# A Study on Korean Sarcasm Detection Model: CoT-Based Korean Sarcasm Corpus Construction and LLM Fine-Tuning

Myung-goo Kang, Jeong-hoon Seo, In-hwan Kim,
Seung-jae Yang, Ik-kyun Ham, Yu-ri Jang, Soo-in Lee

Department of Industrial Management Engineering
Kyung Hee University
Seoul, Korea

June, 2025

# Contents

# Table Contents

# Figure Contents

# Abstract


**A Study on Korean Sarcasm Detection Model:**
**CoT-Based Korean Sarcasm Corpus Construction**
**and LLM Fine-Tuning**

Myunggoo Kang, Jeonghoon Seo, Inhwan Kim,
Seungjae Yang, Ikkyun Ham, Yuri Jang, Sooin Lee
Department of Industrial Management Engineering
Kyung Hee University
Adivised by Sangwoo Bahn

Sarcasm is an ironic linguistic expression in which there is a discrepancy between the apparent meaning and the actual intention, and it is a key recognition target in various natural language processing (NLP) tasks such as sentiment analysis and opinion detection. To build a sarcasm detection model for Korean, this study generated a prompt-based structured training dataset and conducted a comparative experiment between fine-tuning a BERT-based classification model and an LLM-based interpretation model. The data consists of four layers of structure: context, response, label, and explanation, and includes KoCoSa data, an existing representative Korean sarcasm dataset, as well as various real-life language domains such as SNS conversations, reviews, and comments. We built a systematic data generation pipeline through generative language models and LangChain-based prompting design, and applied Chain-of-Thought (CoT) prompting to generate high-quality explanation data, resulting in a total of 4,800 structured datasets. KoBERT was used for classification model training, SKT/KoBERT-110 for LLM model, and Bllossom-3B for LLM model, and each was fine-tuned independently by applying instruction-style input format based on Bllossom-3B. The experimental results showed that the fine-tuned BERT model had higher accuracy and balanced classification performance than the pre-training state, and the recall rate for sarcastic sentences was significantly improved. The LLM model showed strengths in context interpretation and explanation generation, and demonstrated fine-grained analysis of complex sarcastic expressions.This study confirms that a prompt-based structured Korean sarcasm dataset improves detection performance, and also demonstrates the validity of an independent approach that utilizes the unique characteristics of BERT and LLM.



**Key words**

#Sarcasm #Natural language processing #Korean #Prompt Engineering #Chain of Thought #Fine-tuning

# I. Introduction

## 1.1 Theoretical Background

Sarcasm is generally considered a form of irony. McDonald and Pearce (1996) defined sarcasm as "a form of irony used to hurt or criticize someone," while Lee and Katz (1998) viewed it as "a subtype of linguistic irony that expresses a negative and critical attitude toward a victim or group of victims." Both definitions emphasize that sarcasm targets a specific "victim" to express negative emotions. Sarcasm is difficult to discern based solely on sentence structure or word choice; it requires a comprehensive consideration of context, the speaker's intent, and cultural background. In this regard, sarcasm possesses the characteristics of indirectness and implicitness in meaning transmission, which are primarily addressed in the field of pragmatics. In particular, in the online environment, sarcasm is used more frequently in the form of criticism, ridicule, and cynical humor, and functions as an important discourse tool on social media and in communities.

However, the metaphorical nature and context dependence of sarcasm act as major error factors in existing emotion analysis and opinion analysis models. As Băroiu and Trăușan-Matu (2022) pointed out that sarcasm can cause "semantic noise" that can degrade the performance of sentiment classification models, sarcasm can undermine the reliability of automated language processing systems. To address this issue, recent studies have focused on developing automatic detection models that can more accurately identify the presence of sarcasm and related research.

## 1.2 Literature Review

### 1.2.1 Creation of Sarcasm Dataset

Previous studies have divided the methods for constructing sarcasm datasets into two main categories: distant supervision and manual annotation. Among these, distant supervision involves automatically collecting millions of data points from platforms such as Twitter, Reddit, and Amazon based on hashtags (#sarcasm, /s, etc.), enabling large-scale dataset construction (Davidov et al., 2010; González-Ibánez et al., 2011). Twitter is the most widely used platform due to its concise sentence structure and ease of data collection. After collection, non-English tweets, retweets, and hashtags that are not at the end of a sentence are filtered out (González-Ibánez et al., 2011). Barbieri et al. (2014) also analyzed lexical and statistical characteristics to propose an alternative method.

Manual annotation involves human annotators determining whether a statement is sarcastic, typically based on third-party perception, though recently, methods where the speaker directly indicates their intent (iSarcasm; Oprea and Magdy, 2019) have also been utilized. Recently, data from various sources such as Reddit, news, and books have been utilized, and there is a clear trend toward preferring manually refined data over randomly collected data (Băroiu and Trăușan-Matu, 2022). This is interpreted as being due to the high noise level of SNS-based data, which makes it difficult to understand the subtle context of sarcastic expressions.

### 1.2.2 Sarcasm Detection

Sarcasm detection research has gradually developed over the past decade in line with trends in natural language processing. Early studies utilized traditional supervised learning-based classifiers such as SVM (Cristianini and Shawe-Taylor, 2000), logistic regression, decision trees (Quinlan, 1986), naive Bayes, and random forests (Breiman, 2001). Subsequently, these approaches were extended to deep neural network-based methods such as CNN (LeCun et al., 1998), LSTM (Hochreitter and Schmidhuber, 1997), and other deep neural network-based approaches, and recently, pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have become the primary methods.

Deep learning-based sarcasm detection began in earnest around 2015. Bamman and Smith (2015) emphasized the importance of context in understanding sarcasm and proposed an RNN-based classifier, while Rajadesingan et al. (2015) improved accuracy by utilizing user behavior information, and Joshi et al. (2015) developed a feature-based model that captures contextual inconsistencies. In 2016, Zhang et al. (2016) applied deep learning models to Twitter data, and Ghosh and Veale (2016) proposed a neural network-based classifier. Schifanella et al. (2016) attempted multimodal sarcasm detection.

Since 2017, the integration of conversational context and external knowledge has emerged as a major research topic. Ghosh et al. (2017) analyzed the influence of context within consecutive utterances, and Hazarika et al. (2018) proposed the CASCADE model. Felbo et al. (2017)'s DeepMoji demonstrated

excellent performance by fine-tuning after pre-training emoji prediction.

2018 marked a turning point in sarcasm detection research. Khodak et al. (2018) constructed the SARC dataset based on 1.1 million /s-tagged comments, and the sarcasm and irony detection shared task was first introduced in SemEval-2018. Tay et al. (2018) proposed a model that became the starting point for research on improving explainability.

After 2019, the adoption of pre-trained language models (PLMs) became widespread. Potamias et al. (2020) achieved state-of-the-art performance using a Transformer encoder-based architecture, while Baruah et al. (2020) achieved state-of-the-art performance using a BERT-based model that reflects conversational context. Castro et al. (2019) and Cai et al. (2019) proposed hierarchical fusion methods for images and text.

Since 2022, sarcasm detection research has rapidly evolved, centered on Transformer-based language models. Helal et al. (2022) integrated conversational context into a RoBERTa-based model, achieving F1 scores of 99% and 90% on the news headline and MUStARD datasets, respectively. Abu Farha et al. (2022) attempted self-sarcasm detection using ensemble models such as BERT and RoBERTa on SemEval-2022.

In 2024, Montgomery Gole et al. compared the zero-shot sarcasm detection performance of LLMs such as GPT-3.5, GPT-4, Claude, and ChatGPT, but they still showed lower performance than fine-tuned BERT models.

In 2025, Li et al.'s Sarcasm-GPT gained attention. This model constructed a generative sarcasm detection framework by combining prompt generation, RAG, chain-of-thought reasoning, and a multimodal integration pipeline.

As such, sarcasm detection research has rapidly evolved from traditional machine learning to neural network-based approaches and pre-trained language models, and more recently to large-scale generative language models and prompt-based intelligent frameworks. Various strategies have emerged to address challenges in high-dimensional language understanding, such as context comprehension, external knowledge integration, explainability, and multimodal processing. These strategies reflect the recent research trend of recognizing sarcasm not as a simple classification but as a complex semantic interpretation target.

#### 1.2.3 Research on sarcasm detection in Korean

Research on sarcasm detection in Korean began in earnest around 2020. Gong-joo Lee and Ji-eun Kim (2019) focused on rhetorical questions that frequently appear in sports news comments and showed that interrogative endings such as "~ᄒᆞ냐?" can be important clues to satirical expressions. They proposed a method for automatically recognizing satirical comments through machine learning based on the use of the "~냐" ending, which is considered an early example of utilizing Korean-specific morphemes and ending systems for sarca detection.

Subsequent efforts were made to construct Korean corpora related to sarcasm detection. Jeon et al. (2022) released the KOTE (Korean Online Comments Emotion) dataset, which attached 43 detailed emotion labels to over 50,000 Naver News comments. This corpus includes emotion tags similar to sarcasm, such as "cynicism," "mockery," and "parody," demonstrating the possibility of learning the nuances of satirical expressions within a multi-emotion classification system rather than a binary classification system. However, KOTE was not designed specifically for satirical expressions but for general emotion classification, making it difficult to consider it optimized for sarcasm recognition. In particular, it focuses on the surface text of individual comments rather than the overall context of the sentence or the speaker's intent, and the absence of clear answer labels for sarcasm recognition imposed limitations on supervised learning.

To address these limitations, Kim et al. (2024) released KoCoSa (Korean Context-aware Sarcasm Detection Dataset). KoCoSa is the first Korean-language dataset dedicated to sarcasm detection, with 12,824 everyday two-person conversations explicitly labeled as to whether the last utterance was sarcastic. Most importantly, it is designed to determine sarcasm within the context of the speaker and the flow of the conversation, enabling much more sophisticated and realistic sarcasm recognition than existing sentiment corpus-based analysis.

Therefore, KoCoSa is evaluated as the most accurate and useful benchmark data among the Korean sarcasm detection data released to date, and it can serve as a benchmark and starting point for future sarcasm recognition research in Korea and abroad. Based on the structural strengths of KoCoSa, this study aims to further develop and expand it to present new possibilities and practical applications for Korean sarcasm detection.

### 1.3 Research Objectives

To date, sarcasm detection research has shown progress in both dataset construction and detection model development. In particular, in English-speaking countries, various attempts have been made, including the construction of large-scale corpora based on various domains, the introduction of transformer-based language models, and the integration of author, conversation context, and multimodal information. As a result, remarkable achievements have been accumulated in both model performance and explainability.

In contrast, sarcasm detection research in Korean remains in its early stages, with the following limitations clearly evident.

First, most Korean sarcasm-related datasets are derived from general sentiment classification purposes and do not have a structure or task design specialized for sarcasm detection. Second, there are structural imbalances, such as unclear answer labels for sarcasm and explanations provided only for sarcastic responses, which limit the efficiency and explainability of model learning. Third, there is a lack of domain diversity, which means that the models do not reflect real-world language environments and rely too heavily on surface-level information at the sentence level, making it difficult to effectively capture the core elements of sarcasm, such as “irony in context” and “speaker intent.”

In English-speaking countries, multidimensional analysis that integrates the author's intent (intended sarcasm), conversational context, and surrounding information such as images and emojis is actively being conducted, but such attempts are almost non-existent in Korean research. Therefore, there is an urgent need to build a high-quality dataset that accurately reflects the context for Korean sarcasm detection. This study aims to address these limitations by creating a structured, high-quality dataset for training a Korean-based sarcasm detection model. Specifically, by utilizing a generative language model (LLM), we will automatically generate data with a four-part structure consisting of “context–response–sarcasm presence–explanation” to create a sarcasm expression corpus that reflects realistic domains such as everyday conversations, social media, reviews, and comments.

To this end, this study sets the following two specific objectives.

1. Improving the structural limitations of KoCoSa

In the existing KoCoSa dataset, explanations were only provided for satirical responses, limiting learning about non-satirical expressions. This study designs an integrated structure that includes sarcasm status and explanatory rationale for both satirical and non-satirical responses, thereby improving both the efficiency and accuracy of developing an explainable LLM-based sarcasm detection model.

2. Ensuring domain diversity and realism

Sarcasm is an expression that frequently appears in various informal language environments such as social media, comments, and reviews. This study aims to build a dataset that closely resembles real-world language use by collecting contextual information from multiple domains that reflect such colloquial and informal contexts, and automatically generating quadruple structured data based on this information. This will improve the model's generalization performance and increase its applicability in real-world environments.

In conclusion, this study provides a practical infrastructure for developing Korean LLM-based sarcasm recognition models through the construction of such a dataset, and will also contribute to a wide range of language-based AI research, including explanatory sentiment analysis, high-dimensional meaning inference, and social context understanding.

## II. Research Methods

### 2.1 Prompt Engineering

Existing Korean sarcasm datasets have the following limitations in terms of size and sentence structure. Representative public datasets such as KoCoSa have significant limitations in terms of sentence length, contextual diversity, and domain scalability, and most of them focus on short satirical expressions based on everyday conversations. Satirical expressions are metaphorical and highly context-dependent, making it difficult to adequately reflect their nuances using simple binary classification labels alone. As a result, sophisticated relationship-based learning data is required.

This study defined prompts based on a triple structure of “intent–context–expression” and designed a data generation framework accordingly.

Systematic prompt engineering was performed to enable the large-scale generation of structured learning data using generative language models.

Prompt engineering is a technique for carefully designing input sentences so that generative language models can generate text that meets specific objectives. It goes beyond simple instructions, assigning roles to the model, specifying the style, and clearly requiring logical flow and output format. This study adopted prompt engineering as a core data generation method for the following three reasons. First, it was necessary to generate complex structured data that included sarcasm/non-sarcasm responses and explanations for them, rather than simple responses. Second, we wanted to induce a natural style close to conversational language, which was achieved through role-based design and the provision of examples. Third, since it was necessary to control the output values in a predefined structure (JSON format), we enabled structured data generation by linking LangChain and PydanticOutputParser.

This section provides a detailed explanation of the process of constructing four types of datasets based on such prompt engineering. This design process enables the automatic generation of high-quality structured data, which is significant as an independent experimental area distinct from the subsequent model training stage.

#### 2.1.1 KoCoSa-based data generation

KoCoSa (Korean Corpus for Sarcasm Analysis) is a dataset that provides structured data for sarcasm detection based on everyday Korean conversations. In this study, we sought to maintain the structural characteristics of KoCoSa while addressing the limitations of the existing dataset. Specifically, to address the issue that non-sarcasm sentences in the original data lacked explanatory items, we generated explanatory sentences for non-sarcasm sentences and restructured the existing sarcastic sentences into an integrated structure.

The prompt design was structured as a system-user model, assigning the model the role of a “Korean conversation analysis assistant.” The input values are the conversation context and response, and the output values are 1-2 sentences explaining why the response is non-satirical. A zero-shot prompting strategy was applied, and the OpenAI GPT-4o-mini model (temperature=1.0) was used.

The training data and evaluation data consisted of 1,000 and 200 samples, respectively, and were stored in CSV and JSONL formats. We implemented structured output generation using the LangChain framework and PydanticOutputParser, and improved the consistency of the overall data structure and learning efficiency by supplementing the missing information in KoCoSa.

#### 2.1.2 Conversation-based data generation

To reflect the characteristics of satirical expressions frequently found in everyday conversations on social media, we utilized AI Hub's “Topic-based Everyday Conversation Text Dataset.” This dataset includes multi-party colloquial conversations collected from KakaoTalk, Facebook, Instagram, Band, and NateOn. Data generation consisted of three steps.

In the first stage, we summarized and extracted the context based on subject, speaker_type, and text from the JSON raw dialogue, and classified the sentiment as positive or negative. Since satirical expressions primarily occur in situations with negative sentiment, we proceeded with subsequent sentence generation only for cases classified as negative sentiment. We used the GPT-4o-mini model, and set the temperature to 0.2 to ensure an accurate summary of the context.

In the second step, we generated sarcasm and non-sarcasm responses based on the context. Since understanding the emotions and intentions within the context is crucial for satirical sentences, we combined the Chain-of-Thought (COT) method with zero-shot prompting to generate sentences in the following step-by-step process:

(1) Situation summary

(2) Emotion keyword extraction

(3) Sarcastic sentence generation

(4) Non-sarcastic sentence generation

The CoT method contributed to the model generating more logical and natural sarcastic sentences based on contextual judgment. To ensure diversity and creativity in expression, the GPT-4o model (temperature=1.0) was used.

In the third step, explanation was generated by receiving context, response, and label as input values. The reason why each sentence was classified under the corresponding label (Sarcasm/Non-Sarcasm) was explained in a free-form sentence, and it was

designed to be used as a learning signal for improving the model's explanatory power in the future. The GPT-4o-mini model was used in this step.

A total of 1,200 samples (1,000 for training and 200 for evaluation) were generated, and all tasks were automated based on LangChain's Composable Chain structure (prompt → model → parser).

#### 2.1.3 Review-based data generation

To reflect the review domain where satirical expressions frequently appear, this study collected a total of 60,000 negative reviews from Naver shopping reviews (1, 2 stars) and Steam game reviews (including negative labels). A sample of these reviews was used to generate the data.

The prompts were structured in YAML format, with the original review text as the input variable, designed to generate sarcasm and non-sarcasm sentences that align with the intent of the review. Additionally, explanations for both sentences were generated. The GPT-4o and GPT-4o-mini models were used, with the model dynamically selected based on sentence length. The temperature was fixed at 0.5.

Two samples were generated from each review, and a total of 1,200 samples were created, consisting of 1,000 for training and 200 for evaluation. LangChain and LangSmith were used to perform repeated experiments on the prompt, error tracking, and data quality management. Abnormal responses were removed through manual filtering.

#### 2.1.4 Comment-based data generation

The comment domain contains many short and unstructured texts, which can reflect the actual distribution of satirical expressions. For this purpose, this study selected only comments containing 18 negative emotions from the KOTE (Korean Online That-gul Emotions) dataset. The emotion classification criteria were based on the Gunsan University Emotional Vocabulary Dictionary.

Complaint/grievance, annoyance, anger/rage, regret/disappointment, suspicion/distrust, despair, contempt, disgust/repulsion, irritation, bewilderment, defeat/self-loathing, fatigue/exhaustion, hatred/loathing, embarrassment/awkwardness, horror, burden/anxiety, boredom, anxiety/worry

Data generation was performed in three stages. In the first stage, the context was summarized in the form of “a situation where ~,” and satirical expressions were generated using the GPT-4o model. In the second stage, non-sarcasm expressions and their explanations were generated using GPT-4o-mini for the same context. In the final stage, explanations for the satirical sentences were generated using the same model. Each stage used different instances to prevent response dependency and monotony in expression. The temperature was set to 0.5 in all stages.

Based on 500 comments, a total of 1,200 samples were created, consisting of 1,000 for training and 200 for evaluation. The prompt-model structure followed LangChain's Composable Chain method, and expressions such as “presumed to be the case” were inserted into the prompt text to ensure stylistic diversity and logical consistency. This prevented overfitting and maximized the LLM learning effect.

### 2.2 Fine-tuning

Fine-tuning is a transfer learning technique that involves further training a pre-trained language model for a specific task or domain, playing a key role in securing specialized performance that general-purpose language models lack. Pre-trained models learn general language patterns and knowledge through large-scale text corpora, but for tasks involving special language phenomena such as sarcasm detection, additional training using domain-specific data is essential. Sarcasm is a complex linguistic phenomenon involving a mismatch between surface meaning and the speaker's actual intent, requiring both contextual clues and pragmatic inference, making it difficult to achieve sufficient performance with pre-training alone.

In this study, we performed fine-tuning on BERT-based classification models and LLM-based generative models using a structured sarcasm dataset. For BERT-based models, we conducted training specialized for optimizing sarcasm/non-sarcasm binary classification performance using 4,800 custom data points constructed in this study and 14,000 samples from the existing KoCoSa dataset. For LLM models, we also conducted fine-tuning using both custom data and KoCoSa data to enable simultaneous sarcasm detection and interpretation generation. This approach aims to enhance the transparency and practicality of the model by providing not only simple

binary classification but also the basis for sarcasm detection and interpretation.

### 2.2.1 Model Selection

Sarcasm is a linguistic device based on the ironic discrepancy between the speaker's intention and the content of the utterance, and it is difficult to clearly understand the intention of a sentence using only a simple language model (McDonald & Pearce, 1996; Lee & Katz, 1998). Accordingly, this study employed two different approaches for sarcasm recognition. The first approach utilizes a large language model as an interpreter to determine sarcasm through context-based natural language inference, while the second approach employs a pre-trained language model from the BERT family as a classifier to train it as a binary classification task (Devlin et al., 2019).

LLMs leverage extensive pre-trained knowledge based on advanced linguistic reasoning capabilities and can perform prompt-based instructions (Brown et al., 2020). They are particularly advantageous in that they can demonstrate high-level language abilities essential for detecting sarcasm, such as contextual meaning interpretation, pragmatic inference, and common-sense-based judgment (Bamman & Smith, 2015; Ghosh & Veale, 2017). Actual satirical expressions often require judgments based on the speaker's intent rather than direct meaning, and in this regard, LLM can demonstrate more flexible interpretive capabilities.

On the other hand, BERT-based models are structured to optimize the classification of input sentences by vectorizing them, offering the advantage of achieving high performance through fine-tuning even with relatively small amounts of data. When there is a clear correct answer regarding whether something is sarcasm and a large dataset with labels is provided, such classification models provide a reliable baseline for performance evaluation and model comparison (Riloff et al., 2013).

Accordingly, this study aimed to achieve the following objectives by conducting parallel experiments using an interpreter-based approach (LLM) and a classifier-based approach (BERT). First, to examine the extent to which the two approaches are complementary in sarcasm detection through quantitative performance comparison. Second, to analyze how the natural language inference capabilities of LLM can complement the limitations of existing classifiers. Third, to explore the potential for multimodal or hybrid approaches by observing how different model architectures respond to the high-dimensional linguistic phenomenon of sarcasm.

In conclusion, the parallel use of LLM and BERT-based models provides a foundation for approaching the non-surface-level language understanding task of sarcasm detection from various perspectives and serves as a key comparative framework for evaluating the balance between model interpretability and reasoning ability.

### 2.2.2 BERT Fine-Tuning

This study aims to build a model that classifies sarcasm based on KoBERT (skt/kobert-base-v1-110M), a pre-trained language model specialized for Korean, and to improve the model's sarcasm recognition performance by training it on a prompt-based dataset. To this end, a series of experimental procedures were designed, including prompt construction, input preprocessing, model structure, training strategy, evaluation method, and baseline comparison.

First, this study used prompt-based sentences as the model's input text. Unlike existing sentiment classification or sarcasm detection methods that use only sentence-level inputs, this study introduced a structure consisting of “context + response + sarcasm.” This structure aims to guide the model to interpret contextual relationships and evaluate responses according to specific instructions, rather than simply analyzing the vocabulary or style of the sentence. Specifically, the input was structured as follows.

**Table 1** sarcasm classification structure

| | |
|---|---|
| **prompt** | Let's read the next one,<br>Classify whether the following statement is sarcasm or not. |
| **Context** | {context} |
| **response** | {response} |

These prompts guide the model to learn to infer input sentences based on natural language instructions, which is similar to the instruction-based fine-tuning method that has recently gained attention in the field of natural language processing. In this study, we sought to enable the model to understand the high-dimensional language expression of sarcasm in a more structured manner.

In the text preprocessing stage, the labels were refined into a binary classification task based on two classes: “Sarcasm” and “Non-sarcasm.” The labels are mapped to 1 (Sarcasm) and 0 (Non-sarcasm), respectively, and classification learning is conducted based on this. The preprocessed data was pre-separated into training and validation sets to prevent data leakage between the training and evaluation stages.

The model structure was based on the monologg/kobert model from the Hugging Face Transformers library, which is a Korean BERT model pre-trained by SKT Brain. A linear classification head with two output neurons was added on top of this model, and the entire model was constructed using the Auto Model For Sequence Classification class. The input was tokenized using KoBERT Tokenizer, which uses the same vocabulary as the model, and input_ids and attention masks with a maximum length of 128 were generated and used as input. During training, CrossEntropyLoss was used as the loss function, and the AdamW optimizer was applied. The learning rate was set to 2e-5, the batch size was fixed at 16, and after training was completed, performance evaluation was performed using the validation dataset. Accuracy was used as the evaluation metric, and the classification performance was measured by comparing the model's predicted values with the actual correct values for each input sample. In the evaluation loop, input_ids were decoded again through a tokenizer to confirm whether inference was performed based on the data input values. This allowed us to output and analyze the input text, predicted values, and actual values in parallel, thereby ensuring the interpretability of the model's output.

### 2.2.3 LLM Fine-Tuning

For LLM-based fine-tuning, we selected the “Bllossom (llama-3.2-Korean-Bllossom-3B)” model, which is specialized for Korean and based on the LLaMA-3.2 architecture (Park et al., 2025). The main reasons for selecting this model are as follows.

1. It retains the powerful language modeling capabilities of the original LLaMA-3.2 while being further trained to suit the linguistic characteristics of Korean, resulting in excellent contextual understanding and generation quality for Korean text.

2. With a 3B parameter size, it is relatively lightweight, enabling efficient fine-tuning even in limited computing environments.

3. It has pre-trained instruction-following capabilities, making it suitable for prompt-based learning.

The dataset used in this study has a four-part structure consisting of context describing the sarcasm situation, response representing the actual sarcastic utterance, label indicating whether it is sarcasm or not (Sarcasm/Non-Sarcasm), and explanation describing the basis for the judgment. By leveraging the generative nature of LLM models, we designed fine-tuning to enable the model to generate not only simple classification results but also explanations. To this end, we constructed instruction-style prompt templates, with a representative example shown in Figure 1.

However, LLM models have billions of parameters, making it challenging to sufficiently train the entire model with only 4,000 prototype training data points. Furthermore, due to computing resource constraints, it was not feasible to update all parameters simultaneously. To address these issues, we applied the LoRA (Low-Rank Adaptation) technique, a parameter-efficient fine-tuning method, from the PEFT (Parameter Efficient Fine-Tuning) framework. LoRA fixes the weight matrix of the existing model and only trains small adapter matrices that have been decomposed into low ranks, achieving performance close to that of full fine-tuning while updating less than 1% of the total parameters. This enables effective domain adaptation while significantly reducing memory usage and training time (Hu et al., 2022).

During the training process, experiments were conducted with 1, 3, and 5 epochs, considering the constraints on the amount of data, and hyperparameters were adjusted to prevent overfitting while ensuring sufficient training. Specifically, we comprehensively applied techniques for stable learning, including core parameters such as micro_batch_size, learning_rate, warmup_ratio (0.06), gradient_accumulation_steps, etc., as well as fp16 precision, adamw_torch optimizer, and gradient clipping through max_grad_norm. As an evaluation strategy, we secured optimal checkpoints by evaluating each epoch and saving the model every 50 steps.

In addition, we also performed fine-tuning using 14,000 data points from the existing Korean sarcasm dataset, KoCoSa. This was done to quantitatively evaluate whether sufficient performance improvements (sarcasm classification and explanation generation) could be achieved using only the existing dataset, as well as to assess the additional contribution of the explanation-included data constructed in this study. The KoCoSa dataset has a relatively large scale, which is expected to enable more stable training, allowing for a comparative analysis of how data scale and structuring affect LLM fine-tuning performance.

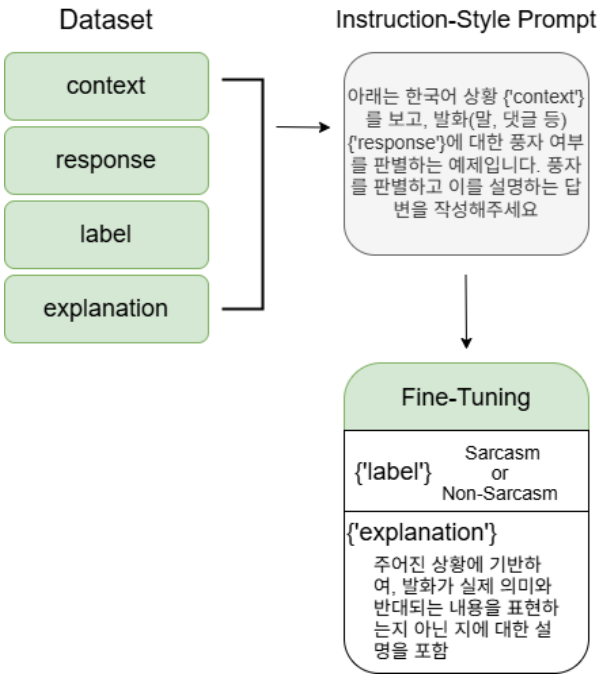


**Figure 1** LLM fine-tuning instruction-style prompt

"Below is an example of determining whether a response (such as speech or comments) is satirical based on the context in Korean. Please determine whether it is satirical and write an explanation."

The model's response was trained to include a label indicating whether it is satirical and an explanation of the basis for that determination.

## III. Results

### 3.1 Final Dataset

In this study, a total of 4,800 sarcasm and non-sarcasm expression datasets were constructed through a total of four prompt-based data generation processes. Each data set consists of four items: context, response, label (satirical/non-satirical), and explanation, and was stored in CSV and JSON formats. The dataset consists of a total of 4,800 samples, evenly distributed across KoCoSa-based, conversation-based, review-based, and comment-based data, each accounting for 25% of the total. The composition is as follows:

**Table 2** Data Corpus

| Data Source | Number of samples for training | Number of samples for evaluation | ratio |
|---|---|---|---|
| KoCoSa Data | 1000 | 200 | 25% |
| Conversation-based data | 1000 | 200 | 25% |
| Review-based data | 1000 | 200 | 25% |
| Comment-based data | 1000 | 200 | 25% |
| total | **4000** | **800** | **100%** |

For all sentences in the final dataset, we described the reasons why each sentence was classified as sarcasm or non-sarcasm. This differs from the existing KoCoSa dataset, which only provided explanations for satirical sentences. Our dataset was designed to enable LLM to learn better by adding logical explanations for non-satirical sentences as well.

Additionally, this dataset includes various domains such as everyday conversations, social media conversations, reviews, and comments, enabling the learning of sarcasm expressions in contexts similar to real-world usage. This enhances the practicality of the dataset for real-world applications after LLM training.

The entire data generation process was executed using LangSmith to track the workflow and manage error responses in real time, thereby refining data quality. As a result, the generated dataset has established a corpus of data with consistency and reliability.

## 3.2 Model Performance

### 3.2.1 BERT

This study utilized KoBERT as a classifier model for detecting Korean sarcasm, and performed fine-tuning on both the existing public dataset (KoCoSa) and the newly constructed prompt-based dataset under the same parameter conditions, then compared their performance. Both datasets consisted of approximately 1,000 training samples and 200 evaluation samples, and the input structure and evaluation method were kept the same.

The experimental results showed that when the KoBERT model was trained for 1 epoch using the KoCoSa dataset, the accuracy was 0.59. On the other hand, when trained for only one epoch with the prompt-based dataset in this study, the accuracy reached 0.82, showing a performance difference of approximately 23 percentage points. This suggests that the data structure has a significant impact on model performance in complex semantic interpretation tasks such as sarcasm detection.

Such performance differences can be interpreted as resulting from differences in data structure. KoCoSa has a structure in which labels are assigned to the last response in a two-person conversation, and in many cases, the context supporting the response is fragmentary or omitted. In addition, there are no explanations for non-satirical responses, making it difficult for the model to learn the logical structure and intentional neutrality of non-satirical expressions. As a result, the BERT model relies solely on fragmentary vocabulary patterns and sentence structures without sufficient linguistic clues for label prediction, resulting in an inability to effectively learn the complex boundaries between sarcasm and non-sarcasm.

In contrast, the dataset proposed in this study includes explicit prompts, and the context and response are clearly distinguished and input into the model in the form of a single instruction. This “instruction-style” input allows pre-trained models such as BERT to learn the context structure in a more structured way and encourages them to perform meaning-based inference rather than simple sentence classification. For example, a sentence such as “Read the following situation and classify whether the following response is sarcasm or not” explicitly requires the model to interpret the context and infer the intention, which acts as a mechanism to effectively bring out BERT's linguistic expression capabilities.

In addition, this dataset includes colloquial expressions from various domains, such as social media conversations, reviews, and comments, exposing the model to a wider variety of expressions and contexts. As a result, BERT can learn the characteristics of sarcasm, such as tone, mockery, exaggeration, and irony, in a sophisticated manner, which led to improved recall, especially in the sarcasm class. The tendency for accuracy to be highest at the beginning of training and then decrease in later epochs suggests that the model tends to overfit to specific language patterns, which can also be interpreted to mean that the prompt-based input applied in this study quickly led to initial learning effects.

In conclusion, this study demonstrated that input format, data structure, and context provision methods have a decisive impact on sarcasm detection performance even under the same BERT structure. Especially in tasks requiring high-dimensional semantic interpretation such as sarcasm, overcoming the limitations of BERT models necessitates directive-based learning structures and explicit context inputs rather than simple classification labels.

### 3.2.2 LLM

This section presents the results of fine-tuning experiments using BLOSSOM-3B, a Korean-specialized medium-sized language model based on the LLaMA-3.2 architecture. Parameter-efficient fine-tuning was performed using the LoRA (Low-Rank Adaptation) technique based on a total of 4,000 structured sarcasm detection datasets, and model performance was comprehensively evaluated from the dual perspectives of sarcasm detection accuracy and explanation generation quality. The experimental design includes tracking performance changes according to the number of epochs (1, 3, 5 epochs) and comparing performance with the existing KoCoSa dataset (14K samples).

**Table 3** Model Performance Comparison

| Model Configuration | Training Data | Number of samples | Accuracy (%) |
|---|---|---|---|
| **BLOSSOM-3B (zero-shot)** | - | - | 50.12 |
| **BLOSSOM-3B + LoRA** | Custom Data (1 epoch) | 4,000 | 58.77 |
| | Custom Data (3 epochs) | 4,000 | 62.22 |
| | Custom Data (5 epochs) | 4,000 | 64.69 |
| | KoCoSa Data | 14,000 | 57.28 |
| **KoBERT** | Custom Data (1 epoch) | 4,000 | 82.45 |
| | Custom Data (3 epoch) | 4,000 | 69.23 |
| | Custom Data (5 epoch) | 4,000 | 61.46 |
| | KoCoSa Data | 14,000 | 59.89 |

In the zero-shot state, the baseline performance of the BLOSSOM-3B model remained at 50.12%, which is at the level of random classification, but gradual performance improvement was observed through LoRA-based fine-tuning. After five epochs of training, the highest accuracy of 64.69% was achieved, showing a performance improvement of 14.57%p compared to the baseline. Notably, the 4,000-sample custom dataset used in this study outperformed the 14,000-sample KoCoSa dataset by 7.41 percentage points. This demonstrates the effectiveness of structured data design based on prompt engineering and multi-domain sampling strategies. However, when compared to the 82% accuracy of the KoBERT model in the same task, the classification performance of medium-sized LLMs still shows a significant gap. This is interpreted as being related to the structural representation capacity limitations (capacity bottleneck) of LLM models in small-scale dataset environments, as pointed out in previous studies by Pu et al. (2023), Han et al. (2024), and Lin et al. (2024), despite the use of LoRA-based PEFT techniques.

**Table 4** BLOSSOM-3B model Description Generation Quality Assessment

| Model Configuration | BLEURT Score (-1.5 - +1.5) | | GPT-4 Score (0 - 10) | |
|---|---|---|---|---|
| | mean | medium | mean | medium |
| **BLOSSOM-3B (zero-shot)** | -0.060 | -0.027 | 2.32 | 2 |
| **BLOSSOM-3B + LoRA (5 epochs)** | 0.388 | 0.403 | 8.49 | 10 |
| **BLOSSOM-3B + LoRA (KoCoSa)** | -0.552 | -1.000 | 7.62 | 9 |

Explanation generation capabilities were analyzed using the BLEURT automatic evaluation metric and GPT-4-based quality evaluation. BLEURT uses a semantic similarity score ranging from -1.5 to +1.5, with values above 0 indicating high semantic consistency (Sellam, 2020). GPT-4 evaluation

comprehensively assesses the logicality, completeness, and appropriateness of explanations on a scale of 0 to 10.

The BLOSSOM model fine-tuned over 5 epochs showed consistent semantic similarity with an average BLEURT score of 0.388 and a median of 0.400, and demonstrated excellent explanation generation capabilities with an average GPT-4 score of 8.49 and a median of 10. In particular, the median score of 10 in the GPT-4 evaluation suggests that the model generated high-quality explanations in most samples.

On the other hand, the model trained solely on the KoCoSa dataset showed relatively low performance, with a BLEURT score of -0.552 and a GPT-4 score of 7.62. This performance degradation is attributed to the structural limitations of the KoCoSa dataset. The fact that only explanations for satirical situations were provided limited the ability to generate explanations for non-sarcasm situations. In particular, the median BLEURT score of -1.000 shows that most of the generated explanations were semantically inappropriate. This is analyzed as being due to difficulties in generating appropriate explanations in various contexts as a result of a monotonous learning pattern centered on sarcasm.

The BLOSSOM model in its zero-shot state as the base model showed a BLEURT score of -0.060 and a GPT-4 score of 2.32, indicating overall poor explanation quality, thereby confirming the necessity of domain-specific training.

The results of this experiment clearly reveal the dual characteristics of medium-sized LLM fine-tuning using LoRA-based PEFT. While it shows limitations compared to traditional encoder models such as KoBERT in terms of classification performance, it demonstrates a significant advantage in complex language generation tasks such as explanation generation. Notably, the four-layer structure (context-response-label-explanation) of the dataset design in this study made a decisive contribution to improving explanation generation quality.

Non-Sarcasm and Sarcasm examples for short Context are shown in Figures 2 and 3.

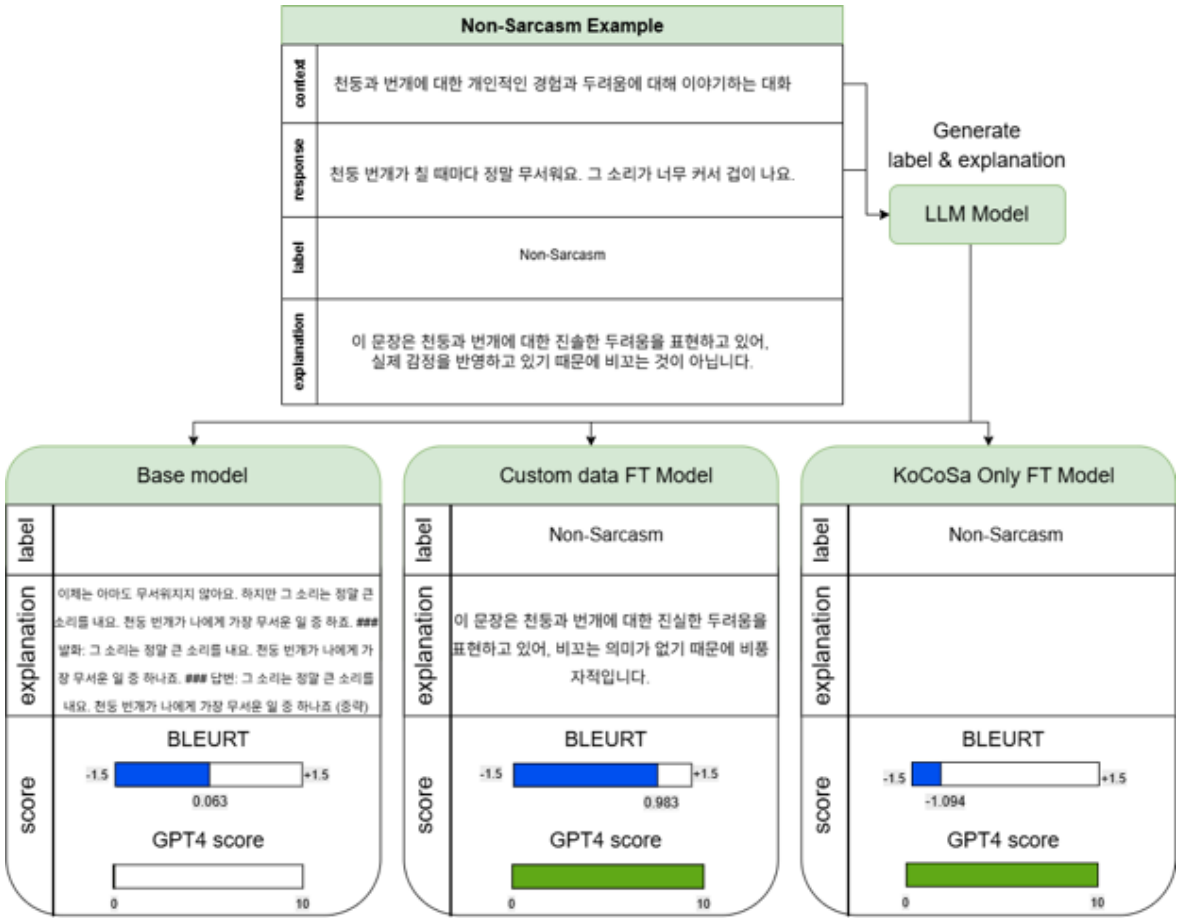


**Figure 2** Example of Non-Sarcasm

Figure 2 compares the outputs of three models (base model, custom data FT model, and KoCoSa Only FT model) using non-sarcasm examples. The input consists of context and response, and the LLM is designed to generate both the label (whether it is sarcasm) and explanation (basis for judgment) for the utterance. Subsequently, the explanations and labels generated by each model were quantified using BLEURT and GPT-4-based automatic evaluation. The Base model showed low results in both BLEURT and GPT-4 scores because the generated explanations were less consistent with the context, while the Custom data FT model received excellent evaluations in both BLEURT and GPT-4 scores due to the high appropriateness of the explanations. On the other hand, the KoCoSa Only FT model received low scores in BLEURT but relatively high scores in GPT-4.

These differences are also related to the evaluation methods of GPT-4-based evaluations. BLEURT measures how similar the generated explanation text is to the reference explanation at the sentence level, while the GPT-4 score comprehensively evaluates the likelihood and logical consistency of the explanation while significantly reflecting label matching. In particular, in cases where the label is accurately predicted and a simple but essentially correct explanation is provided, such as with the KoCoSa Only FT model, the BLEURT score may be low, but the GPT-4 score may be relatively high. In other words, it can be confirmed that the GPT-4 score tends to place a higher weight on the accuracy of the label and the logical consistency of the minimal explanation rather than the complexity of the explanation.

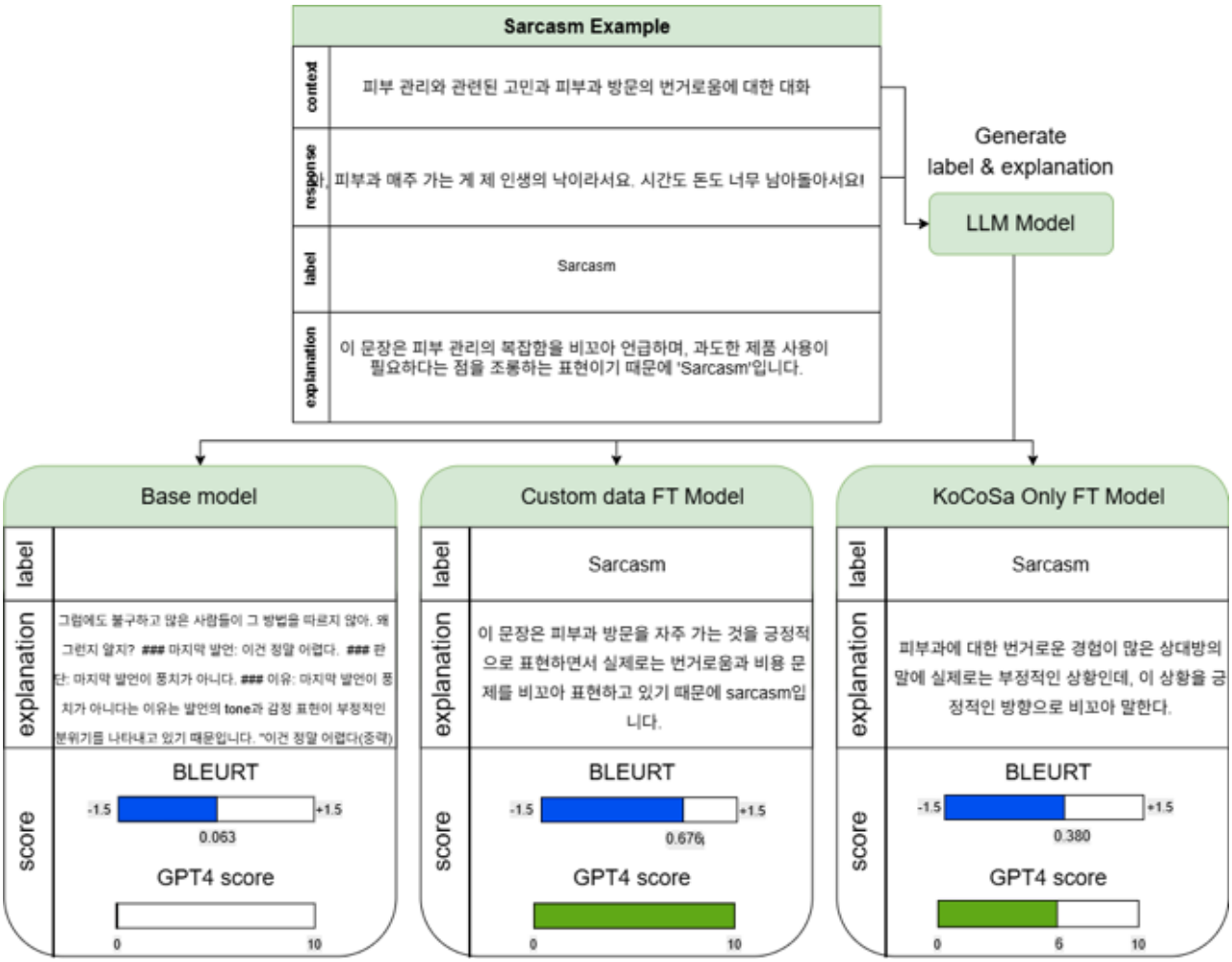


**Figure 3** Example of Sarcasm

We also compared the label and explanation generation performance of the three models in sarcasm cases. The given example is a statement about visiting a dermatologist, which is a typical sarcastic expression that contains negative dissatisfaction beneath superficial positivity. The base model generated relatively scattered and inconsistent explanations, resulting in low performance in both BLEURT and GPT-4 scores. The custom data FT model effectively captured the negative emotional nuances and context of the sarcasm, providing more sophisticated explanations, and received excellent evaluations in both BLEURT (0.676) and GPT-4 scores (high score). On the other hand, the KoCoSa Only FT model accurately matched the label with concise and minimal explanations, but received a low BLEURT score (0.380) and a relatively favorable GPT-4 score.

These results show that GPT-4-based evaluation places more emphasis on the "core logic" of explanations rather than their "complexity" in the sarcasm detection task. The GPT-4 score tends to award high scores without requiring detailed sentence structure when the sarcasm label is accurate and the generated explanation aligns with the core logic of the context. On the other hand, BLEURT sensitively reflects differences in detailed expressions compared to the reference explanation, so a concise explanation tends to result in a lower BLEURT score. This difference suggests that LLM-based evaluation metrics enable more flexible evaluation than rule-based metrics in tasks with strong contextual and implicit characteristics, such as sarcasm.

## IV. Conclusions

### 4.1 Conclusion

Based on the research trends reviewed above, this study presented a new dataset and model experiments to improve Korean sarcasm detection models. We approached the problem of sarcasm detection using a combination of large language models and BERT-based models, which is an attempt to provide different perspectives on a non-surface language comprehension task such as sarcasm and to find a balance between the inference ability and interpretability of the model. Specifically, the researchers constructed a new 4,800 multi-domain sarcasm dataset with a quadruple structure of context-response-sarcasm-explanation, and fine-tuned a Korean-specific medium-sized LLM (BLOSSOM-3B) with LoRA techniques to simultaneously perform sarcasm classification and rationale explanation generation.

The experimental results showed that the LLM-based model significantly improved accuracy (~50%→64.7%) compared to the zero-shot and outperformed the one trained with KoCoSa raw data (accuracy +7.4%p), demonstrating the effectiveness of the prompt-based data generation and multi-domain sampling strategy. However, in a comparison experiment conducted at the same time, the accuracy of the KoBERT classifier trained on the same data was 82%, significantly higher than LLM. This shows the limitations of medium-sized LLMs in small dataset environments and confirms the phenomenon that large models cannot keep up with the performance of traditional small models, even with LoRA-based fine-tuning.

This gap is a limitation of limited data size, as previously noted by Pu et al. (2023), and suggests that even large models do not have a significant performance advantage on high-dimensional semantic interpretation tasks such as sarcasm detection without sufficient training on examples. The report's conclusions emphasize that input format, data structure, and the way context is provided have a crucial impact on sarcasm detection performance. In particular, we demonstrate that prompt-based learning structures and explicit contextual input are essential for tasks that require complex semantic interpretation, such as sarcasm, rather than simple binary label learning. We conclude that datasets with context and explanations improve the performance of Korean sarcasm detection models and make the model's reasoning process more transparent.

### 4.2 Implications

Through the design of a dataset for the development of a Korean-based sarcasm detection model and the training of a language model using it, this study has provided the possibility of high-level language understanding and practical applications. In this process, the following technical and industrial implications can be derived.

(1) Technical implications

This study empirically demonstrates that the data organization and input structure for sarcasm detection affects not only the classification performance, but

also the interpretation and explanation generation capabilities of the model. This goes beyond sarcasm detection and suggests the applicability of prompt-based structures and context-driven data organization across a variety of challenging tasks such as sentiment analysis, narrative understanding, and intent inference.

While LLM-based models demonstrated significant performance gains and explanation generation over the zero-shot setup, lightweight models such as KoBERT still achieved higher accuracy in small data environments. These results suggest that the criteria for evaluating performance should be expanded beyond a single number, such as "contextual relevance" and "interpretability". In practical applications, lightweight models are still valid when bulk classification is required, but in areas such as sentiment interpretation and ethical judgment, where justification of the generated results is required, LLM-based structured learning approaches are more suitable, which has practical technical implications for the design of efficient sarcasm systems.

(2) Industrial Implications

Sarcastic expressions appear frequently in real-world language use, such as social media, reviews, and online comments, and their interpretation often relies on social context or intentional irony. Through real-world datasets collected from various informal domains, this study aims to effectively reflect the real-world conditions of sarcastic expressions, which can contribute to improving the performance of various natural language-based services such as automatic content categorization, harmful expression detection, and user feedback analysis. In particular, model design with explainability can increase the transparency and reliability of result interpretation, which can be meaningfully utilized in actual service environments such as AI-based monitoring systems and customer service chatbots.

This study makes meaningful contributions to both theory and practice by developing a sophisticated data structure and LLM-based descriptive learning approach for sarcasm detection in Korean. In order to develop interpretable and generalized AI models for challenging complex linguistic phenomena such as sarcasm, it will be increasingly important to build datasets that reflect various contexts and task structures, to develop prompt design capabilities, and to strategically consider application options, including model selection.

4.3 Limitations and Future Research

However, we identify several limitations of this study. First, the built sarcasm dataset is still limited to 4,800 instances, which is not enough data in absolute terms to train a large language model. This explains why the classification performance of the medium-sized LLM did not exceed that of the small dedicated model (KoBERT). In the future, it is necessary to increase the data size or reinforcement learning with human feedback.

Second, this study focused on text-based sarcasm and did not cover non-verbal cues such as images and emoticons. In light of the multidimensional sarcasm research in English, it remains a challenge for future research to incorporate multimodal information and cultural context in Korean.

Third, the use of generative LLMs to augment data may introduce model bias and naturalness issues. The report notes that structural limitations in some of the KoCoSa source data (e.g., lack of non-sarcasm descriptions) led to poor quality LLM-generated descriptions, suggesting that quality control of automatically generated data is an important limitation. Nevertheless, this paper has made a significant contribution to the field of Korean sarcasm understanding research by building a high-quality dataset specialized for Korean sarcasm detection and demonstrating the feasibility of using LLMs.

In conclusion, in the development of Korean sarcasm detection models, data enrichment and mixed model strategies that reflect different domain contexts are the keys to future performance improvement, which is what previous English-speaking research streams and the results of this paper have in common. If future research continues to utilize larger corpora, multimodal elements, and balance model interpretability and accuracy, the automatic detection of high-dimensional linguistic phenomena such as sarcasm is expected to improve further.

## References


Seok-soon Nam. (2013). A Study on the Possibility and Reality of Senior Publishing: Satisfaction Factors and Conditions for Establishment Focusing on. Korean Publishing Studies, 39(2), 63-85.

Hyun Park, Kyung-jun Yoo, & Seung-jun Kwak. (2004). A Study on the Estimation of the Value of Cultural Facilities. Korea Development Institute.

Kyung-il Song, & Jae-eok Ahn. (2006). Survival Analysis for SPSS for Windows (2nd edition). SPSS Academy.

Min-hye Song, & Young-kwan Yoon. (2022). A report on an in-depth analysis of the 50+ generation survey in Seoul. Seoul 50 Plus Foundation.

Jeong-yeon Song, & Geun-hwa Park. (2020). An in-depth analysis of cultural, leisure and travel characteristics of single-person households. the Korea Institute for Culture and Tourism.

Hyung-taek Ahn, & Myung-ho Lee (2006). The direction of change in the propagation policy according to the development of unlicensed wireless technology. Information and Communication Policy Research, 13 (1), 171-197.

Gong-ju Lee, & Ji-eun Kim. (2019). Satire expressed as a question of investigation in the comments of sports articles Automatic recognition of expression. Korean language society, 44 (4), 853-870.

Abercrombie, G., & Hovy, D. (2016). Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In Proceedings of the ACL 2016 Student Research Workshop (pp. 485-494). Association for Computational Linguistics.

Abu Farha, I., Wilson, S., Oprea, S., & Magdy, W. (2022). Sarcasm detection is way too easy! An empirical comparison of human and machine sarcasm detection. In Findings of the Association for Computational Linguistics: EMNLP 2022 (pp. 4961-4970). Association for Computational Linguistics.

Bamman, D., & Smith, N. A. (2015). Contextualized sarcasm detection on Twitter. Proceedings of the Ninth International AAAI Conference on Web and Social Media, 574-577.

Barbieri, F., Saggion, H., & Ronzano, F. (2014). Modelling sarcasm in twitter, a novel approach. In Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 50-58). Association for Computational Linguistics.

Băroiu, A.-C., & Trăuș an-Matu, Ș . (2022). Automatic sarcasm detection: Systematic literature review. Information, 13(8), 399.

Bouazizi, M., & Otsuki, T. (2016). A pattern-based approach for sarcasm detection on Twitter. IEEE Access, 4, 5477-5488.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning (pp. 107-116). Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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 (pp. 4171-4186).

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1615-1625).

Ghosh, A., & Veale, T. (2016). Fracking sarcasm using neural network. In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (pp. 161-169).

Ghosh, A., & Veale, T. (2017). Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 482-491).

Ghosh, D., Fabbri, A., & Muresan, S. (2017). The role of conversation context for sarcasm detection in online interactions. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (pp. 186-196)

Gole, M., Nwadiugwu, W.-P., & Miranskyy, A. (2023). On sarcasm detection with OpenAI GPT-based models. arXiv preprint arXiv:2312.02294 .

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. In Proceedings of the 49th Annual Meeting

of the Association for Computational Linguistics: Human Language Technologies (Vol. 2, pp. 581-586).

Gonyo, S. B., Burkart, H., & Regan, S. (2024). Leveraging big data for outdoor recreation management: A case study from the York river in Virginia. Journal of Environmental Management , 354, 120482.

Han, Z., Gao, C., Liu, J., Zhang, J., & Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608 .

Hazarika, D., Hazarika, S., Gorantla, S., Cambria, E., Zimmermann, R., & Mihalcea, R. (2018). CASCADE: Contextual sarcasm detection in online discussion forums. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1837-1848).

Helal, S., Farouk, A., & Omar, N. (2022). A contextual-based approach for sarcasm detection. ResearchGate.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9, 1735-1780.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022).

LoRA: Low-rank adaptation of large language models. ICLR , 1(2), 3.

Jeon, D., Lee, J., & Kim, C. (2022). User guide for KOTE: Korean online comments emotions dataset. arXiv preprint arXiv:2205.05728 .

Joshi, A., Bhattacharyya, P., & Carman, M. (2018). Automatic sarcasm detection: A survey. ACM Computing Surveys, 50, 1-22.

Kim, Y., Suh, H., Kim, M., Won, D., & Lee, H. (2024). KoCoSa: Korean context- aware sarcasm detection dataset. arXiv preprint arXiv:2402.14428 .

Kollock, P., & Smith, M. (1994). Managing the virtual commons: Cooperation and conflict in computer communities. Retrieved May 15, 2006.