Ingénierie des Systèmes d'Information
Vol. 31, No. 8, August, 2026, pp. 2643-2651
Journal homepage: http://iieta.org/journals/isi

Sarcasm Detection with Small-Scale Decoder-Only Language Models via Low-Rank Adaptation-Based Parameter-Efficient Adaptation

Wahyu Tisno Atmojo[1*], Pulung Nurtantio Andono[2], Abdul Syukur[2], Ahmad Zainul Fanani[2]

[1] Faculty of Sains and Technology, Universitas Pradita, Tangerang 15810, Indonesia
[2] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

Corresponding Author Email: wahyu.tisno@pradita.ac.id

https://doi.org/10.18280/isi.310811

Received: 21 April 2026
Revised: 29 July 2026
Accepted: 12 August 2026
Available online: 31 August 2026

Keywords:
sarcasm detection, small-scale language models, parameter-efficient fine-tuning, Low-Rank Adaptation, decoder-only architecture, natural language processing

ABSTRACT

Sarcasm detection remains a challenging problem in natural language processing (NLP) because the intended meaning of sarcastic expressions often differs from their literal interpretation. Although large language models (LLMs) provide strong semantic understanding capabilities, their computational requirements limit their application in resource-constrained environments. This study investigates the effectiveness of small-scale decoder-only language models combined with parameter-efficient adaptation for sarcasm detection. Three models, namely Qwen3-1.7B, Llama-3.2-1B, and Gemma-2-2B, are evaluated using a balanced news headline dataset containing 23,448 samples. Two experimental settings are considered: zero-shot generative classification and Low-Rank Adaptation (LoRA)-based parameter-efficient fine-tuning (PEFT) with a Multilayer Perceptron (MLP) classification head. The experimental results reveal that zero-shot small-scale LLMs suffer from predictive collapse caused by strong bias toward a dominant class. After LoRA adaptation, all evaluated models achieve substantial improvements, demonstrating enhanced semantic representation and classification capability. Among them, Gemma-2-2B obtains the best performance with an F1-score of 0.9599 and an area under the receiver operating characteristic curve (AUC-ROC) value of 0.9810. Statistical analysis using McNemar's test confirms that the performance improvements are significant ($p < 0.05$). These findings demonstrate that small-scale decoder-only language models can achieve competitive sarcasm detection performance through efficient adaptation strategies, providing a practical direction for lightweight NLP systems.

1. INTRODUCTION

The emergence of social media has transformed the way people communicate, making it very easy for them to express themselves [1]. Individuals are free to express their opinions, humour, insults, and other forms of expression using sarcasm [2]. Sarcasm is the act of saying the opposite of what you mean in order to make someone else feel foolish or to show them that you are angry [3]. Detecting sarcasm is crucial in natural language processing (NLP), particularly in text classification systems [4-6]. This is because a common feature of sarcasm is irony—that is, the discrepancy between the literal meaning and the intended meaning of a text [7]. User-generated content on social media platforms such as Twitter, Reddit and online forums is typically informal and context-dependent, which makes it even more difficult to detect sarcasm [8]. In face-to-face conversations, recognizing sarcasm is fairly easy, as it is aided by facial expressions, tone of voice and body language [1, 9]. However, it becomes difficult and complicated when communication takes place in writing, such as on social media [10]. Detecting sarcasm is tricky because these cues are absent when text is posted on social media [11]. The challenge becomes far greater when identifying sarcasm in shared content that consists not only of text but also of images, videos, and audio [11-13]. In news articles, sarcasm often employs unexpected combinations of ideas, an ironic tone, or exaggerated situations to convey criticism [14, 15]. The failure of language models to capture these nuances can lead to significant misinterpretations in sentiment analysis and broader contextual understanding [16].

With the advancement of artificial intelligence, the use of large language models (LLMs) based on a decoder-only architecture has become the new standard for a wide range of text classification tasks [12, 17, 18]. However, there is a major challenge in striking a balance between the model's reasoning capabilities and computational efficiency [19]. Gigantic models with parameters exceeding 4B often place an extremely heavy load on Video Random Access Memory (VRAM), making them difficult to implement on standard Graphics Processing Unit (GPU) infrastructure [20]. On the other hand, models with fewer than 1 billion parameters have been shown empirically to exhibit a deficit in the pragmatic reasoning required to detect contextual inconsistencies in sarcastic text [21]. This study focuses on parameter classes 1B to 2B, which are considered to represent the optimal balance between performance and efficiency. The primary focus of this study is to determine which of the three leading models—Qwen3-1.7B-Base (Alibaba), Llama-3.2-1B (Meta), and

Gemma-2-2B (Google) is capable of generating the most accurate semantic representations for sarcasm classification. These three models were selected because they represent the diversity of large-scale training corpora and multilingual capabilities, including Indonesian.

Although research on LLMs is advancing rapidly, the majority of current studies still focus on massive models with over 7 billion parameters, leaving a knowledge gap regarding the effectiveness of small-scale models in complex pragmatic tasks. Existing methods face significant challenges because large-scale models require extremely high computational costs, while small models often fail to capture the nuances of sarcasm due to limitations in pragmatic reasoning and in-depth semantic analysis. This phenomenon leads to predictive collapse in zero-shot conditions, where small-scale models tend to get trapped in dominant class biases due to their inability to distinguish between literal and intended meanings. The Low-Rank Adaptation (LoRA)-Multilayer Perceptron (MLP) framework was chosen as a solution because it combines the parameter adaptation efficiency of LoRA with a structured MLP classification layer, enabling 1B–2B-scale models to achieve strong semantic representations and correct classification bias without requiring massive computational resources.

Sarcasm detection, however, is far from straightforward. The core difficulty lies in the inherent mismatch between what is said and what is actually meant, a gap that requires more than lexical analysis to bridge [22]. Unlike sentiment or topic classification, sarcasm often masquerades as a genuine statement, making it nearly invisible to models that rely solely on surface-level features [23]. The problem is further compounded in written communication, where vocal inflection, pauses, and facial expressions, cues that humans instinctively use to recognize irony, are entirely absent. Even when context is available, interpreting sarcasm frequently demands world knowledge or cultural familiarity that current models struggle to encode reliably [24]. From a practical standpoint, the field also faces a resource dilemma: LLMs have shown promise in capturing these pragmatic subtleties, yet their computational footprint makes them impractical for many real-world applications. Meanwhile, smaller and more efficient models tend to underperform precisely because sarcasm operates at a level of abstraction that shallow representations fail to capture [25]. These overlapping constraints, linguistic, contextual, and computational, form the central problem space that this study seeks to address.

To provide a comprehensive evaluation, this study tests the model under two integrated evaluation conditions: zero-shot generative as a baseline and parameter-efficient fine-tuning (PEFT) via the LoRA method combined with an MLP Classifier Head. This methodology draws on a cross-learning-paradigm evaluation framework for social media classification tasks. By comparing the performance of these three models on a balanced news headlines dataset, this study aims to provide recommendations on the most effective small-scale LLMs for NLP tasks that are sensitive to pragmatic context. The evaluation was conducted across two integrated testing conditions: a generative baseline and PEFT using LoRA [26], combined with a task-specific MLP Classifier Head. PEFT enables adaptation of pretrained models by updating a relatively small number of additional parameters [27], while LoRA specifically introduces trainable low-rank updates to selected pretrained layers [26] to assess the effectiveness of the model's adaptation to a specific task.

Despite the rapid advancement of LLMs, existing studies predominantly focus on models with over 7 billion parameters, leaving a critical knowledge gap regarding the effectiveness and failure modes of small-scale models (1B–2B parameters) in pragmatic tasks like sarcasm detection. Furthermore, baseline zero-shot deployments often suffer from severe classification bias, while standard fine-tuning can be computationally prohibitive. To address these challenges, this study proposes a LoRA-MLP hybrid framework for small-scale decoder-only LLMs. The main contributions of this work are threefold:

•Empirical analysis of predictive collapse: We identify and document the phenomenon of predictive collapse in small-scale base LLMs under zero-shot conditions for sarcasm detection.

•Hybrid adaptation framework: We propose a parameter-efficient LoRA-MLP architecture that attaches a specialized classifier head to frozen 8-bit LLM backbones, achieving high accuracy with minimal parameter updates.

•Cross-ecosystem small-LLM benchmarking: We provide a systematic evaluation comparing leading 1B–2B models (Qwen3-1.7B, Llama-3.2-1B, and Gemma-2-2B) across zero-shot and fine-tuned paradigms on a standardized news headlines corpus.

2. METHOD

2.1 Research workflow

This study follows a systematic approach comprising five main stages: (1) collection and selection of datasets, (2) data pre-processing, including undersampling and stratified splitting, (3) comparative experiments on three LLM models under two testing conditions, (4) comprehensive evaluation using classification metrics, and (5) in-depth analysis to determine which model is the best.

2.2 Dataset characteristics and preprocessing

This study uses the News Headlines Dataset for Sarcasm Detection, which consists of the sarcasm and non-sarcasm classes. The data was taken from a previous study conducted by Misra and Arora [21], where the data source consists of headlines collected from two news websites: The Onion and HuffPost. The reason for using this dataset is that the headlines are written by professionals in a formal style, with no spelling mistakes or informal language, unlike in social media-based datasets [28-30]. This reduces vocabulary sparsity and also increases the likelihood of finding pre-trained embeddings to improve performance [29]. The dataset will then undergo data balancing: as the initial distribution is unbalanced (11,724 sarcastic vs 14,985 non-sarcastic), and the number of data points between classes is unbalanced, undersampling will be performed by taking the smallest number of classes present in the dataset, namely the "Sarcastic" class, so that the ratio becomes 1:1, with 11,724 non-sarcastic classes and 11,724 sarcastic classes. A random undersampling technique is used to achieve a 1:1 ratio, resulting in a total of 23,448 samples. Random undersampling was chosen to efficiently achieve a balanced 1:1 class ratio, given that the number of "Sarcastic" data points (11,724) was still considered sufficient to train a 1B-2B-scale model without significant loss of information. The datasets before and after undersampling are shown in

Table 1.

Following class balancing, the dataset was split using the stratified split technique with a ratio of 80:10:10. The stratified split ensures that a 50:50 class distribution is maintained proportionally across each subset (training, validation, and test), thereby preventing any imbalance in class representation within any subset. The stratified split with a ratio of 80:10:10 is shown in Table 2.

2.3 Model selection

We evaluated three decoder-only models across a parameter range of 1B–2B, representing architectures from different global developers:

•Qwen3-1.7B-Base: A pretrained base language model from the Qwen3 family developed by the Qwen Team [31].

•Llama-3.2-1B: A lightweight text-only language model released by Meta as part of the Llama 3.2 family [32].

•Gemma-2-2B: A lightweight open language model from the Gemma 2 family developed by Google DeepMind, with a 2-billion-parameter model configuration [33].

These three models were selected because they represent three major ecosystems: Alibaba, Meta, and Google, with diverse training corpora and strong multilingual capabilities, and fall within the 1B–2B parameter range, which is the sweet spot between efficiency and performance. The main specifications of the selected models are summarized in Table 3.

Table 1. Dataset distribution table

Condition	Sarcasm	Not Sarcasm	Total	Ratio
Before Undersampling	11.724	14.985	26.709	1:1,28
After Undersampling	11.724	11.724	23.448	1:1 (50:50)

Table 2. Distribution of the dataset following an 80:10:10 stratified split

Subset	Total Sample	Sarcasm	Not Sarcasm	Function
Training set (80%)	18.758	9.379	9.379	Training Model (for LoRA fine-tuning)
Validation set (10%)	2.345	1.172	1.173	Early stopping (for LoRA fine-tuning)
Test set (10%)	2.345	1.172	1.173	Final evaluation (for zero-shot and LoRA fine-tuning)

Note: LoRA = Low-Rank Adaptation.

Table 3. Specifications of the model used

Model	Developer	Parameter	Hidden Size
Qwen3-1.7B	Alibaba Cloud	1.7B	2.048
Llama-3.2-1B	Meta AI	1B	2.048
Gemma-2-2B	Google DeepMind	2B	2.304

2.4 Experimental design

The experiments were conducted under two main scenarios to test the model's semantic representation capabilities, as shown in Table 4.

Table 4. Experimental design

Condition	Description	Sample Case	Prediction Mechanism
Zero-shot (baseline)	Task instructions + characteristics	There isn't any	Generate Teks → Parse 'Sarcasm/News'
LoRA fine-tuning + MLP head	Fine-tuning the entire training set with the INT8 + LoRA adapter backbone	Train set (18.758)	Sigmoid MLP Classifier Head (BCELoss)

Note: MLP = Multilayer Perceptron, LoRA = Low-Rank Adaptation.

2.5 Zero-shot generative (baseline)

In this scenario, the model is provided with task instructions and descriptions of the characteristics of sarcasm and news, without any additional training. Predictions are made using greedy decoding with `max_new_tokens = 10`. The text output is then parsed into binary labels: 1 for 'Sarcasm' and 0 for 'News'. The design of the Zero-Shot Generative test is shown in Table 5.

In the zero-shot setting, the prediction mechanism used is text generation with output parsing. The model is generated using greedy decoding with `max_new_tokens = 10` tokens, producing short response texts. The output is then parsed: if it contains 'Sarcasm' → label 1; if it contains 'News' → label 0; if neither → fallback to News (0). The evaluation metrics and results for the zero-shot setting are accuracy, F1 score, and macro F1. Evaluation is performed using an unseen test set (10% of the test data) by mapping the prediction results into a Confusion Matrix to obtain the values for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Table 5. Zero-shot generative testing design

Component	Contents
Assignment instructions	Classify headlines as 'Sarcasm' or 'News'
Characteristics of sarcasm	Exaggerated/absurd, ironic tone, humorous/satirical, unexpected combinations
News characteristics	Neutral tone, informative/descriptive, reports real events without irony
Reference example	None
Output format command	"Reply ONLY with 'Sarcasm' or 'News'. Do not explain."
Decoding strategy	Greedy decoding, max_new_tokens = 10

2.6 Low-Rank Adaptation fine-tuning with a Multilayer Perceptron Classifier Head

The LoRA fine-tuning with MLP Classifier Head test design is divided into four main blocks: the Preprocessing Layer (Tokenisation & Padding), where the raw news text (raw input text) is first converted into a sequence of tokens that can be understood by the machine. A padding process is

applied to ensure all inputs have a uniform sequence length before entering the model. The second block is the LLM Backbone with LoRA Adapters; this is the core component comprising the transformer layers of the selected model with LoRA adapters inserted. Only the trainable weights in the adapters are updated, whilst the main LLM layers are frozen in INT8 precision for computational efficiency. The third block is the EOS Pooling Layer. This layer takes the final vector representation of the closing token to summarize the entire semantic meaning of the sentence into a single compact vector representation. The fourth block is the MLP Classifier Head. The pooled vector is then passed to an additional neural unit (MLP) specifically designed for classification. This block contains four operations: a Linear Layer, which transforms the feature dimensions from the model's hidden size to 256 units; ReLU and Dropout, to introduce non-linearity and prevent overfitting; and a Sigmoid Output Layer, to convert the final numerical score into a probability value between 0 and 1. Values close to 1 are classified as Sarcasm, whilst values close to 0 are classified as News. The test design is shown in Figure 1.

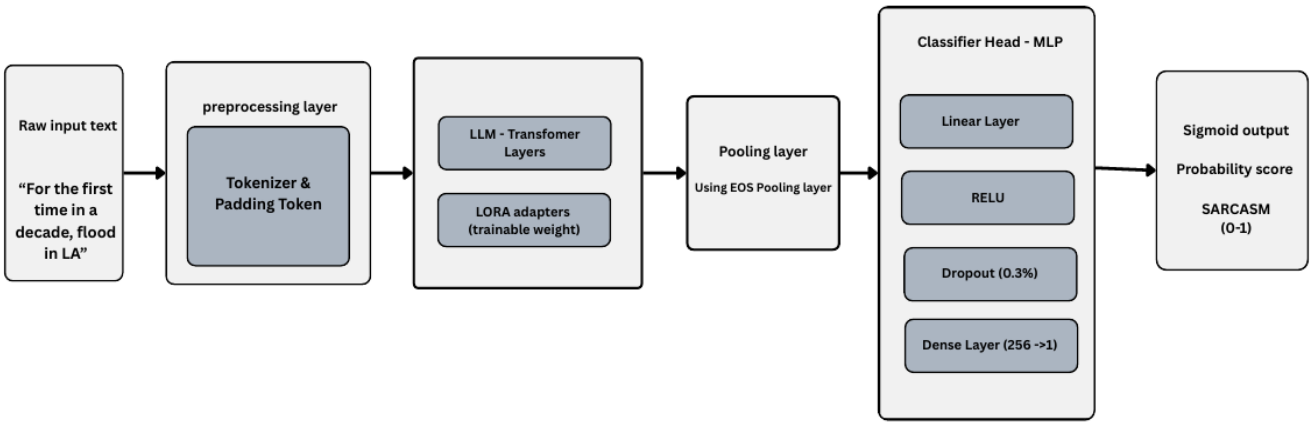

Figure 1. Test design architecture for Low-Rank Adaptation (LoRA) fine-tuning with a Multilayer Perceptron (MLP) Classifier Head

To improve accuracy, the researchers applied PEFT using the LoRA method.

•LoRA Configuration: Using *rank* (*r*) = 8, *alpha* = 16, and the module target at q_{proj} and v_{proj}.

•Architecture: The model uses INT8 quantization. The output from the *transformer layers* is processed through an EOS Pooling Layer and fed into the MLP Classifier Head. The weight update in the transformer layer is defined as:

$$h = W_{0x} + \Delta W x = W_{0x} + BAx$$

where, h is the output, W_0 is the frozen weight matrix, ΔW is the weight update matrix, and B and A are low-rank matrices.

•Classifier head: Consists of a Linear Layer (with 256 hidden units), ReLU activation, Dropout ($p = 0.10$), and a final linear layer with Sigmoid activation to produce binary probabilities.

•Hyperparameters: Training was carried out using a learning rate of 2×10^{-4}, an effective batch size of 16, and Binary Cross-Entropy Loss (BCELoss) for a maximum of 10 epochs with early stopping.

2.7 Evaluation metrics

Model performance was evaluated quantitatively using accuracy, macro precision, macro recall, and macro F1-score. In addition, a confusion matrix was used to analyze classification errors (FPs and FNs) in order to identify model bias in specific classes.

All experiments were conducted on an NVIDIA RTX 3090 GPU with 24GB VRAM. The models were trained using a learning rate of 2×10^{-4}, a batch size of 16, and a dropout rate of 0.1. Early stopping was applied with a patience of 3 epochs. The selected models represent diverse architectures from major AI ecosystems (Alibaba, Meta, and Google), enabling comprehensive comparison. Because the models were evaluated on the same held-out test instances, statistical significance was assessed using McNemar's test on the paired classification outcomes. The test evaluates whether the two compared classifiers exhibit significantly different error patterns on the same test instances. Statistical significance was determined using a threshold of $p < 0.05$.

3. RESULT

3.1 Zero-shot generative performance

Under zero-shot conditions, the three models were tested using a single prompt template containing task instructions and descriptions of the characteristics of Sarcasm and News. The results of the first test, conducted under zero-shot generative conditions, revealed significant differences in performance among the three models in capturing the semantic representation of sarcasm without additional training. A summary of the evaluation metrics is presented in Table 6.

Table 6. Performance of zero-shot large language models (LLMs) (1B–2B parameters)

Model	Parameter	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
Qwen3-1.7B-Base	1.7 Billion	0.5365	0.6887	0.5367	0.4197
Llama-3.2-1B	1 Billion	0.507	0.5708	0.5068	0.363
Gemma-2-2B	2 Billion	0.4998	0.4722	0.4996	0.3362

Table 7. Comparison of models without fine-tuning

Model	TN (News → News)	FP (News → Sarc)	FN (Sarc → News)	TP (Sarc → Sarc)
Qwen3-1.7B-Base	103	1,070	17	1,155
Gemma-2-2B	1,168	5	1,17	2
Llama-3.2-1B	1,152	21	1,135	37

Note: TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

Qwen3-1.7B-Base has the highest accuracy and detects almost all instances of sarcasm (TP = 1,155, Sarcasm recall = 0.5367), but almost always misclassifies 'News' as 'Sarcasm' (FP = 1.070). Conversely, Gemma-2-2B and Llama-3.2-1B detect almost no sarcasm at all (TP of 2 and 37, respectively). This contrasting pattern suggests that without fine-tuning, base models tend to fall into a single dominant class, as shown in Table 7.

3.2 Analysis of model bias and confusion matrix

The distribution of predictions under zero-shot conditions shows that all three models tend to 'collapse' into a single dominant class. This phenomenon can be seen in the following confusion matrix:

Qwen3-1.7B-Base: This model demonstrates very high sensitivity to the sarcasm class, with a TP rate of 1,155 and a Recall of 0.5367. However, the model fails to distinguish normal news, resulting in a very high FP rate (1.070), whereby almost all news text is classified as sarcasm. The confusion matrix for Qwen3-1.7B is shown in Figure 2.

Llama-3.2-1B and Gemma-2-2B: In contrast to Qwen3, these two models are barely capable of detecting sarcasm in generative tasks. Llama-3.2-1B produced only 37 TPs, whilst Gemma-2-2B produced just 2 TPs. The majority of the test data was classified as News, resulting in very high FN rates (1.135 for Llama and 1.168 for Gemma). The confusion matrix for Llama-3.2-1B and Gemma-2-2B is shown in Figure 3.

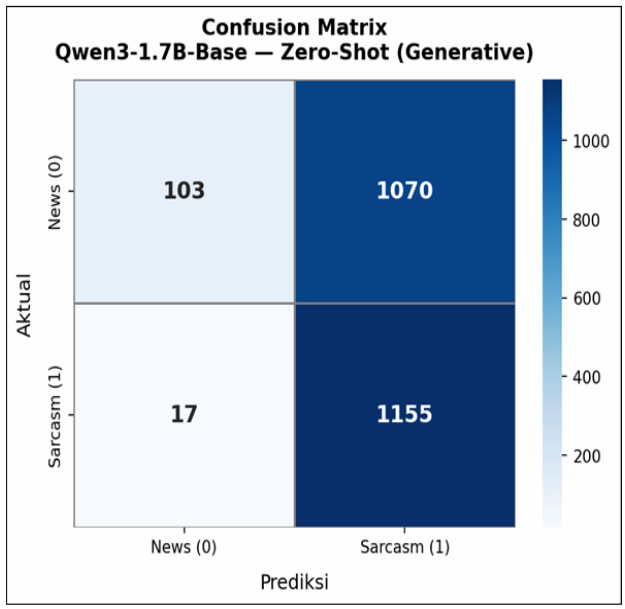

Figure 2. Confusion matrix for qwen3-1.7B-base

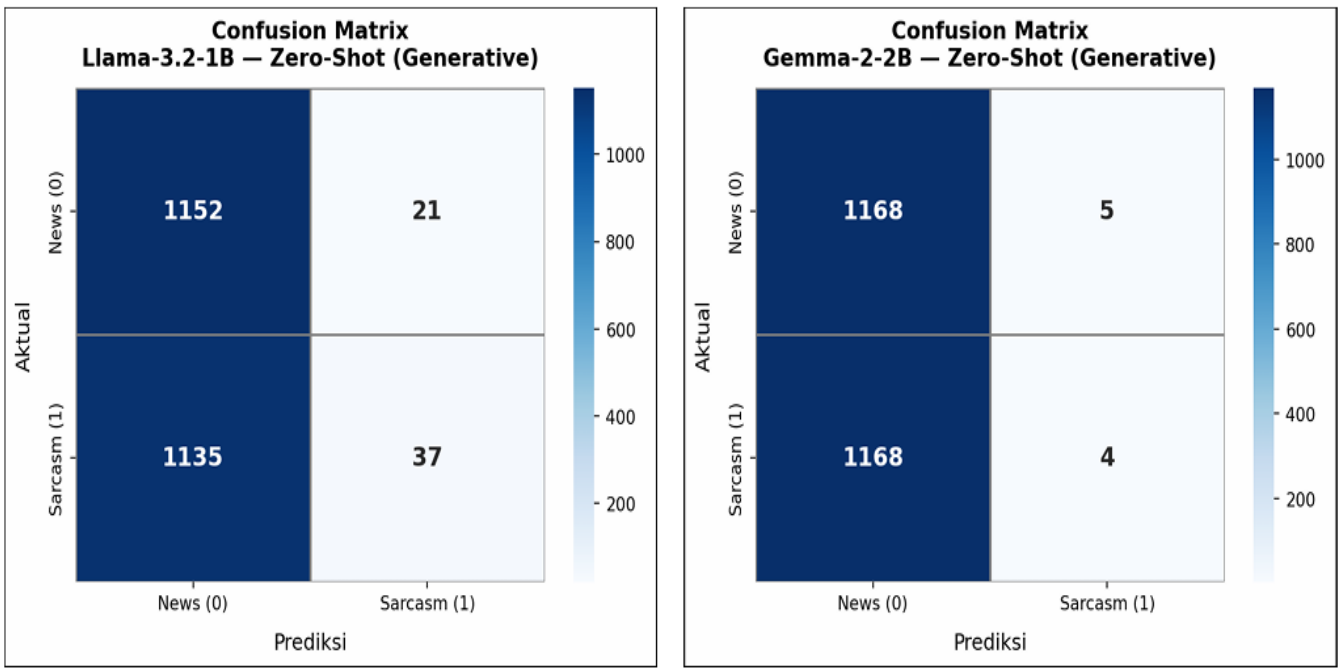

Figure 3. Confusion matrix for Llama-3.2-1B and Gemma-2-2B

Table 8. Test results using Low-Rank Adaptation (LoRA) fine-tuning + Multilayer Perceptron (MLP) Classifier Head

Model	Best Epoch	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Qwen3-1.7B	7	0.9390	0.9393	0.9390	0.9390	0.9658
Gemma-2-2B	8	0.9599	0.9599	0.9599	0.9599	0.9810
Llama-3.2-1B	4	0.9463	0.9465	0.9463	0.9463	0.9783

Note: AUC-ROC = area under the receiver operating characteristic curve.

3.3 Second test results—Low-Rank Adaptation fine-tuning + Multilayer Perceptron Classifier Head

In the LoRA fine-tuning scenario, all three models were fine-tuned using the LoRA configuration (r = 8, alpha = 16) with an INT8-quantised backbone and an MLP Classifier Head. The results of the fine-tuning, including Accuracy, Precision, Recall, F1-Score, and AUC-ROC, are shown in Table 8.

Table 8 shows that Gemma-2-2B achieved the highest F1 score on the test set (0.9599), followed by Llama-3.2-1B (0.9463) and Qwen3-1.7B (0.9390). The superior performance of Gemma-2-2B can be attributed to several architectural and training factors. First, Gemma-2-2B employs grouped-query attention and sliding window attention mechanisms, which enhance its ability to capture long-range dependencies critical for detecting subtle sarcastic patterns. Second, its pre-training on a diverse multilingual corpus may have exposed it to a broader range of figurative language constructs. In contrast, Qwen3-1.7B's strong zero-shot sarcasm bias may stem from its instruction-following pre-training objectives, which predispose it toward certain token sequences. Llama-3.2-1B's earlier convergence at epoch 4 suggests a risk of underfitting, potentially limited by its smaller parameter budget. These differences underscore that model architecture and pre-training design, beyond parameter count alone, are decisive factors in pragmatic NLP tasks such as sarcasm detection.

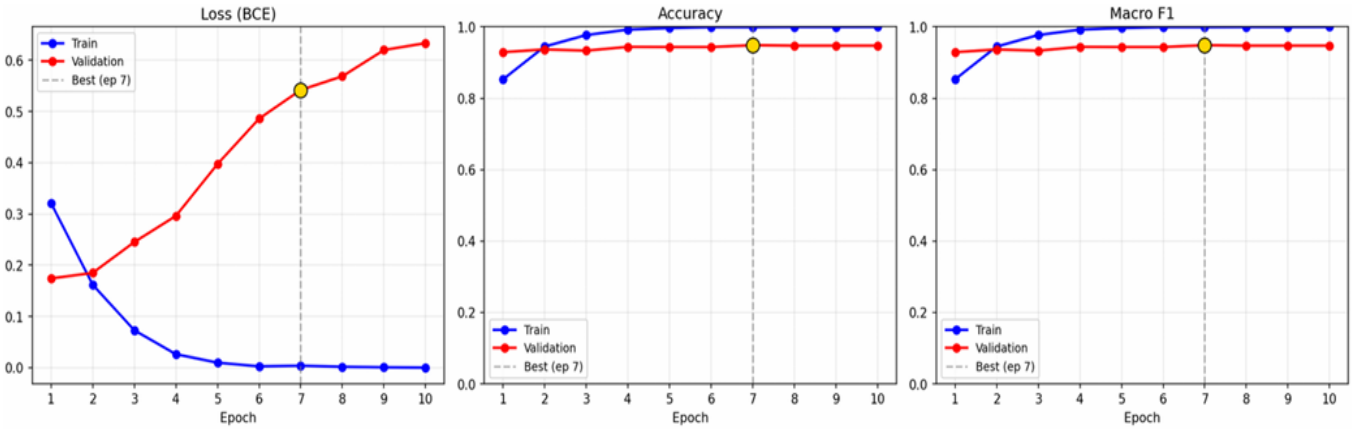

Figure 4. Training and validation curve for Qwen3-1.7B base + Low-Rank Adaptation (LoRA)

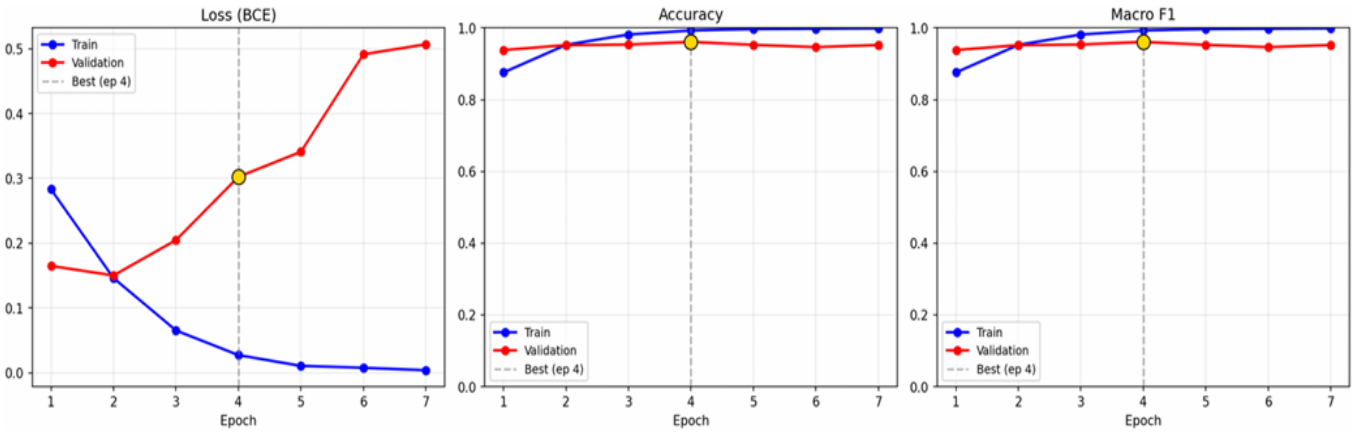

Figure 5. Training and validation curve for Llama 3.2 1B + Low-Rank Adaptation (LoRA)

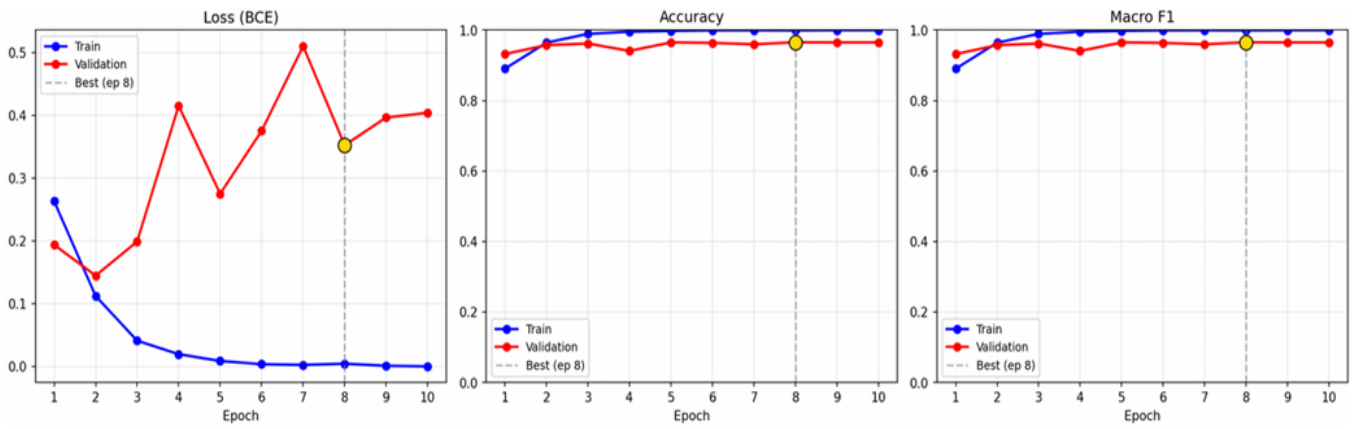

Figure 6. Training and validation curve for Gemma 2 2B + Low-Rank Adaptation (LoRA)

The AUC-ROC for all three is above 0.96, indicating excellent discrimination/classification ability between the two classes. Training was then carried out for all three models using LoRA, with a maximum of 10 epochs and early stopping set at 3 epochs. The Qwen3-1.7B model achieved its best performance at the 7th epoch with a validation F1 score of 0.9488, as shown in Figure 4.

The Llama 3.2 1B model achieved its best performance in the fourth epoch, with a validation F1 score of 0.9612, as shown in Figure 5.

The Gemma 2 2B model achieved its best performance in the 8th epoch, with a validation F1 score of 0.9655, as shown in Figure 6.

3.4 Comparison of the performance of the proposed model

Table 9 presents a contextual comparison between the results reported in previous studies on news-headline sarcasm detection and the results obtained in this study using LoRA fine-tuning with an MLP Classifier Head. Because dataset versions, preprocessing procedures, data splits, and evaluation protocols may differ across studies, the values in Table 9 are not treated as a strictly controlled benchmark. The comparison is therefore intended to provide contextual evidence rather than a definitive numerical superiority claim.

Table 9. Contextual comparison with previous news-headline sarcasm detection studies

Study	Model	Dataset	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Ali et al. [2]	GMP-LSTM	News Headlines	92,5486	97,5473	98,2501	98,3913
Roy et al. [34]	CBMPABiLSTM	News Headlines	95,84	96,14	95,25	96,20
This study	Qwen3-1.7B + LoRA + MLP	News Headlines	93,90	93,90	93,93	93,90
This study	Llama-3.2-1B + LoRA + MLP	News Headlines	94,63	94,63	94,65	94,63
This study	Gemma-2-2B + LoRA + MLP	News Headlines	95,99	95,99	95,99	95,99

Note: GMP-LSTM = Global Max Pooling–Long Short-Term Memory, CBMPABiLSTM = Convolutional Blocks–MaxPooling–Attention–Bidirectional Long Short-Term Memory, MLP = Multilayer Perceptron, LoRA = Low-Rank Adaptation.

Table 9 provides contextual evidence regarding the performance of the proposed models in relation to previous studies on news-headline sarcasm detection. Among the three models evaluated in this study, Gemma-2-2B achieved the highest test F1-score of 95.99%, followed by Llama-3.2-1B with 94.63% and Qwen3-1.7B with 93.90%. These results indicate strong performance on the evaluated benchmark; however, they should not be interpreted as definitive evidence of superiority over previous studies because differences in dataset versions, preprocessing, data splits, and evaluation protocols may affect the reported results. Furthermore, the consistency of accuracy, precision, and recall values across the three models indicates a balanced distribution of predictions between the “sarcasm” and “news” classes.

These findings demonstrate that small-scale LLMs in the 1–2 billion parameter range can achieve high detection accuracy when paired with structured adaptation strategies such as LoRA and an MLP Classifier Head, offering a computationally viable solution for deployment on constrained hardware.

4. CONCLUSION

This study successfully evaluated the semantic representations of three small-scale LLMs, Qwen3-1.7B-Base, Llama-3.2-1B, and Gemma-2-2B, for the task of classifying sarcasm in news articles. Based on the results of the experiments conducted, it can be concluded that under zero-shot generative conditions, Qwen3-1.7B-Base achieved the highest accuracy of 0.5365 and a macro F1-score of 0.4197 compared to its competitors. However, an in-depth analysis using a confusion matrix revealed that this performance was driven by a strong bias towards the sarcasm class, where the model successfully detected almost all sarcasm data TP = 1,155 but failed to accurately identify normal news FP = 1.070.

Conversely, the Llama-3.2-1B and Gemma-2-2B models exhibited failure patterns that ran counter to the highly conservative classification tendencies observed in the News class. These findings provide empirical evidence that language models with a parameter range of 1B–2B have limitations in pragmatic reasoning without additional adaptation. The use of PEFT techniques such as LoRA with an MLP Classifier Head has proven to be a crucial step in correcting class bias and aligning the model’s understanding with the complex linguistic features of sarcasm.

Limitations and threats to validity: A key limitation of this study lies in the scope of the evaluation dataset. The News Headlines Dataset comprises curated, grammatically structured headlines from professional sources (The Onion and HuffPost). While this minimizes lexical noise and spelling variations, formal headlines lack the informal constructs typical of social media platforms, such as slang, emojis, chaotic punctuation, and conversational context. Consequently, while our LoRA-MLP framework demonstrates strong efficiency on structured text, performance claims regarding "real-world deployment" should be contextualized. Future work must validate this framework on noisy, highly informal social media corpora (e.g., Twitter, Reddit) and multimodal contexts.

As a recommendation, although Qwen3 demonstrates the potential for more nuanced representations in zero-shot scenarios, practical implementation for sarcasm detection on small-scale models still requires a structured fine-tuning phase. Further research could explore the use of more diverse datasets or more specific optimisation of LoRA hyperparameters to balance precision and recall in these efficient models. This study is limited to the use of the News Headlines dataset; future research is expected to explore datasets from social media, which feature a wider variety of informal language. In addition, testing could also be carried out on larger public datasets, using multilingual data, or on other recent small-scale LLM variants to further validate the LoRA-MLP framework.

REFERENCES

[1] Defede, N., Magdaraog, N.M., Thakkar, S.C., Bizel, G. (2021). Understanding how social media is influencing

the way people communicate: Verbally and written. International Journal of Marketing Studies, 13(2): 1-11. https://doi.org/10.5539/ijms.v13n2p1

[2] Ali, R., Farhat, T., Abdullah, S., et al. (2023). Deep learning for sarcasm identification in news headlines. Applied Sciences, 13(9): 5586. https://doi.org/10.3390/app13095586

[3] Annan, A., Eiden, A.L., Wang, D., et al. (2025). Evaluating large language models for sentiment analysis and hesitancy analysis on vaccine posts from social media: Qualitative study. JMIR Formative Research, 9: e64723. https://doi.org/10.2196/64723

[4] Ashwitha, A.S.G.S.H.R., Shruthi, G., Shruthi, H.R., Manjunath, T.C. (2021). Sarcasm detection in natural language processing. Materials Today: Proceedings, 37: 3324-3331. https://doi.org/10.1016/j.matpr.2020.09.124

[5] Avvaru, A., Vobilisetty, S., Mamidi, R. (2020). Detecting sarcasm in conversation context using transformer-based models. In Proceedings of the Second Workshop on Figurative Language Processing, pp. 98-103. https://doi.org/10.18653/v1/2020.figlang-1.15

[6] Bade, G.Y., Kolesnikova, O., Oropeza, J.L., Tash, M.S. (2026). Evaluating the capability of base and large-scale language models for multilingual sarcasm detection. PeerJ Computer Science, 12: e3584. https://doi.org/10.7717/peerj-cs.3584

[7] Bucher, M.J.J., Martini, M. (2024). Fine-tuned'small'LLMs (still) significantly outperform zero-shot generative AI models in text classification. arXiv preprint arXiv:2406.08660. https://doi.org/10.48550/arXiv.2406.08660

[8] Corradini, F., Leonesi, M., Piangerelli, M. (2025). State of the art and future directions of small language models: A systematic review. Big Data and Cognitive Computing, 9(7): 189. https://doi.org/10.3390/bdcc9070189

[9] Dubey, P., Dubey, P., Bokoro, P.N. (2025). Unpacking sarcasm: A contextual and transformer-based approach for improved detection. Computers, 14(3): 95. https://doi.org/10.3390/computers14030095

[10] Filik, R., Țurcan, A., Thompson, D., Harvey, N., Davies, H., Turner, A. (2016). Sarcasm and emoticons: Comprehension and emotional impact. Quarterly Journal of Experimental Psychology, 69(11): 2130-2146. https://doi.org/10.1080/17470218.2015.1106566

[11] Gao, X., Nayak, S., Coler, M. (2025). Spoken in jest, detected in earnest: A systematic review of sarcasm recognition-multimodal fusion, challenges, and future prospects. IEEE Transactions on Affective Computing, 16(4): 2526-2544. https://doi.org/10.1109/TAFFC.2025.3612205

[12] Hassan, M.A., García-Méndez, S., de Arriba-Pérez, F. (2026). Contribution to sarcasm detection in Arabic using natural language processing techniques. Applied Sciences, 16(6): 2724. https://doi.org/10.3390/app16062724

[13] Helal, N.A., Hassan, A., Badr, N.L., Afify, Y.M. (2024). A contextual-based approach for sarcasm detection. Scientific Reports, 14(1): 15415. https://doi.org/10.1038/s41598-024-65217-8

[14] Jin, X., Yang, Y., Wu, Y., Xu, Y. (2024). Research on sarcasm detection technology based on image-text fusion. Computers, Materials & Continua, 79(3): 5225-5242. https://doi.org/10.32604/cmc.2024.050384

[15] Huang, C.W., Lu, H., Gong, H., et al. (2024). Investigating decoder-only large language models for speech-to-text translation. arXiv preprint arXiv:2407.03169. https://doi.org/10.48550/arXiv.2407.03169

[16] Kampelopoulos, D., Tsanousa, A., Vrochidis, S., Kompatsiaris, I. (2025). A review of LLMs and their applications in the architecture, engineering and construction industry. Artificial Intelligence Review, 58(8): 250. https://doi.org/10.1007/s10462-025-11241-7

[17] Liu, B. (2024). Comparative analysis of encoder-only, decoder-only, and encoder-decoder language models. In Proceedings of the 1st International Conference on Data Science and Engineering, pp. 524-530. https://doi.org/10.5220/0012829800004547

[18] Liu, J., Chang, Y., Wu, Y. (2025). R-LoRA: Randomized multi-head LoRA for efficient multi-task learning. In EMNLP (Findings), pp. 660-674. https://doi.org/10.18653/v1/2025.findings-emnlp.35

[19] Lu, M., Dong, Z., Guo, Z., et al. (2025). A multi-modal sarcasm detection model based on cue learning. Scientific Reports, 15(1): 10261. https://doi.org/10.1038/s41598-025-94266-w

[20] McAuley, L., Glenwright, M. (2025). Humor styles predict self-reported sarcasm use in interpersonal communication. Behavioral Sciences, 15(7): 922. https://doi.org/10.3390/bs15070922

[21] Misra, R., Arora, P. (2023). Sarcasm detection using news headlines dataset. AI Open, 4: 13-18. https://doi.org/10.1016/j.aiopen.2023.01.001

[22] Prabhu, J.J., Velur, S.P.R. (2025). Sarcasm detection in conversational contexts: A comprehensive review with a logistic regression baseline study. Premier Journal of Science, 15: 100125. https://doi.org/10.70389/pjs.100125

[23] Mamun, M.B., Tsunakawa, T., Nishida, M., Nishimura, M. (2024). Hate speech detection by using rationales for judging sarcasm. Applied Sciences, 14(11): 4898. https://doi.org/10.3390/app14114898

[24] Pexman, P., Reggin, L., Lee, K. (2019). Addressing the challenge of verbal irony: Getting serious about sarcasm training. Languages, 4(2): 23. https://doi.org/10.3390/languages4020023

[25] Peykani, P., Ramezanlou, F., Tanasescu, C., Ghanidel, S. (2025). Large language models: A structured taxonomy and review of challenges, limitations, solutions, and future directions. Applied Sciences, 15(14): 8103. https://doi.org/10.3390/app15148103

[26] Hu, E.J., Shen, Y., Wallis, P., et al. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. https://doi.org/10.48550/arXiv.2106.09685

[27] Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. (2019). Parameter-efficient transfer learning for NLP. arXiv preprint arXiv.1902.00751. https://doi.org/10.48550/arXiv.1902.00751

[28] Song, C., Zhang, Y., Gao, H., Yao, B., Zhang, P. (2025). Large language models for subjective language understanding: A survey. arXiv preprint arXiv:2508.07959. https://doi.org/10.48550/arXiv.2508.07959

[29] Soliman, G., Zaki, H., Kilany, M. (2025). A comparative analysis of encoder only and decoder only models for challenging LLM-generated STEM MCQs using a self-

evaluation approach. Natural Language Processing Journal, 10: 100131. https://doi.org/10.1016/j.nlp.2025.100131

[30] Gupta, A., Thomas, B., Asnani, H., et al. (2025). Small language models (SLMs) can still pack a punch: A survey (updated 2026). arXiv preprint arXiv:2501.05465. https://doi.org/10.48550/arXiv.2501.05465

[31] Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 technical report. arXiv preprint arXiv:2505.09388. https://doi.org/10.48550/arXiv.2505.09388

[32] Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783

[33] Team, G., Riviere, M., Pathak, S., et al. (2024). Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118. https://doi.org/10.48550/arXiv.2408.00118

[34] Roy, M.C., Bisoy, S.K., Sahoo, P.K., Kumawat, G. (2026). Advancing news headline sarcasm detection through hybrid neural networks. Discover Computing, 29(1): 12. https://doi.org/10.1007/s10791-025-09877-8