Effective Intended Sarcasm Detection Using Fine-Tuned Llama 2 Large Language Models

Fachry Dennis Heraldi and Fariska Zakhralativa Ruskanda

September 6, 2024

Effective Intended Sarcasm Detection Using Fine-tuned Llama 2 Large Language Models

[1]Fachry Dennis Heraldi, [2]Fariska Zakhralativa Ruskanda
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
[1]fd.heraldi@gmail.com, [2]fariska.zr@staff.stei.itb.ac.id

Abstract*— Detecting sarcasm in English text is a significant challenge in sentiment analysis due to the discrepancy between implied and explicit meanings. Previous studies using Transformer-based models for intended sarcasm detection show room for improvement, and the development of large language models (LLMs) presents a substantial opportunity to enhance this area. This research leverages the open-source Llama 2 LLM, released by Meta, fine-tuned to develop an effective sarcasm detection model. Our proposed system design generalizes the use of Llama 2 for text classification but is specifically designed for sarcasm detection, sarcasm category classification, and pairwise sarcasm identification. Data from the iSarcasmEval dataset and additional sources, totaling 21,599 samples for sarcasm detection, 3,457 for sarcasm category classification, and 868 for pairwise sarcasm identification, were used. Methods include prompt development, fine-tuning using Parameter Efficient Fine-tuning (PEFT) with Quantized Low Rank Adaptation (QLoRA), and zero-shot approach. Our model demonstrates significant improvements, sarcasm detection model and pairwise sarcasm identification model are surpassing top models on previous study: F1-score of 0.6867 for sarcasm detection, Macro-F1 of 0.1388 for sarcasm category classification, and accuracy of 0.9 for pairwise sarcasm identification. Results demonstrate that Llama 2, combined with external datasets and effective prompt engineering, enhances intended sarcasm detection. The PEFT technique with QLoRA reduces memory requirements without compromising performance, enabling model development on devices with limited computational resources. This research underscores the importance of context and intention in intended sarcasm detection, with dataset labeling discrepancies remaining a significant challenge.

Keywords—sarcasm detection, large language models, Llama 2, fine-tune, prompt engineering

I. Introduction

Sarcasm is a form of speech or writing where the intended meaning differs from the literal one [1]. It is often used to criticize or compliment someone indirectly, such as saying, "You're early!" to someone who is late or "You're such a terrible tennis player!" to a tennis competition winner [2]. Sarcasm often relies on the tone, context, and prior knowledge shared between the speaker and listener, making it difficult for algorithms to accurately interpret. The problem of detecting sarcasm in English text remains a significant challenge in the field of natural language processing (NLP) due to its implicit nature and context dependency. The ambiguity in sarcasm poses a challenge for sentiment analysis systems as it can lead to misinterpretations, impacting system performance [3], [4]. Detecting sarcasm without clear context can result in misclassifying negative sentiment as positive [5], [6]. Researchers have developed new techniques, including dataset improvements and advanced machine learning models such as Transformers, to address this challenge [7]. Effective sarcasm detection can benefit other natural language processing (NLP) applications such as summarization, dialogue systems, and review analysis [5].

Farha et al. introduced iSarcasmEval at SemEval 2022, recent study focusing on intended sarcasm detection in English and Arabic [8]. iSarcasmEval is the pioneering shared task aimed at detecting intended sarcasm, with data sourced and annotated by the authors of the texts themselves. Participants were asked to provide and label their own words as sarcastic or not, thus capturing intended sarcasm. It's important to note that an utterance intended as sarcastic by its author may not be perceived as such by people from different backgrounds. When a tweet is deemed sarcastic by its author, this phenomenon is referred as intended sarcasm [9]. The study highlighted the capabilities and limitations of existing models, providing a benchmark for future advancements, suggested that further improvements could be achieved with deeper machine learning methods and additional datasets [3]. Băroiu & Trăușan-Matu [10] found that fine-tuned GPT-3 outperformed other Transformer-based models in sarcasm detection on multimodal datasets. Despite the progress made in the previous studies, there is still room for improvement in sarcasm detection performance.

Large Language Models (LLMs) have demonstrated exceptional capabilities in understanding and processing complex language structures, making them suitable for advanced NLP tasks [11], [12]. Models such as GPT-3 and its derivatives, have achieved superior performance in various NLP tasks, including sarcasm detection. The ability of LLMs to capture nuanced language features and contextual information makes them ideal candidates for improving sarcasm detection systems. Llama 2, developed by Touvron et al. from Meta [12], is a next generation of autoregressive language model following Llama 1. Llama 2 is available in both base and fine-tuned versions, each with three parameter sizes: 7 billion (7B), 13 billion (13B), and 70 billion (70B). Designed for both commercial and research purposes, Llama 2 is open access. Fine-tuned models are optimized for assistant chat systems, providing more accurate and contextual responses in human-machine interactions. However, the base model can be adapted and fine-tuned for specific NLP tasks, including text classification. Observations indicate that fine-tuning the base model yields promising performance. For instance, using the Massive Multitask Language Understanding (MMLU) benchmark, the fine-tuned model achieved a score of 68.9%, compared to 63.4% for the base model. This demonstrates the significant potential of fine-tuning Llama 2 for specific tasks such as sarcasm detection.

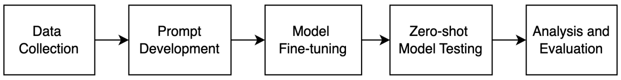

Fig. 1. Proposed method general flow

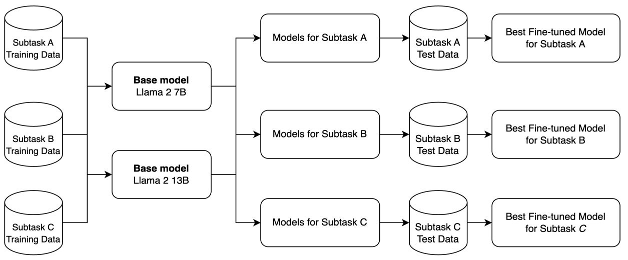

Fig. 2. Fine-tuned model development flow for each subtask

This research builds upon these insights by proposing the use of the Llama 2 model, fine-tuned with additional external datasets to enhance its sarcasm detection capabilities. Considering the challenges and opportunities identified, this research aims to address key questions in the field of sarcasm detection. Our primary objective is to build and evaluate a fine-tuned Llama 2 model to improve sarcasm detection, sarcasm category classification, and pairwise sarcasm identification in English text. Furthermore, the research proves the impact of incorporating external datasets during the fine-tuning process on the model's performance. By leveraging Llama 2 and diverse datasets, this research aims to develop a sarcasm detection model that surpasses the performance of best models in previous study on intended sarcasm detection.

II. Related Works

Sarcasm detection has drawn considerable attention in the field of NLP due to its complexity and the challenge it presents for sentiment analysis systems. Early research in this domain focused on traditional machine learning techniques such as Support Vector Machines (SVM) and Feedforward Neural Networks (FFNN) to classify sarcasm. However, these methods struggled with the inherent ambiguity and context-dependency of sarcastic statements. The introduction of BERT (Bidirectional Encoder Representations from Transformers) significantly advanced the field by leveraging deep learning techniques to capture contextual information more effectively. BERT, introduced by Devlin et al. utilizes a bidirectional approach to understand context from both directions in a sentence [13], enhancing the model's ability to interpret sarcasm. BERT-based models have shown improved performance in sarcasm detection tasks compared to traditional machine learning models [8].

Study on intended sarcasm detection further highlighted the advancements and challenges in sarcasm detection [8]. This study featured three subtasks:

1. **Subtask A - Sarcasm Detection**: Determine if a given text is sarcastic or not.
2. **Subtask B - Sarcasm Category Classification**: Classify a text into one or more of the categories defined by Leggit and Gibbs [14]: (1) sarcasm: contradicts the situation, directed at an addressee with a critical attitude; (2) irony: contradicts the situation, not necessarily critical, and may or may not be directed at an addressee; (3) satire: directed at an addressee who appears to be supported, but the text conveys mockery, contempt, or disagreement; (4) understatement: does not contradict the situation but minimizes its importance; (5) overstatement: does not contradict the situation but exaggerates its significance; (6) rhetorical question: a question implying an answer that contradicts the situation. Note that texts can belong to multiple categories.
3. **Subtask C - Pairwise Sarcasm Identification**: Given a sarcastic text and its non-sarcastic rephrase (two texts that convey the same meaning), identify which is the sarcastic one.

```
"<s>[INST] <<SYS>>
You are an expert in detecting sarcasm in text. Your task is to
analyze the given text and determine whether it is sarcastic or
not. Follow these steps:
- Analyze the text and identify the linguistic cues and
contextual elements that might indicate sarcasm.
- Provide a brief explanation of your analysis.
- Based on your analysis, determine whether the text is
sarcastic or not without any further explanation.
Provide your answer in the following format:
Analysis:
[Your analysis here]
Sarcasm Detection:
[Without any further explanation, only respond with either
"SARCASTIC" or "NOT SARCASTIC" as the final answer]
<</SYS>>
Below is the text to analyze:
<input>
[/INST] <To be completed by the model>"
```

Fig. 3. Llama 2 chat model prompt with chain of thought for sarcasm detection

Each subtask in the previous study provides a comprehensive benchmark for evaluating different models. The highest F1-score for sarcasm detection was 0.605 [15], achieved using an ensemble learning approach combining three transformer-based models: RoBERTa [16], DeBERTa [17], and XLM-RoBERTa [18]. The best Macro-F1 score for sarcasm category classification was 0.1630 [19], obtained with an ensemble of RoBERTa and BERTweet [20]. The highest accuracy for pairwise sarcasm identification was 0.870 [21], achieved using an ensemble of ERNIE-M [22] and DeBERTa.

Gole et al. [23] proposed a methodology for sarcasm detection using a GPT-based model from OpenAI, outlining four stages: prompt development, fine-tuning, zero-shot testing, and data analysis. In the prompt development stage, they designed prompts for both training and zero-shot testing to ensure the model received appropriate context for predictions. Fine-tuning stage involved selecting a model with specific parameter sizes and hyperparameters, training it with sarcasm datasets. During zero-shot testing, prompts were used as prefixes for inference, directing the model to return specific tokens only. Data analysis was then conducted using appropriate evaluation metrics to assess model performance.

Subsequently, Wang et al. [24] focused on designing a zero-shot text classification workflow using LLMs, detailing three main stages: data collection, LLM utilization, and classification result analysis. They highlighted the significant

Table I. Dataset, Model, and Model ID Pairs for Sarcasm Detection

Dataset	Number of Data	Model	Model ID
iSarcasmEval 2022	3462 (2592 label 0, 867 label 1)	Llama 2 7B	a-ise-22-7b
		Llama 2 13B	a-ise-22-13b
SemEval2018 Task3	4618 (2396 label 0, 2222 label 1)	Llama 2 7B	a-se-18-t3-7b
		Llama 2 13B	a-se-18-t3-13b
iSarcasm2020	3519 (2920 label 0, 599 label 1)	Llama 2 7B	a-is-20-7b
		Llama 2 13B	a-is-20-12b
Multimodal Sarcasm	10000 (5000 label 0, 5000 label 1)	Llama 2 7B	a-ms-7b
		Llama 2 13B	a-ms-13b
Combined	21599 (12911 label 0, 8688 label 1)	Llama 2 7B	a-ag-7b
		Llama 2 13B	a-ag-13b

TABLE II. DATASET, MODEL, AND MODEL ID PAIRS FOR SARCASM CATEGORY CLASSIFICATION

Dataset	Number of Data	Model	Model ID
iSarcasmEval 2022	3457	Llama 2 7B	b-ise-22-7b
		Llama 2 13B	b-ise-22-13b

differences between LLM-based workflows and traditional machine learning models for text classification. Traditional models involve complex steps such as data labeling, tokenization, and feature extraction. In contrast, LLMs streamline this process by directly using raw text data with prompts for zero-shot learning, allowing immediate classification results. This approach simplifies the workflow and demonstrates the efficiency of LLMs in handling text classification tasks without extensive preprocessing and training phases.

Another notable development is the Parameter-Efficient Fine-Tuning (PEFT) technique, specifically Low-Rank Adaptation (LoRA) [25]. LoRA enables efficient adaptation of pre-trained models by focusing on a small subset of parameters, reducing computational costs and memory requirements. Prompt engineering has also emerged as a crucial technique in optimizing LLMs for specific tasks. Techniques such as Chain of Thought (CoT) prompting guide the model by giving instruction that makes model explicitly think through a step-by-step reasoning process before providing a conclusion, enhancing its ability to generate relevant and accurate responses [26]. CoT can be applied with simple instructions such as "Let's think step by step" or by demonstrating a sequence of chained instructions to reach the expected answer. These advancements in model architecture and fine-tuning methodologies underscore the evolving landscape of sarcasm detection research. The field of sarcasm detection has progressed from traditional machine learning approaches to sophisticated Transformer-based models like LLMs. The integration of advanced techniques such as PEFT and prompt engineering continues to push the boundaries of what is possible in this challenging domain, offering new opportunities for improving the performance and reliability of sarcasm detection systems.

III. SYSTEM OVERVIEW

In this research, we delve into developing a sarcasm detection model leveraging the capabilities of the Llama 2, which requires fine-tuned adjustments to excel in specialized tasks. Our system's design is depicted in Fig. 1, which presents a schematic overview of the proposed method. The approach adopted for this research is broken down into several key stages, reflecting a comprehensive strategy to achieve best performing models on each subtask in intended sarcasm detection.

TABLE III. DATASET, MODEL, AND MODEL ID PAIRS FOR PAIRED SARCASM IDENTIFICATION

Dataset	Number of Data	Model	Model ID
iSarcasmEval 2022	867	Llama 2 7B	c-ise-22-7b
		Llama 2 13B	c-ise-22-13b

A. Data Collection

The data collection stage involving primary data from iSarcasmEval and supplementary external datasets. The iSarcasmEval dataset, divided into training and testing sets for three subtasks, serves as the primary source, with its testing data benchmarking model performance. Yuan et al. [15] enhanced their model by incorporating three publicly accessible external datasets: the iSarcasm dataset [27], a multimodal sarcasm detection dataset [28], and the SemEval-2018 Task 3 dataset [29]. These external datasets are considered for sarcasm detection to address data imbalance and provide diverse sarcasm contexts, improving model robustness. For sarcasm classification and pairwise sarcasm identification, only the iSarcasmEval dataset is utilized due to the lack of suitable external datasets. The inclusion of external datasets in sarcasm detection subtask not only balances data distribution but also enriches the variety of sarcastic contexts, enhancing the model's overall performance and generalizability.

B. Prompt Development

In this stage, prompt development is carried out separately for each task, encompassing both fine-tuning and zero-shot approaches. Effective and well-crafted prompts help the model understand the context and characteristics of sarcasm in text, leading to accurate predictions. For the fine-tuning process, prompts are designed to wrap the input text with the corresponding labels, guiding the model to respond with designated outputs. These prompts enabling the model to learn from the training dataset and adapt to the specific tasks it faces. Conversely, zero-shot prompts are crafted to ensure that the fine-tuned model can respond accurately to given tasks without further training. These prompts focus on providing clear and directed instructions to the model. Techniques such as chain of thought (CoT) can be employed, guiding the model to generate structured responses as if it were reasoning step-by-step before arriving at a conclusion. An instance of CoT prompt Llama 2 chat model formatted for sarcasm detection is shown on Fig. 3.

C. Model Fine-tuning

In the fine-tuning stage, two base models are selected for training on each subtask: the Llama 2 model with 7 billion parameters (Llama 2 7B) and the Llama 2 model with 13 billion parameters (Llama 2 13B). The preprocessed training data is fed into these models according to the experimental scenarios. These scenarios are constructed based on the dataset and the type of model used. The experimental

TABLE IV. PERFORMANCE OF EXPERIMENTAL MODELS FOR SARCASM CATEGORY CLASSIFICATION

Model	Model ID	Macro-F1	F1-sarcasm	F1-irony	F1-satire	F1-undestatement	F1-overstatement	F1-rhetorical question
Llama 2 7B	b-ise-22-7b	0.1388	**0.6167**	0.000	0.000	0.000	0.000	**0.2162**
Llama 2 13 B	b-ise-22-13b	0.1041	0.5677	0.0571	0.000	0.000	0.000	0.000
Llama 2 7B Chat	b-chat-7b	0.0631	0.2785	0.0492	0.0167	0.000	0.000	0.0345
Llama 13 B Chat	b-chat-13b	0.0830	0.3229	0.0373	0.0594	0.000	0.0313	0.0471
Best Model	b-ref-1	**0.1630**	0.4828	**0.1863**	**0.0667**	0.000	**0.0870**	0.1556
Second Best Model	b-ref-2	0.0875	0.2314	0.1622	0.0392	0.000	**0.0870**	0.0923

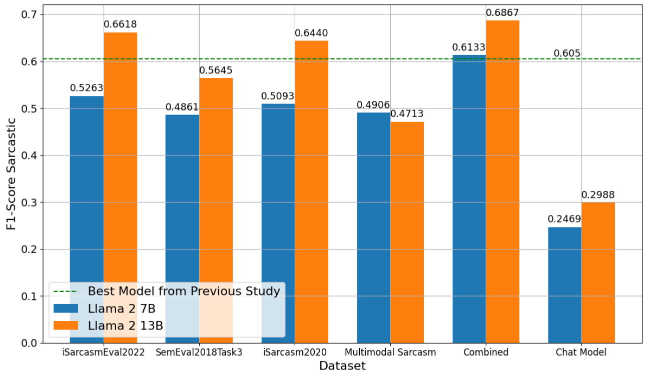

Fig. 4. F1-score sarcastic comparison of Llama 2 models on sarcasm detection fine-tuned across different datasets.

workflow is illustrated in Fig. 2. The process begins with training the base models using the training data for each subtask. Multiple fine-tuned models are developed based on the training data configurations and the number of parameters used. Each fine-tuned model is then evaluated using the corresponding test data. The evaluation results are compared to determine the best fine-tuned model for each subtask. This systematic approach ensures that the model configurations yielding the highest performance are identified and selected.

D. Zero-shot Model Testing

In this stage, the fine-tuned models are evaluated using zero-shot testing by providing prompts similar to those used during training, with each test data input processed individually. The end of each prompt is adjusted to allow the model to generate text, which is then interpreted as the predicted label for the input data. Additionally, we employed chat models developed by Meta for dialogue use cases—specifically, the Llama 2 7B Chat and Llama 2 13B Chat models—which can be used directly without additional fine-tuning. These chat models are prompted in a way that aligns their responses with the specific task objectives. The generated responses are then converted into numerical labels suitable for each task. For example, in sarcasm detection, the model is prompted to output either "SARCASTIC" or "NOT SARCASTIC." These outputs are subsequently encoded as numeric labels, with "SARCASTIC" as 1 and "NOT SARCASTIC" as 0.

TABLE V. FINE-TUNING HYPERPARAMETERS

Hyperparameter	Value	Description
Batch Size	4	Number of samples processed before the model is updated
Gradient Accumulation	4	Number of steps to accumulate gradients before updating the model weights
Optimizer	Paged AdamW 32-bit	A memory-optimized variant of the Adam optimizer for large models
Learning Rate	1e-4	Initial learning rate used for training
Max Gradient Norm	0.3	Maximum value for gradient clipping to prevent exploding gradient
Warm-up Ratio	0.03	Fraction of steps used for learning rate warm-up
Learning Rate Scheduler	Cosine	A learning rate scheduler that decreases the learning rate following a cosine curve.
Number of Training Steps	1000	Total number of training iterations
LoRA Rank	8	Rank of low-rank matrices in LoRA
LoRA Alpha	16	Scaling Factor for LoRA updates
LoRA Dropout	0.05	Dropout probability in LoRA layers

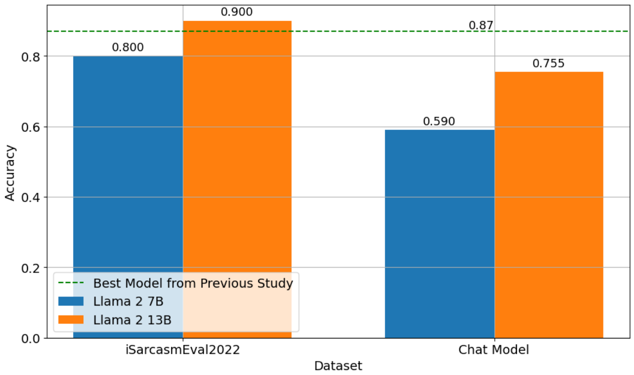

Fig. 5. Accuracy comparison of Llama 2 models on pairwise sarcasm identification.

E. Analysis and Evaluation

The model's predictions on the test data obtained from the zero-shot process are thoroughly analyzed in this stage. These predictions are compared against the ground truth labels to evaluate the model's performance. Analysis involves applying appropriate evaluation metrics for each subtask, compared with the performance of best model achieved in previous study. Additional analysis is conducted by highlighting notable test cases, such as instances where most models fail to predict correctly. The prediction results of each model for these challenging texts are analyzed to understand their characteristics or patterns. This failure analysis aims to identify potential causes of prediction errors.

IV. EXPERIMENTS

The experiment objectives are to identify the optimal combinations of training data and model parameters for each subtask and analyze how these factors affect model performance. The experiments also validate the feasibility of creating high-performance models using configurations based on PyTorch tutorial references [30]. A key focus is demonstrating the importance of prompt engineering in leading model outputs toward desired responses. Lastly, the experiments compare the response characteristics of Llama 2 fine-tuned models with Llama 2 chat models.

A. Data Processing

The data processing and experimental scenario development stage involves preparing and fine-tuning the models for each subtask. The models built for sarcasm detection are listed in Table 1, sarcasm category classification in Table 2, and pairwise sarcasm identification in Table 3. Each model is fine-tuned using the corresponding dataset. These models are optimized specifically for each subtask, ensuring that they can handle the distinct complexities of each task effectively. The datasets used for sarcasm detection show

TABLE VI. MISCLASSIFIED TEXTS IN SARCASM DETECTION

No.	Text	True Label	Predicted Label
1.	Politics is like waves, they depend on the wind where it blows.	sarcastic	not sarcastic
2.	Really bro, thanks for the update?	not sarcastic	sarcastic

TABLE VII. MISCLASSIFIED TEXTS IN SARCASM CATEGORY CLASSIFICATION

No.	Text	True Label	Predicted Label
1.	You look fantastic in that new dress. It shows off your figure	sarcasm and satire	non categorical
2.	What was the point in vaccines? Just to pay pharmaceutical companies for nothing	non categorical	sarcasm and rhetorical question

a predominance of label 0 (non-sarcastic) over label 1 (sarcastic). Conversely, the Multimodal Sarcasm dataset has a balanced distribution of labels, offering a more diverse context for analysis. Combining all these datasets results in a larger, varied dataset that maintains label diversity, with a majority being non-sarcastic but sufficiently representative for sarcastic labels. Each training dataset for sarcasm detection is sampled from different sources, reflecting various text characteristics and sarcasm labels. These comprehensive datasets are integral to training robust models capable of accurately detecting sarcasm in diverse contexts.

B. *Fine-tuning Process*

We utilized an NVIDIA Tesla V100 SXM2 32GB GPU for our computations. Due to the limited memory capacity of our GPU, full-precision fine-tuning of the Llama 2 models was not feasible. To overcome the computational challenges associated with fine-tuning large language models, we employed Parameter Efficient Fine-Tuning (PEFT) with Quantized Low Rank Adaptation (QLoRA). This method significantly reduces memory usage while preserving model performance. Using the bitsandbytes library, we quantized the Llama 2 models to 4-bit precision, which drastically lowered their memory requirements. For example, the Llama 2 7B model's memory usage dropped from an estimated 120GB for full-precision fine-tuning to 16GB in actual usage during training. Similarly, the Llama 2 13B model's memory requirement was reduced from an estimated 240GB to 26GB. QLoRA was applied across all linear layers of the Llama 2 architecture, including gate_proj, down_proj, up_proj, q_proj, v_proj, k_proj, and o_proj, to achieve this efficiency. Key hyperparameters used in our fine-tuning process are detailed in Table V.

C. *Results*

The performance of our models developed for sarcasm detection is presented in Fig. 4, using the F1-score as the evaluation metric. Several configurations of our fine-tuned Llama 2 models achieved higher F1-scores compared to the best achieved model in previous study. Overall, Llama 2 13B models outperformed the Llama 2 7B models on almost all datasets. This indicates that increasing the model parameters positively impacts performance. Notably, our a-ag-13b model showed the best performance, surpassing the best model F1-score by 0.0817. Chat models, however, exhibited lower F1-scores compared to the fine-tuned models, indicating the significant impact of fine-tuning and dataset configuration. The performance for sarcasm category classification, using Macro-F1 as the evaluation metric and detail F1-score for each category are shown in Table IV. Our b-ise-22-7b model achieved the best performance in sarcasm category classification, with a Macro-F1 score of 0.1388, which is competitive though slightly lower than the best model in previous study by 0.0242. It performed significantly better than the second-best model by 0.0513. The F1-sarcasm and F1-rhetorical question scores for this model were notably higher than those of other models. For pairwise sarcasm identification, accuracy was used as the evaluation metric, and the results are shown in Fig. 5. Our c-ise-22-13b model achieved the highest accuracy of 0.90, outperforming the best achieved model's score of 0.87.

TABLE VIII. MISCLASSIFIED TEXTS IN PAIRWISE SARCASM IDENTIFICATION

No.	Text Pair	True Label	Predicted Label
1.	Text 0: I know you're busy but can I have a cuppa please? Text 1: I think my husband has lost the ability to find the kitchen	text 1	text 0
2.	Text 0: As usual, the government has responded in adequate time to prevent the spread of the omicron variant and has a plan on how to keep infections to a minimum Text 1: As before, the spread of a new (Omicron) variant has taken the government by surprise and it has yet to devise a plan to keep the infections to a minimum	text 0	text 1

V. DISCUSSION

We analyzed the characteristics of texts that most models failed to predict correctly. The purpose of this failure analysis is to identify patterns or factors that might cause the models to struggle with certain texts. For sarcasm detection, we examined test texts that were labeled as sarcastic but were predicted as non-sarcastic by most models, and vice versa. The results are shown in Table VI. From the evaluation, it is evident that texts labeled as sarcastic but predicted as non-sarcastic can contain subtle cues that are challenging to detect without additional context. For instance, metaphors and nuanced language can obscure the intended sarcasm. In contrast, texts labeled as non-sarcastic but predicted as sarcastic can involve exaggerated statements or ironic punctuation, such as inverted question marks, which the models mistakenly interpret as sarcasm due to their previous training experiences.

In sarcasm category classification, texts that were labeled with sarcasm and satire but not categorized by the models, as shown in Table VII, contained compliments intended as insults or statements with an underlying critical tone that were missed without the necessary social context. On the other hand, non-categorical texts misclassified into sarcasm categories typically involved rhetorical questions or critical tones, which the models, influenced by their training data, incorrectly interpreted as sarcastic. Additionally, our analysis revealed that none of the models were able to recognize "understatement" sentences as shown in Table IV. This can be attributed to the factor that iSarcasmEval dataset contains very few examples of understatement compared to other categories. In the training set, only 10 from 867 of the samples were labeled as understatement, making it challenging for models to learn this category effectively. For pairwise sarcasm identification, we analyzed pairs of texts where the ground truth label was not correctly predicted by most models. The results, as shown in Table VIII, indicate that the failure to detect sarcasm in paired texts can stems from the models' inability to capture subtle ironic expressions or the context implied between paired statements.

VI. Conclusions

The experimental results highlight key findings. Our Llama 2 13B model fine-tuned with a combined dataset achieved an F1-score of 0.6867 for sarcasm detection, surpassing previous study bests by 0.0817. For sarcasm category classification, our Llama 2 7B model fine-tuned with the iSarcasmEval dataset achieved a Macro-F1 score of 0.1388, exceeding second-best models from previous study by 0.0513. In pairwise sarcasm identification, our Llama 2 13B model finetuned with the iSarcasmEval dataset reached an accuracy of 0.9, outperforming previous study bests by 0.03. These results show that combined datasets and larger models generally enhance performance. Implementing PEFT with QLoRA 4-bit quantization reduced memory requirements while maintaining performance, resulting Llama 2 viable on resource-limited devices. Prompt engineering and CoT techniques improved contextual analysis, though some misclassifications highlighted the need for better context understanding.

Future research should focus on enhancing model generalization and performance in sarcasm detection by incorporating a more diverse and extensively labeled dataset, particularly in the sarcasm category classification, to ensure data quality and represent a broader range of contexts. Further exploration of advanced LLMs that better capture sarcasm context than Llama 2 should be conducted, along with systematic comparisons to pinpoint models that deliver superior sarcasm detection performance. Lastly, the development of systems using the proposed solution approach can be aimed for a general-purpose text classification that surpasses traditional methods in efficiency while maintaining competitive model performance.

Acknowledgment

We would like to express our gratitude and appreciation to the School of Electronics and Informatics Engineering (STEI) of Bandung Institute of Technology (ITB) for their invaluable support and resources provided throughout this research. Their assistance and provision of essential facilities have been instrumental in the successful completion of this research.

References

[1] J. Sinclair, *Collins COBUILD advanced learner's English dictionary*, 4th ed. Glasgow, Kraków: HarperCollins Publishers ; Express Pub. - EGIS [dystr.] Glasgow, Kraków, 2003.

[2] R. Filik, A. Ţurcan, D. Thompson, N. Harvey, H. Davies, and A. Turner, "Sarcasm and emoticons: Comprehension and emotional impact," *Quarterly Journal of Experimental Psychology*, vol. 69, no. 11, pp. 2130–2146, Nov. 2016.

[3] I. A. Farha, W. Zaghouani, and W. Magdy, "Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic," 2021. [Online]. Available: https://www.appen.com/

[4] S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov, "SemEval-2014 Task 9: Sentiment Analysis in Twitter," 2014. [Online]. Available: https://dev.twitter.com

[5] D. Davidov, O. Tsur, and A. Rappoport, "Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon," Association for Computational Linguistics, 2010.

[6] S. Muresan, R. Gonzalez-Ibanez, D. Ghosh, and N. Wacholder, "Identification of nonliteral language in social media: A case study on sarcasm," *J Assoc Inf Sci Technol*, vol. 67, no. 11, pp. 2725–2737, Nov. 2016.

[7] A. Vaswani *et al.*, "Attention Is All You Need," Jun. 2017.

[8] I. A. Farha, S. V. Oprea, S. R. Wilson, and W. Magdy, "SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic," 2022.

[9] S. V. Oprea and W. Magdy, "Exploring Author Context for Detecting Intended vs Perceived Sarcasm," Association for Computational Linguistics. [Online]. Available: https://www.reddit.com

[10] A. C. Băroiu and Ștefan Trăușan-Matu, "How capable are state-of-the-art language models to cope with sarcasm?," in *Proceedings - 2023 24th International Conference on Control Systems and Computer Science, CSCS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 399–402.

[11] M. R. Douglas, "Large Language Models," Jul. 2023.

[12] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018.

[14] J. S. Leggitt and R. W. Gibbs, "Emotional reactions to verbal irony," *Discourse Process*, vol. 29, no. 1, pp. 1–24, 2000.

[15] M. Yuan, M. Zhou, L. Jiang, Y. Mo, and X. Shi, "stce at SemEval-2022 Task 6: Sarcasm Detection in English Tweets," 2022.

[16] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019.

[17] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.03654

[18] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.02116

[19] X. Du, D. Hu, M. Jin, L. Jiang, Y. Mo, and X. Shi, "PALI-NLP at SemEval-2022 Task 6: iSarcasmEval-Fine-tuning the Pre-trained Model for Detecting Intended Sarcasm," 2022.

[20] D. Quoc Nguyen, T. Vu, A. Tuan Nguyen, and V. Research, "BERTweet: A pre-trained language model for English Tweets," 2020.

[21] Y. Han *et al.*, "X-PuDu at SemEval-2022 Task 6: Multilingual Learning for English and Arabic Sarcasm Detection," 2022.

[22] X. Ouyang *et al.*, "ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora." [Online]. Available: https://github.

[23] M. Gole, W.-P. Nwadiugwu, and A. Miranskyy, "On Sarcasm Detection with OpenAI GPT-based Models," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.04642

[24] Z. Wang, Y. Pang, and Y. Lin, "Large Language Models Are Zero-Shot Text Classifiers," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.01044

[25] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," Jun. 2021.

[26] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.14735

[27] S. Oprea and W. Magdy, "iSarcasm: A Dataset of Intended Sarcasm," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.03123

[28] headacheboy, "data-of-multimodal-sarcasm-detection," GitHub. Accessed: May 05, 2024. [Online]. Available: https://github.com/headacheboy/data-of-multimodal-sarcasm-detection

[29] C. Van Hee, E. Lefever, and V. Hoste, "SemEval-2018 Task 3: Irony Detection in English Tweets," in *Proceedings of the 12th International Workshop on Semantic Evaluation*, M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, Eds., New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 39–50. doi: 10.18653/v1/S18-1005.

[30] Y. Belkada *et al.*, "Finetune LLMs on your own consumer hardware using tools from PyTorch and Hugging Face ecosystem," PyTorch. Accessed: May 05, 2024. [Online]. Available: https://pytorch.org/blog/finetune-llms/