KoCoSa: Korean Context-aware Sarcasm Detection Dataset

Yumin Kim[*1], **Heejae Suh**[*2], **Mingi Kim**[*3], **Dongyeon Won**[*2], **Hwanhee Lee**[†3]
[1]Dept. of Applied Statistics [2]Dept. of Computer Science and Engineering
[3]Dept. of Artificial Intelligence, Chung-Ang University
{kimym7801, linkyouhj, mingi8233, dnjsehddus99, hwanheelee}@cau.ac.kr

Abstract

Sarcasm is a way of verbal irony where someone says the opposite of what they mean, often to ridicule a person, situation, or idea. It is often difficult to detect sarcasm in the dialogue since detecting sarcasm should reflect the context (i.e., dialogue history). In this paper, we introduce a new dataset for the Korean dialogue sarcasm detection task, KoCoSa(Korean Context-aware Sarcasm Detection Dataset), which consists of 12.8K daily Korean dialogues and the labels for this task on the last response. To build the dataset, we propose an efficient sarcasm detection dataset generation pipeline: 1) generating new sarcastic dialogues from source dialogues with large language models, 2) automatic and manual filtering of abnormal and toxic dialogues, and 3) human annotation for the sarcasm detection task. We also provide a simple but effective baseline for the Korean sarcasm detection task trained on our dataset. Experimental results on the dataset show that our baseline system outperforms strong baselines like large language models, such as GPT-3.5, in the Korean sarcasm detection task. We show that the sarcasm detection task relies deeply on the existence of a sufficient context. We will release the dataset at https://anonymous.4open.science/r/KoCoSa-2372.

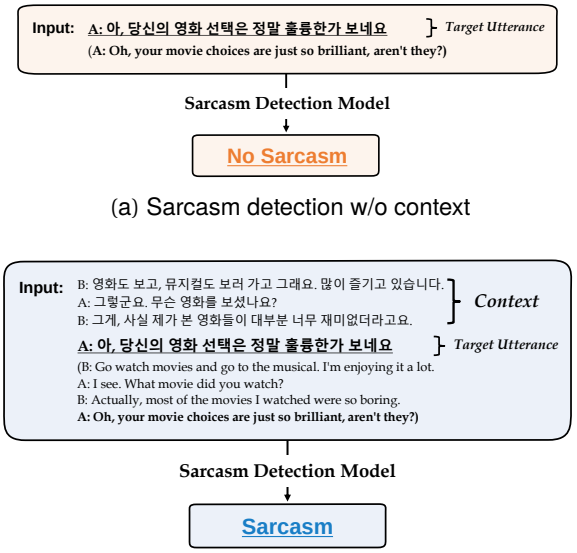
Keywords: Context-aware Sarcasm Detection, Dataset Construction, Korean Dataset

1. Introduction

Sarcasm is a form of verbal irony characterized by saying something contrary to the text's literal meaning, which is widely used in everyday dialogues to humorously criticize a specific situation or object (Filik et al., 2016). When developing a dialogue system, misunderstanding this sarcasm may lead to fatal errors (Riloff et al., 2013). Hence, to prevent such errors, it is often important to develop a sarcasm detection system.

Sarcasm detection poses a different challenge compared to general sentiment analysis tasks, primarily due to its sensitivity to the presence or absence of context (Ghosh et al., 2017; Avvaru et al., 2020). For example, as demonstrated in Figure 1a, even humans may find it challenging to determine whether the target response in the dialogue is sarcasm when the context is absent. But when the dialogue context is provided, as illustrated in Figure 1b, we can easily understand that the last response is sarcasm used to tease A for not choosing a good movie. As in this example, context is particularly significant in sarcasm detection tasks (Bamman and Smith, 2015). Therefore, it is necessary to develop a context-aware sarcasm detection system that can utilize dialog history.

Meanwhile, a majority of datasets utilized for training sarcasm detection systems originate from English Twitter and Reddit comments (Moores and Mago, 2022). However, these datasets inherently diverge from daily dialogues that predominantly involve interactions among acquaintances, making it difficult to use the sarcasm detection system trained on these datasets for daily dialogues. In addition, sarcasm also reflects cultural differences that exhibit the nuances of each country's culture. When annotated by annotators from different countries, these differences result in a degree of inconsistency. (Liu et al., 2014; Joshi et al., 2016) Therefore, relying solely on *translationese* when



(a) Sarcasm detection w/o context

(b) Sarcasm detection w/ context

Figure 1: Examples on Korean sarcasm detection results for a target utterance, (a) without the context and (b) with context, respectively.

[*]Equal Contribution.
[†]Corresponding author.

investigating sarcasm detection methods in various languages is not appropriate, as the *translationese* may not capture the linguistic nuances of sarcasm. As a result, there is an imperative need for monolingual datasets, especially considering that research on English sarcasm detection has been notably more extensive compared to other languages (Rahma et al., 2023).

In this paper, we introduce a new Korean context-aware sarcasm detection dataset composed of 12.8k daily dialogues. We construct the dataset through a comprehensive and effective data generation pipeline facilitated by both LLMs and human revision. We utilize two kinds of Korean datasets, which comprise daily Korean dialogues and message exchanges, adequately representing the social and cultural background. Using these source datasets, we first extract situations from the dialogues and identify the level of politeness of each utterance. By integrating this contextual information into LLMs as prompts, we generate new dialogues that include both sarcastic responses and corresponding explanations. Then, we filter harmful and offensive content from the dataset. Finally, expert annotators review the correctness of labels and remove abnormal dialogues to improve the dataset's overall quality further.

We conduct a series of experiments with our dataset to benchmark various baseline systems, including the strong baseline systems trained with our dataset and several Korean LLMs, for the Korean context-aware sarcasm detection task. Notably, the pre-trained Korean language model KLUE-RoBERTa (Park et al., 2021) fine-tuned with our dataset significantly outperforms GPT-3.5 and achieves performance nearly equivalent to GPT-4 (OpenAI, 2023). Also, we reveal a significant drop in task performance when context information was omitted, underscoring the importance of considering contextual cues for accurate sarcasm detection. However, it was evident that the model's performance fell behind human capabilities, emphasizing the continued need for research into sarcasm detection methodologies. The main contributions of the paper can be summarized as follows.

- We propose a comprehensive dataset generation pipeline for the context-aware sarcasm detection task using LLMs and human revision.
- We introduce a new large-scale Korean Context-aware Sarcasm detection dataset (KoCoSa) through the proposed pipeline, which is composed of 12.8k daily dialogues.
- We provide a decent analysis of the Korean context-aware sarcasm detection task through this dataset, including the strong baseline system for the task.

2. Related Work

2.1. Sarcasm Detection Dataset

Numerous English datasets have been released for sarcasm detection (Bamman and Smith, 2015; Rajadesingan et al., 2015; Wallace et al., 2015; Hazarika et al., 2018), with many of them incorporating context as a crucial factor in determining whether an utterance is sarcastic or not. Meanwhile, datasets for non-English languages are less abundant than in English, but some are available. Recently, several datasets and research endeavors related to Arabic have been pursued (Farha and Magdy, 2020; Elgabry et al., 2021; Faraj and Abdullah, 2021; Farha and Magdy, 2021). Also, few research and datasets for Chinese and Czech have been conducted (Ptáček et al., 2014; Lin and Hsieh, 2016; Gong et al., 2020; Xiang et al., 2020). To the best of our knowledge, in the case of Korean, *Kocasm* Kim and Cho (2019) is the only dataset for sarcasm detection. However, this dataset lacks the necessary context, highlighting the need for additional Korean datasets. In our work, we first introduce a Korean sarcasm dataset enriched with comprehensive context.

2.2. Context-aware Sarcasm Detection

Due to its linguistic nature, many previous studies related to context-aware sarcasm detection have been conducted. Multiple studies have affirmed that the efficacy of sarcasm detection significantly improves with the consideration of context (Bamman and Smith, 2015; Ghosh and Veale, 2017; Poria et al., 2016; Ghosh et al., 2018). Similarly, Baruah et al. (2020) emphasized that incorporating both the last utterance in the context and the target response proved to be the most effective approach in enhancing sarcasm detection performance. In addition, Dong et al. (2020) notes that considering relative context improved sarcasm detection performance. These findings collectively emphasize the crucial role of context-aware approaches in advancing the field of sarcasm detection. In this work, we aim to develop a context-aware sarcasm detection system, especially in Korean daily dialogues, by incorporating contextual information.

2.3. Data Augmentation Leveraging LLMs

Previous research has established the efficacy of large pre-trained language models(LLMs) like GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023) in data augmentation. Similarly, utilizing LLMs for data labeling and annotation has proven to be

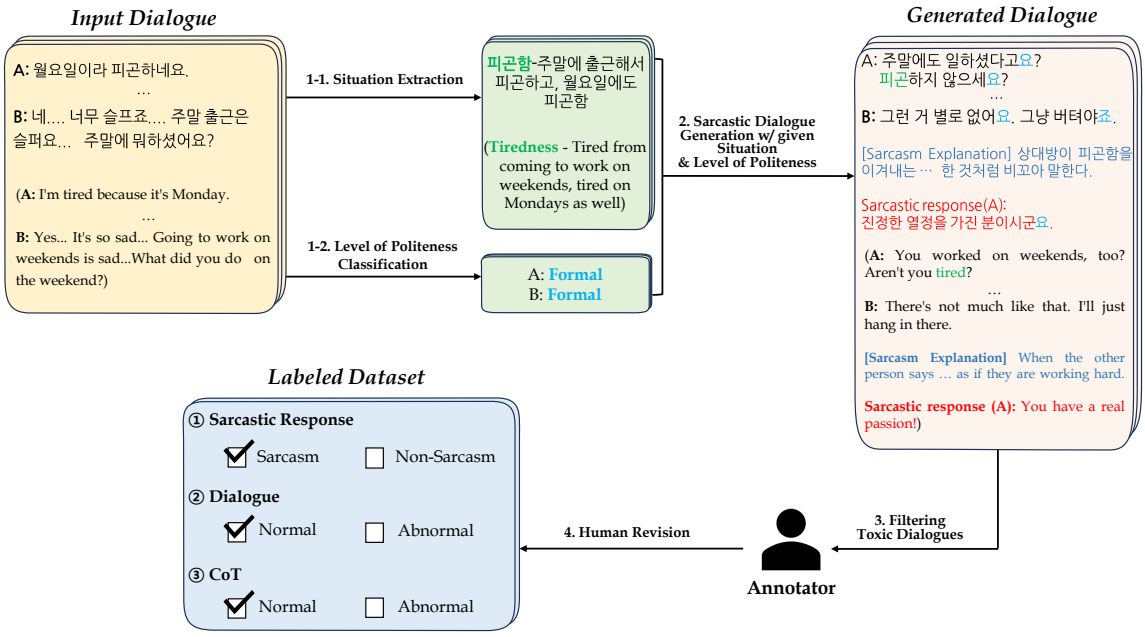


Figure 2: The overall pipeline of KoCoSa dataset construction. In the *Generated Dialogue* example, *light blue letters* represent the honorific ending of a word in Korean. This figure is best viewed in color.

highly effective in terms of both cost and time efficiency, as demonstrated by (Wang et al., 2021) and (He et al., 2023). Furthermore, in terms of Zero-Shot augmentation, Kumar et al. (2020) have argued that transformer-based pre-trained language models excel in data augmentation. Recent research has witnessed successful endeavors in data augmentation and labeling by employing LLMs with zero-shot or few-shot prompting (Bonifacio et al., 2022; Dai et al., 2022; Ubani et al., 2023). In this paper, we present a comprehensive pipeline that leverages LLMs for sarcasm detection, providing a holistic approach to data augmentation.

3. Dataset Construction

In this section, we describe a comprehensive construction process of our dataset as illustrated in Figure 2. Our proposed approach involves harnessing large language models, GPT-4 and GPT-3.5 (OpenAI, 2023), to generate sarcastic dialogues from source dialogues. However, we observe that simply instructing LLM to generate sarcastic dialogue has critical problems. First, we demonstrate that asking LLM to simply create a new dialogue which includes the sarcastic response for the last utterance, results in sarcastic dialogues with almost similar theme. Similarly, we find that adding only a sarcastic response at the end of an existing dialogue with LLM often generates an unnatural sarcastic utterance that doesn't match the context. Since sarcasm needs to be provided with an appropriate situation, adding sarcastic utterances to a random dialogue mostly creates an abnormal response. To solve these problems, we propose a new pipeline for generating sarcastic dialogue using LLMs from the existing dialogue as shown in Figure 2. Rather than simply creating the dialogue from scratch, we construct diverse and natural dialogues that contain sarcastic responses in the last utterance by utilizing the situation extracted from the source dialog. (§3.2) After automatic filtering on undesired samples (§3.3), we ask human annotators to review and label these dialogues to construct a high-quality sarcasm detection dataset. (§3.4)

3.1. Collecting Source Dialogue

We use two primary Korean source dialogue corpora to construct KoCoSa: NIKL Messenger Corpus (National Institute of Korean Language, 2022a) (24.4%) and NIKL Online Text Message Corpus (National Institute of Korean Language, 2022b) (75.6%). The data formats of both corpora are nearly identical, with very minor differences. The subject matter of this source dialogue corpora pertains to daily dialogue. We find that in cases involving three or more participants, the conversation lacks consistency due to the unordered nature of utterance sequences in online messengers. Hence, we only use the dialogues that are made between two people for the dataset.

Input Prompt

You are Korean. You create natural Korean dialogues proficiently. Please consider the level of politeness.
Sarcasm: someone says something but means the opposite in a mocking or ironic way, often using tone and context to convey the real meaning.
Task Description: Create a completely new Korean dialogue related to the provided summary. Then, generate a sarcastic sentence in response to the final utterance of the dialogue. Provide an explanation of how to respond sarcastically to the generated dialogue. Then, write a sarcastic response(about 10 to 15 words) without any additional context.

Example 1. Situation: 저녁 메뉴-계란 프라이를 태워 먹지 못하는 상황(*Dinner menu - Couldn't eat because of burnt fried eggs*)
Level of politeness: A-반말(Informal), B-반말(Informal))
A: 요리는 잘 돼가? (*How's the cooking going?*)
...
B: 계란 후라이가 조금 탔어.(*The fried eggs are a little burnt.*)
Sarcasm Explanation: 계란프라이가 바싹 타버렸다는 마지막 A의 말에 실제로는 부정적인 상황인데, 이 상황을 긍정적인 방향으로 비꼬아 말한다. (*It's actually a negative situation when A said that the fried egg was burnt out, but A sarcastically calls this situation in a positive direction.*)
Sarcastic response(A): 이거 정말 바삭바삭하겠는걸.(*It's going to be really crispy.*)

Example 2. Situation: {situation},
Level of politeness: A-{h_a}, B-{h_b}
A:

Table 1: Prompt used to generate sarcastic dialogue D' in section 3.2. The words colored blue respectively indicates S, h_a, and h_b, extracted from D.

3.2. Generating Sarcastic Dialogues

We provide input dialogue D from the source dataset to GPT-3.5 to extract the situation S (§3.2.1). Then, we ask GPT-4 to generate a new dialogue D' based on S extracted from the previous step to include a sarcastic response R (§3.2.2).

3.2.1. Situation Extraction

We first extract the situation of the given input dialogue D, utilizing GPT-3.5 to extract S in a fixed format($S = \{t; s\}$, by two-shot in-context learning, t and s each representing the main theme and brief summary) as shown in Figure 2. We observe that if a dialogue is generated from t only, the issue of diversity arises. Similarly, when generating content solely based on s, there is a tendency to overly focus on the summary's contents, leading to an increase in abnormalities as the dialogue's subject rapidly shifts. For these reasons, we utilize both t and s for a new dialogue generation. We truncate a total of a maximum of 6 turns as input dialogue D from the source corpus to mitigate multiple distinct situations that may exhibit abrupt transitions in the dialogue.

3.2.2. Generating New Dialogue from Situation

We create a new dialogue D' using GPT-4 based on the situation generated from the previous step. As depicted in Table 1, we configure the sample of the input prompt to create a dialogue with an extracted situation. We fix the level of politeness(h_a,h_b, each indicating the level of politeness of speaker A and B) of each speaker as *Informal* or *Formal* by examining whether the input dialogue was spoken in level of politeness.

Consistent level of politeness. In Korean, ensuring a consistent level of politeness is of utmost importance. Hence, when a single speaker switches between formal and informal language, it results in a less natural dialogue flow. To resolve the issue, we aim to maintain the level of politeness of each speaker in original dialogue by obtaining the formality of the speaker from a pre-trained formal classifier.[1]. We incorporate this formal information into the input prompt to create a dialogue that preserves the formality of speakers in each utterance.

Generating Dialogues with LLMs. Using the extracted situation S and level of politeness h_a and h_b, we generate a new dialogue D' that contains a sarcastic response in the last utterance using GPT-4. We employ the Chain-of-Thought(CoT) prompting (Wei et al., 2022) in conjunction with a two-shot generation approach. As a way of CoT prompting, we instruct GPT-4 to write *Sarcasm Explanation* that explains why the last utterance would be sarcastic considering the context before creating a sarcastic response. Afterward, a sarcastic response is generated that fits the explanation. We present the full prompt in Table 1.

3.3. Data Filtering

To improve the quality and moderativity of the dataset, we automatically filter toxic or abnormal dialogues in advance to the annotation process.

[1] https://huggingface.co/j5ng/kcbert-formal-classifier

	Generated	Label
Context	A: 퇴근이 왜 이렇게 늦어지는 거야?(*Why is leaving work so late?*) B: 너도 늦게 퇴근했나 봐. 나랑 같이 저녁 먹을래?(*I guess you left work late, too. Would you like to have dinner with me?*) A: 음, 그래. 뭐 먹을까?(*Well, yeah. What should we eat?*) B: 오늘 피곤하니까 그냥 편의점에서 라면 사 와서 끓여먹자(*I'm tired today, so let's just buy ramen from the convenience store and eat it.*)	**Normal**
Response	A: 그래, 우리 건강에 정말 좋겠다! (*Yeah, it must be great for our health!*)	**Sarcasm**
Sarcasm Explanation	B의 마지막 대화에서 건강에 좋지 않은 라면을 추천했기 때문에, A는 이를 아이러니하게 비꼬아 말한다.(*Because '**B**' recommended unhealthy ramen in his last utterance, '**A**' ironically sarcastically says this.*)	**Normal**

Table 2: An example of the generated dataset with annotation results. We deprecate the contexts labeled as abnormal in the released version.

3.3.1. Filtering Toxic Dialogue

Sarcastic utterances occasionally take the form of offensive language, so it is necessary to filter out data that may cause potential harmfulness when the dataset is disclosed. We guarantee the dataset to be moderate through a total of two steps. First, We apply the moderation API of OpenAI[2] to all data. Among 17,073 samples, we removed 23(0.0013%) dialogues that were labeled as "improper." Then, we additionally filtered five samples that included swear words through the manual data inspection process. For swear word filtering, we use a pre-defined *frequently used toxic words* list.

3.3.2. Filtering Abnormal Dialogue

To ensure dataset quality, a final step involves applying automatic abnormal dialogue detection by instructing GPT-3.5 to check if the given dialogue is incongruent. Any data identified as abnormal through detection is discarded after manual review.

3.4. Human Annotation

We describe the human annotation process for the generated dialogue in this section. We deliberately select exceptional annotators to ensure the attainment of high-quality annotations. Furthermore, we furnish comprehensive annotation guidelines, encompassing detailed explanations and a few illustrative samples.

3.4.1. Annotator Selection

We utilize a portion of the data as a preliminary survey for selecting proficient Korean native annotators. In the survey, all participants (30 people) conducted annotations on the same data. We choose 10 annotators by prioritizing those with a high degree (over 65%) of agreement with the majority vote, thus ensuring the quality of annotations.

3.4.2. Annotation Guideline

We offer a comprehensive guideline that encompasses a detailed definition of sarcasm, along with relevant examples for various cases to annotators. We provide full dialogue including *Sarcasm Explanation* for the annotation process. We ask annotators to choose three distinct types of labels for each dialog: *Sarcasm Detection*, *Context Abnormality Detection*, and *Sarcasm Explanation Revision* as follows.

- **Sarcasm Detection** (*Sarcasm / Non-Sarcasm / Abnormal*): Does the last response have sarcastic nuance while congruent with context?
- **Context Abnormality Detection** (*Normal / Abnormal*): Does the provided context seem to be a natural dialogue?
- **Sarcasm Explanation Revision** (*Normal / Abnormal*): Does the explanation offer a suitable description of why the response is sarcasm?

Finally, we only use the dialogues where the labels of both *"Context Abnormality Detection"* and *"Sarcasm Explanation Revision"* are natural to construct the dataset. Among them, we use the data where the label of *Sarcastic Response* is also "Sarcasm" as the "Sarcasm" label in the dataset. For the "Non-sarcasm," we first include the data where the response is *Non-Sarcasm/Abnormal* and the context is *Normal*. Additionally, we use the dialog

[2] https://platform.openai.com/docs/guides/moderation

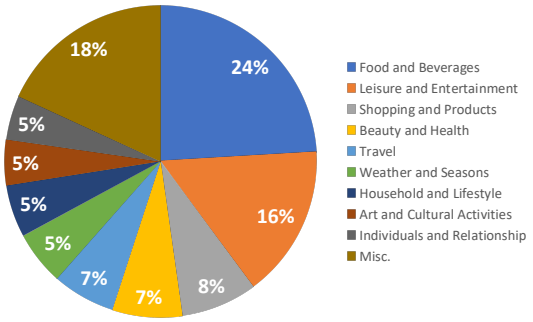


Figure 3: Topic diversity of Online Text Message Corpus and Messenger Corpus. Topics that account for less than 5% are grouped as Misc.

in cases where the context is *Normal* but the response is *Abnormal* by removing the last sarcastic response which is a cause of abnormal. This approach of constructing non-sarcasm data helps address the issue of having data comprised solely of failed attempts at generating sarcasm dialogue.

Total Dialogues	12824
Sarcasm	7608(59.3%)
Non-Sarcasm	5216(40.7%)
Average Turns per Dialogue	4.3
Max Turns	10
Min Turns	2
Tokens per Dialogue	40.3
Tokens per Utterance	9.3
Tokens per Explanation	18.9

Table 3: Overall Statistics of KoCoSa.

4. Dataset Analysis

4.1. Overall Statistics

KoCoSa dataset contains a total of 12,824 dialogues as shown in Table 3. For the last-utterances that are used for the sarcasm detection task, 7,608(59.3%) are labeled as *sarcasm*, and 5,216 dialogues(40.7%) are *Non-sarcasm*. The average number of utterances in the dialogue varies from a minimum of 2 to a maximum of 10 and an average of 4.3. The initial number of data before the human annotation process is 14,788, and the final dataset consists of 12,824, representing only 13.28% data loss due to the abnormality of dialogues. This demonstrates the high production efficiency of our dataset construction pipeline.

Table 4 compares KoCoSa with other sarcasm detection datasets. KoCoSa represents a pioneering contribution as it incorporates both dialog context and explanation of sarcasm. Furthermore, KoCoSa distinguishes itself by drawing its source content from online messenger, in contrast to other datasets that primarily derive from online communities like Reddit and Twitter.

We also investigate the topic diversity of dialogues in KoCoSa by analyzing the topics of each dialogue of the source corpora. As shown in Figure 3, source corpora of KoCoSa contains daily dialogues on a variety of topics. Especially, dialogues related to *food and beverage* and *leisure and entertainment* are the top-2 topics accounting for 40% of the total. Since the dialogues in the KoCoSa dataset are created by extracting the situation from these source corpus, the KoCoSa dataset also contains these diverse topics.

4.2. Human Quality Evaluation

Although our intention is to create natural dialogues that contain sarcastic responses in the dataset, the resulting dataset unexpectedly contains instances labeled as *Non-sarcasm* and *Abnormal*. We conduct a comprehensive analysis of each case and categorize them accordingly.

4.2.1. Label Change

We aim to generate dialogues to include sarcastic responses with the proposed pipeline. However, after the human annotation process, we encounter a substantial amount of instances labeled as *Non-Sarcasm*. We systematically categorize them into two distinct types: *Casual Dialogue*, representing regular dialogues with typical responses, and *Direct Criticism*, denoting openly critical statements without any attempt to hide their intent. Among these instances, we notice that *Direct Criticism* is notably shaped by particular word choices. For instance, in the example provided in Table 5, the word '*extravagant*' straightforwardly conveys a critical tone, resulting in its labeling as Non-Sarcasm.

4.2.2. Abnormal Case Study

We created a new dialogue using LLMs instead of simply modifying the dialogue of the existing dataset to construct KoCoSa. We found that a considerable proportion of these LLM-generated dialogues are labeled as abnormal after the human annotation. Hence, we manually analyzed the types of these abnormal dialogues within a subset of KoCoSa. These abnormal cases can be broadly categorized into two distinct types:

- **Contextual Awkardness**: The unnatural choice of words or context either semantically, morphologically, or both.
- **Format Misalignment**: Cases that do not adhere to the dataset format such as the absence of dialogue, usage of English text, or generation of two or more dialogues, etc.

As demonstrated in Table 6, the majority of cases labeled as abnormal are caused by contextual awk-

Dataset	Language	Source	Sarcastic	Total	Context	Explanation
Muresan et al. (2016)	English	Twitter	0.9K	2.7K	N	N
Oraby et al. (2016)	English	Internet Argument	3.3K	6.5K	N	N
Peled and Reichart (2017)	English	Twitter	3K	3K	N	N
SARC (Khodak et al., 2018)	English	Reddit	1.34M	533M	Y	N
Kocasm (Kim and Cho, 2019)	Korean	Twitter	4.7k	9.3K	N	N
Ghosh et al. (2020)	English	Twitter	3.1K	6.2K	Y	N
Ghosh et al. (2020)	English	Reddit	3.4K	6.8K	Y	N
iSarcasm (Oprea and Magdy, 2020)	English	Twitter	0.8K	4.5K	N	N
Gong et al. (2020)	Chinese	News Comment	2.5k	91.8K	Y	N
ArSarcasm-v2 (Abu Farha et al., 2021)	Arabic	Twitter	3.0K	15.5K	N	N
Misra and Arora (2023)	English	News Headline	13.6K	28.6K	N	N
KoCoSa (Ours)	**Korean**	**Online Message**	**7.6K**	**12.8K**	**Y**	**Y**

Table 4: Comparison of sarcasm datasets. The *Explanation* column indicates whether an explanation for the sarcasm is included in the data.

Change Type(%)	Sarcasm ⟶ Non-Sarcasm
Casual Dialogue (80%)	A: 간식 좀 먹을 게 있어? (*Do we have any snacks?*) ... B: 그냥 인터넷으로 주문하면 되지 않을까? (*Why not just order online?*) **A:** 아니야 직접 가서 사는게 훨씬 편하잖아 (*No, going there in person is much more convenient*)
Direct Critism (20%)	A: 이번 주말에 뭐해? (*What are you up to this weekend?*) ... B: 그냥 200만 원 정도 생각하고 있어. (*I'm thinking of around 2 million won, just casually.*) **A:** 와우, 허세 정말 한가득이네. (*Wow, that's quite extravagant*)

Table 5: Sarcasm attempted but labeled as Non-Sarcasm Cases. The ratios next to each type represent the categorization proportions for the random 50 *Non-Sarcasm* dialogues.

wardness. These issues stem from the inherent characteristics of the source dataset, which encompasses a wide range of topics. This diversity poses a significant challenge in conducting topic extraction and identifying a singular, predominant theme. In order to address the issue of topic diversity, we truncate the dialogues in the source dataset into 6 turns. Despite this approach, we discover that certain remnants of the inherent characteristics of the source dataset persisted. We also observe that a small portion of dialogues are abnormal due to format misalignment.

5. Experiment

5.1. Experimental Setup

Baselines We provide sarcasm detection scores on the KoCoSa dataset using two zero/few-shot models, GPT-3.5 and GPT-4. Additionally, we fine-tune the KLUE-RoBERTa$_{BASE}$ and KLUE-RoBERTa$_{LARGE}$, which is a pre-trained Korean language model on the Korean Language Understanding Evaluation (KLUE) dataset (Park et al., 2021), to our dataset for building strong baseline systems. We set the number of epochs as 5 and use a batch size of 16 for fine-tuning both KLUE-RoBERTa$_{BASE}$ and KLUE-RoBERTa$_{LARGE}$. For GPT-3.5/4, we experiment with in-context learning settings by giving 0/4/8 examples.

Abnormality Type	# of Cases (%)
Contextual Awkardness	788(97.04)
Format Misalignment	24(2.96)
Total	**812**

Table 6: Causes of Abnormal Dialogue

Data Preparation We split KoCoSa into Train/Dev/Test in the proportions of about 8:1:1. To mitigate dataset bias, we exchange the name of speakers *A* and *B* for 50% of the dataset, as most of the sarcastic responses' speaker in the dataset is *A*. The input data passed to the models consists of contexts, and sarcastic responses, where the output is the label for the sarcasm detection task. We do not use explanation data for developing sarcasm detection systems in our work, leaving it for future work.

Evaluation Metric We use six metrics considering the imbalanced label distribution in the dataset; Balanced Accuracy, Weighted F1, Precision-[*Sarcasm, Non-Sarcasm*] and Recall-[*Sarcasm, Non-Sarcasm*].

5.2. Results

As reported in Table 7, the baseline scores of models are competitive with the results from the human evaluation, although they fall slightly below the human evaluation score. GPT-3.5 shows about 50% on the Balanced Accuracy representing almost random guess and the lowest *Recall-S* in all of zero-shot (20.5%), 4-shot (20%), and 8-shot learning (6%). Meanwhile, GPT-4 shows the competitive score to Human Evaluation. We build KoCoSa utilizing GPT-4, raising concerns about it being tailored to fit GPT-4's tendencies. However,

Model	Balanced Acc.	Weighted-F1	Precision-S	Recall-S	Precision-N	Recall-N
Zero/Few-shot						
GPT-3.5(zero-shot)	53.5	40.6	71.2	20.5	40.2	86.5
GPT-3.5(4-shot)	51.8	39.4	66.4	20.0	39.2	83.5
GPT-3.5(8-shot)	49.4	27.2	57.7	6.0	37.8	**92.9**
GPT-4(zero-shot)	73.2	72.6	**83.3**	68.8	60.5	77.6
GPT-4(4-shot)	75.0	76.6	80.5	82.3	70.2	67.7
GPT-4(8-shot)	74.5	75.1	81.9	76.3	65.4	72.6
Fine-tuning						
KLUE-RoBERTa$_{BASE}$	74.0($\pm$0.2)	74.7($\pm$0.2)	71.5($\pm$0.2)	**93.4**($\pm$0.5)	**87.2**($\pm$0.7)	54.7($\pm$0.7)
KLUE-RoBERTa$_{LARGE}$	74.9($\pm$0.3)	75.5($\pm$0.3)	74.6($\pm$0.3)	85.0($\pm$0.6)	80.0($\pm$0.6)	64.8($\pm$0.6)
Human Evaluation	**80.2**	**80.3**	83.0	80.5	77.1	79.9

Table 7: Sarcasm detection score on the test set of KoCoSa for various models. Note that KLUE-RoBERTa models are fine-tuned for the training set of KoCoSa. *Precision-S* and *Recall-S* represent scores for the *Sarcasm* label. *Precision-N* and *Recall-N* show scores for the *Non-Sarcasm* label.

Context	Model	Human
Only Response	73.2	62.2
Last 1 Utterance + Response	73.2	-
Last 2 Utterance + Response	74.9	-
Last 3 Utterance + Response	75.8	-
Full Context	**76.0**	**80.2**

Table 8: Balanced accuracy of sarcasm detection among utterance length of the dialogue context using KLUE-RoBERTa$_{LARGE}$.

Topic	Balanced Acc.		Weighted F1	
	KLUE	GPT	KLUE	GPT
Food and Beverages	71.9	73.2	71.7	73.9
Leisure and Entertainment	73.6	72.8	74.3	73.0
Individuals and Relationship	76.1	76.8	76.8	77.2
Beauty and Health	73.5	74.7	74.5	76.2

Table 9: Performance across various topics in the KoCoSa test set, using KLUE-RoBERTa$_{LARGE}$ and GPT-4 4-shot learning.

human revision reveals that about 40% are abnormal or non-sarcasm. Furthermore, GPT-4's detection performance still falls short of human capabilities.

As shown in Table 7, Also, despite the huge gap in the model size, fine-tuning models show competitive performance compared to zero/few-shot models. In particular, KLUE-RoBERTa$_{LARGE}$ with 337M parameters outperforms most of the GPT-3.5 and GPT-4 models except the 4-shot GPT-4 model regarding Balanced Accuracy. The 71.5% and 74.6% on the *Precision-S* of each KLUE-RoBERTa$_{BASE}$ and KLUE-RoBERTa$_{LARGE}$ are lower than 87.2% and 80% on *Precision-N* of each model. Meanwhile, in zero/few-shot models, *Precision-S* is higher than *Precision-N*.

5.3. Context Length Dependency

To demonstrate the importance of context in the sarcasm detection task, we compare scores based on different lengths of utterance turns within the context. Unlike the baseline experiment setup which includes *Full context*, we design an experiment to provide varying amounts of contexts to the model for validating the effectiveness of context length dependency in the sarcasm detection task. Starting from the setup w/o context (*Only Response*), we increase the amount of context to *Full context* to train KLUE-RoBERTa$_{LARGE}$ model and report the results. As shown in Table 8, using the full context obtains the best Balanced Accuracy of 76.0. From the lowest Balanced Accuracy of 73.2 when only the response is used, the performance steadily improves as more context utterances are incorporated. It implies that language models benefit from sufficient contextual information to enhance the accuracy of sarcasm detection.

5.4. Topic Dependency

We evaluate the consistency in the detection performance of each system among the topic of dialogues in the sarcasm detection task, by measuring the performance of KLUE-RoBERTa$_{LARGE}$ and GPT-4 4-shot model on the KoCoSa test set for each topic. Given the similarity between the topics in the source dialogue and those generated in KoCoSa(§4.1), we categorize the KoCoSa using the source dialogue topics.

As depicted in Table 9, we suggest that sarcasm detection performance does not vary with the dialogue's topic. Specifically, dialogues centered around *Individuals and Relationship* reveal

the highest sarcasm detection performance for both KLUE-RoBERTa$_{LARGE}$ and GPT-4. Meanwhile, dialogues about *Food and Beverages* show the lowest Balanced Accuracy 71.9% for KLUE-RoBERTa$_{LARGE}$ and *Leisure and Entertainment* show the lowest Balanced Accuracy 72.8% from GPT-4. Notably, the gaps between the highest and lowest scores are 4.2%p and 5.1%p in both Balanced Accuracy and Weighted F1. Therefore we demonstrate that there is not much difference in performance between topics.

6. Conclusion

In this paper, we propose a new Korean context-aware sarcasm detection dataset KoCoSa composed of 12,824 dialogues. We construct this dataset through a comprehensive pipeline that leverages large language models, automatic dataset filtering, and human annotations. Our dataset construction pipeline represents an efficient framework capable of generating diverse conversational content while maintaining minimal loss. We also provide baseline performance on our dataset including GPT models, and train strong baseline systems using KLUE-RoBERTa models for the dataset. We expect that our proposed dataset will serve as a valuable resource for advancing research in Korean sarcasm detection. Furthermore, we believe that the dataset generation pipeline described in this paper can facilitate researchers in exploring language-specific sarcasm detection in low-resource languages.

7. Ethics Statement

We carefully supervise the process of generating KoCoSa so as not to pose any ethical issues. We initiated the annotation process after obtaining approval from the Institutional Review Board (IRB). Moreover, we ensured the ethical integrity of our data by leveraging OpenAI's Moderation API. Additional human verification was conducted to account for cases where toxic expressions remained in the filtered dialogues. We removed offensive language and inappropriate content, guided by the feedback provided by annotators. Also, we pay data annotators above the minimum wage, ensuring fair and equitable compensation for their work.

Acknowledgement

This research was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program (Chung-Ang University)).

8. Bibliographical References

Adithya Avvaru, Sanath Vobilisetty, and Radhika Mamidi. 2020. Detecting sarcasm in conversation context using transformer-based models. In *Proceedings of the second workshop on figurative language processing*, pages 98–103.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter. In *proceedings of the international AAAI conference on web and social media*, volume 9, pages 574–577.

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Context-aware sarcasm detection using bert. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 83–87.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: few-shot dense retrieval from 8 examples (2022). *arXiv preprint arXiv:2209.11755*.

Xiangjue Dong, Changmao Li, and Jinho D Choi. 2020. Transformer-based context-aware sarcasm detection in conversation threads from social media. *arXiv preprint arXiv:2005.11424*.

Hazem Elgabry, Shimaa Attia, Ahmed Abdel-Rahman, Ahmed Abdel-Ate, and Sandra Girgis. 2021. A contextual word embedding for arabic sarcasm detection with random forests. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 340–344.

Dalya Faraj and Malak Abdullah. 2021. Sarcasmdet at sarcasm detection task 2021 in arabic

using arabert pretrained model. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 345–350.

Ibrahim Abu Farha and Walid Magdy. 2020. From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 21–31.

Ruth Filik, Alexandra Ţurcan, Dominic Thompson, Nicole Harvey, Harriet Davies, and Amelia Turner. 2016. Sarcasm and emoticons: Comprehension and emotional impact. *Quarterly Journal of Experimental Psychology*, 69(11):2130–2146.

Aniruddha Ghosh and Tony Veale. 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*, pages 482–491.

Debanjan Ghosh, Alexander R Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. *arXiv preprint arXiv:1707.06226*.

Xiaochang Gong, Qin Zhao, Jun Zhang, Ruibin Mao, and Ruifeng Xu. 2020. The design and construction of a chinese sarcasm dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5034–5039.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413*.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Shih-Kai Lin and Shu-Kai Hsieh. 2016. Sarcasm detection in chinese using a crowdsourced corpus. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing* (*ROCLING 2016*), pages 299–310.

Peng Liu, Wei Chen, Gaoyan Ou, Tengjiao Wang, Dongqing Yang, and Kai Lei. 2014. Sarcasm detection in social media based on imbalanced classification. In *Web-Age Information Management: 15th International Conference, WAIM 2014, Macau, China, June 16-18, 2014. Proceedings 15*, pages 459–471. Springer.

Bleau Moores and Vijay Mago. 2022. A survey on automated sarcasm detection on twitter. *arXiv preprint arXiv:2202.02516*.

OpenAI. 2023. Gpt-4 technical report.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Lucy Park, Alice Oh, Jung-Woo Ha (NAVER AI Lab), Kyunghyun Cho, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 213–223.

Alaa Rahma, Shahira Shaaban Azab, and Ammar Mohammed. 2023. A comprehensive review on arabic sarcasm detection: Approaches, challenges and future trends. *IEEE Access*.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

Byron C Wallace, Eugene Charniak, et al. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*), pages 1035–1044.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Rong Xiang, Xuefeng Gao, Yunfei Long, Anran Li, Emmanuele Chersoni, Qin Lu, and Chu-Ren Huang. 2020. Ciron: a new benchmark dataset for chinese irony detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5714–5720.

9. Language Resource References

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.

Gong, Xiaochang and Zhao, Qin and Zhang, Jun and Mao, Ruibin and Xu, Ruifeng. 2020. *The design and construction of a Chinese sarcasm dataset*.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (*LREC 2018*), Miyazaki, Japan. European Language Resources Association (ELRA).

Kim, Jiwon and Cho, Won Ik. 2019. *Kocasm: Korean Automatic Sarcasm Detection*. GitHub.

Rishabh Misra and Prahal Arora. 2023. Sarcasm detection using news headlines dataset. *AI Open*, 4:13–18.

Smaranda Muresan, Roberto González-Ibáñez, Debanjan Ghosh, and Nina Wacholder. 2016. Identification of nonliteral language in social media: A case study on sarcasm. *Journal of the Association for Information Science and Technology*, 67:n/a–n/a.

National Institute of Korean Language. 2022a. *NIKL Messenger Corpus* (*v.2.0*).

National Institute of Korean Language. 2022b. *NIKL Online text message Corpus* (*v.1.0*).

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn

Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.

Peled, Lotem and Reichart, Roi. 2017. *Sarcasm SIGN: Interpreting Sarcasm with Sentiment Based Monolingual Machine Translation*. Association for Computational Linguistics. [link].