으로 탐지하기 위해, KoCoSa 데이터셋을 활용한 다각적인 언어 분석과 문맥(Context) 인지형 딥러닝 모델링을 수행하였다.

첫째, 언어 데이터 분석을 통해 반어법의 언어학적 특성을 규명하였다. 형태소 분석 결과, 반어적 문장에서는 과도한 긍정 어휘나 감탄사, 특정 종결 어미가 빈번하게 사용됨을 확인하였다. 또한 의미론적 분석을 통해 문맥과 응답 간의 ’의미적 불일치’와 ’감성적 대비’가 반어법을 결정짓는 핵심 기제임을 정량적으로 입증하였다. 이는 반어법 탐지가 단순한 문장 분류를 넘어, 대화의 흐름과 화자의 의도를 파악하는 고도의 추론 과정임을 시사한다.

둘째, 딥러닝 모델 실험을 통해 문맥 정보 활용의 중요성과 도메인 특화 모델의 유효성을 검증하였다. 실험 결과, 문맥 정보를 함께 학습한 ’Context-Aware’ 모델이 단일 발화만 사용한 모델보다 우수한 성능을 보였다. 특히, 한국어 구어체와 온라인 댓글 특성을 반영한 KcBERT 모델이 거대 언어 모델인 GPT 기반의 접근 방식(In-context Learning 등)보다 월등히 높은 정확도를 기록하였다. 이는 특수한 언어적 뉘앙스와 문화적 맥락이 강한 한국어 반어법 탐지 과제에서는, 범용 모델보다는 해당 도메인에 특화된 모델을 미세조정(Fine-tuning)하는 전략이 더욱 효과적임을 보여준다.

셋째, 실제 웹 환경에서의 적용 가능성을 확인하였다. 영화 ’리얼’의 리뷰 데이터를 수집하여 수행한 정성적 평가에서, 본 연구의 모델은 학습 데이터에 없던 새로운 도메인에서도 ”나만 당할 수 없다”와 같은 고맥락 풍자와 반어적 칭찬을 성공적으로 탐지해 내었다. 이는 구축된 시스템이 제한된 실험 환경을 넘어 실제 온라인 공간의 혐오 표현 필터링이나 여론 분석 시스템으로 확장될 수 있는 실질적인 잠재력을 가짐을 의미한다.

종합적으로 본 프로젝트는 한국어 반어법의 언어적 특성을 데이터에 기반하여 체계적으로 분석하고, 이를 바탕으로 문맥을 이해하는 강건한 탐지 모델을 구축했다는 점에서 의의가 있다. 향후 연구로는 텍스트 정보에 국한되지 않고 이미지나 이모티콘 등 비언어

적 단서를 함께 고려하는 멀티모달 반어법 탐지로의 확장을 제안하며, 이를 통해 인간의 복잡한 의사소통 의도를 더욱 정교하게 파악하는 인공지능 시스템으로 발전시킬 수 있을 것이다.

참고 문헌

References

- [1] Yumin Kim, Heejae Suh, Mingi Kim, Dongyeon Won, and Hwanhee Lee. Kocosa: Korean context-aware sarcasm detection dataset. *arXiv preprint arXiv:2402.14428*, 2024.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [3] Junbum Lee. Kcbert: Korean comments bert. In *Annual conference on human and language technology*, pages 437–440. Human and Language Technology, 2020.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [5] Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, jiwung Hyun, and Sungho Park. A technical report for polyglot-ko: Open-source large-scale korean language models, 2023.
- [6] Naver Blog (hyunjee1110). 영화 ‘리얼’ 배경 지식 및 감상평. <https://blog.naver.com/hyunjee1110/223840133347>, 2025. Accessed: 2025-12-17.