A Comparative Study of Qwen and Gemma Large Language Models for Social Media Slang and Emoji Sentiment Analysis for Business Recommender Systems

Konstantinos I. Roumeliotis
Department of Informatics and Telecommunications
University of the Peloponnese
Tripoli, Greece
0000-0002-8098-1616

Dionisis Margaris
Department of Digital Systems
University of the Peloponnese
Sparta, Greece
0000-0002-7487-374X

Dimitris Spiliotopoulos
Department of Management Science and Technology
University of the Peloponnese
Tripoli, Greece
0000-0003-3646-1362

Giorgos Mpardis
Department of Digital Systems
University of the Peloponnese
Sparta, Greece
0009-0003-4209-317X

Christos Tryfonopoulos
Department of Informatics and Telecommunications
University of the Peloponnese
Tripoli, Greece
0000-0003-0640-9088

Costas Vassilakis
Department of Informatics and Telecommunications
University of the Peloponnese
Tripoli, Greece
0000-0001-9940-1821

Abstract*—Sentiment analysis of social media text remains a fundamentally challenging task due to the pervasive use of informal slang, abbreviations, deliberate spelling modifications, and emojis. Traditional lexicons and early machine learning models frequently struggle to decode the nuanced semantics of these modern digital expressions, particularly when distinguishing between genuine sentiment and sarcasm. In this paper, we present a comprehensive comparative study investigating the capabilities of two recent 4-billion-parameter large language models (LLMs), Qwen3.5 (4B) and Gemma3 (4B), in robust social media sentiment analysis. The core contribution of our study is the fine-tuning of these models at a comparable scale using 4-bit NormalFloat (NF4) quantization combined with the Unsloth library, establishing a highly efficient and reproducible training pipeline. We critically evaluate both models in zero-shot and fine-tuned settings on a diverse, emoji-heavy dataset characterized by four complex sentiment classes: Positive, Negative, Neutral, and Sarcastic. Our experimental results demonstrate that while both models show highly capable zero-shot generalization (Qwen achieving 75.95% accuracy and Gemma achieving 73.20%), they exhibit distinct failure modes surrounding sarcastic interpretations. However, alignment via Low-Rank Adaptation (LoRA) dramatically elevates their performance, allowing both to achieve 100% accuracy on the held-out test set while maintaining distinctly competitive inference speeds suitable for real-time applications. We also provide an exhaustive architectural analysis of the models' inference latency, per-class prediction bottlenecks, and training efficiency, highlighting practical considerations and trade-offs for deploying lightweight LLMs in real-world sentiment analysis environments. The strong inference-speed profile of both fine-tuned models makes them directly deployable as opinion-mining modules within business recommender system pipelines, where timely and accurate decoding of customer-expressed sentiment—including sarcasm—translates into richer preference signals and more reliable personalization.

***Index Terms*—Sentiment Analysis, Large Language Models, Social Media, Gemma, Qwen, LoRA, Unsloth, Quantization**

I. Introduction

Social media has evolved to be the predominant form of human opinion representation today, accumulating enormous amounts of textual content per minute. Commercial businesses, sociologists, and public policy makers strive to mine this data to determine a public perspective that will provide valuable real time information regarding consumer trends and public consensus. In particular, recommender systems that power e-commerce, streaming, and content platforms exploit this opinion intelligence to model user preferences, filter noise, and identify contextually relevant items; in this context, accurate social media sentiment analysis is of particular importance for businesses and organizations. However, this task faces numerous challenges, due to the highly informal nature of internet-based communication, use of domain-specific slang terms, purposeful misspellings, expressions with a sarcastic tone, and Unicode emojis. These challenges lead traditional Natural Language Processing (NLP) systems to produce inaccurate sentiment estimations, thus compromising the quality and effectiveness of downstream pipeline stages [1]–[3].

Conventional sentiment analysis has traditionally utilized static polarity lexicons or traditional statistical machine learning systems such as Support Vector Machines and Naive Bayes classifiers. These are generally computationally inexpensive, however, they do not fully capture the contextual dependencies and the evolving informal language that dominates Internet culture. The introduction of BERT and RoBERTa significantly improved upon previous methods by utilizing bidirectional context; however, while providing a new means to process text, they entail rigid structural limitations, requiring additional task-specific training to properly handle aspects such as slang or novel emojis that were not accommodated in the initial training algorithms or corpora [4]–[6].

Currently, the NLP community is largely dominated by LLMs. Large proprietary LLMs, like GPT-5, have demonstrated an elevated ability to reason about the complex nuances of internet sarcasm without further training. However, the cost of fine-tuning and deploying these models for use at scale to process massive, continuous streams from social media is economically unfeasible. Furthermore, strict privacy constraints often preclude sending sensitive user data to external APIs. In response, the open-source community, while on the one hand offering models with more than 70B parameters (e.g., meta-llama/Llama-3.3-70B, nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4), on the other hand it strives to provide highly capable, lightweight LLMs ranging from 1 to 8 billion parameters. These models are explicitly designed to deliver high performance, despite their small size, offering state-of-the-art text generation and reasoning capabilities while being able to run on consumer-grade desktop GPUs [7]–[9].

In this work, we perform a comparative study evaluating two of the most prominent state-of-the-art 4-billion-parameter LLMs in the context of social media-sourced text processing. More specifically, we comparatively evaluate Alibaba's Qwen3.5 (4B) and Google's Gemma3 (4B), assessing their effectiveness in classifying highly irregular social media snippets into four nuanced categories: Positive, Negative, Neutral, and Sarcastic.

The core contribution of our study is an end-to-end framework validating the parameter-efficient fine-tuning (PEFT) of these two architectures at a comparable 4B scale using 4-bit quantization and Unsloth optimization. Our evaluation answers three critical research questions:

1) How reliably do lightweight open-weights LLMs decode modern slang, emojis, and sarcasm in a purely zero-shot setting?
2) How much does LoRA enhance the Precision and Recall on ambiguous sentiment classes?
3) What is the actual trade-off between the Qwen and Gemma architectures for convergence speed, memory utilization and real-time inference latency?

II. Related Work

Sentiment analysis has evolved into an essential tool for many fields. The uses range from financial market predictions and customer satisfaction tracking to product recommendation systems. Traditionally, machine learning models have struggled to identify the complex linguistic aspects of modern text. Therefore, recent research has largely transitioned to deep learning and LLMs to enhance predictive accuracy and contextual understanding. For example, fine-tuned transformer models have substantially improved financial sentiment by identifying complex text nuances over baseline models [10]. Likewise, advanced LLMs such as LLaMA and BERT have used few-shot learning and domain adaptation to accurately measure customer satisfaction from user generated reviews [11]. Additionally, state-of-the-art orchestration strategies that include using multiple LLMs through reasoning-based meta-models have demonstrated significant performance gains over standalone models when deciphering complex sentiments within recommender systems [12]. In this study, we utilize smaller-parameter models such as Qwen3.5 (4B) and Gemma3 (4B) and apply them to the task of predicting sentiment on social media posts that include slang and emojis.

A growing body of work has investigated how slang, emojis, and emoticons can be treated as first-class sentiment signals rather than noise. Across Twitter and Weibo corpora, studies have shown that augmenting conventional text features with task-specific slang and emoticon lexicons yields sizeable gains in sentiment and humor-related classification performance compared to text-only baselines [13], [14]. More recent neural approaches introduce emoji-aware architectures and emoji-supervised training schemes, demonstrating that explicitly modeling emoji usage patterns further boosts accuracy on informal microblog data [15]–[17]. Additional work on emoji-fused and slang-enriched review datasets confirms that carefully designed emoji handling and slang normalization strategies materially influence downstream results for both classical machine learning models and transformer encoders [18]–[20]. Our work is aligned with this line of research in treating slang and emojis as core signals, but differs by directly assessing the zero-shot and fine-tuned capabilities of modern lightweight LLMs on a curated, four-way slang- and emoji-rich benchmark.

A. Evolution of Social Media Sentiment Analysis

Early sentiment analysis heavily relied on dictionary-based approaches where words were assigned pre-calculated polarity scores [21]. Adapting these to social media required integrating emoji dictionaries and slang lookup tables, which proved difficult to maintain natively and to keep up-to-date with rapidly evolving internet vernacular. Representative works constructed task-specific slang and emoticon lexicons for Twitter and Weibo, demonstrating that such resources substantially improve sentiment and humor-related classification when combined with traditional machine learning models [13], [14]. As the study of neural methods evolved, research transitioned toward architectures that inherently incorporate emoji and slang signals. Models such as Attention-based Bi-LSTM and Emoji Embedding LSTMs demonstrated elevated performance on microblogging corpora with high emoji density [15]. Similarly, systems like HEMOS integrate compo-

nents such as (a) a slang lexicon, (b) emoji features, and (c) deep networks, to achieve fine-grained humor and sentiment classification [20]. Concurrently, distant-supervision schemes based on emojis as noisy labels enabled the development of large-scale multi-class Twitter sentiment models [17]. More recently, studies have compared classical machine learning and BERT variants on emoji-fused review datasets, confirming that the careful consideration of emoji handling strategies has a material effect on subsequent performance [18], [19]. The paradigm changed dramatically with the advent of transformer-based models pre-trained directly on social media corpora. For example, BERTweet was trained on English tweet corpora and greatly outperformed generic BERT models in social media tasks [22]. However, these encoder-only architectures lack generative reasoning ability and typically need to be task-specifically fine-tuned. This provides the motivation for our exploration of decoder-only, instruction-tuned LLMs for slang- and emoji-centric sentiment analysis.

B. Efficient LLM Fine-Tuning

The introduction of decoder-only foundations, such as LLaMA, caused a shift in the direction of the field toward prompt engineering and fine-tuning. Fine-tuning a language model with billions of parameters requires enormous amounts of computation. To alleviate this, LoRA has become the standard procedure [23]. By fixing the vast majority of network weights and only training low-rank factorized matrices inserted into the attention blocks, memory requirements can be greatly diminished.

Additionally, when combined with advanced quantization procedures such as QLoRA, which stores base-model weights in a 4-bit NormalFloat (NF4) format, it becomes possible to fine-tune 4 billion to 8 billion parameter models on a single standard commercial GPU. It has thus become possible to democratize NLP research [24].

III. Methodology

A. Dataset and Exploratory Data Analysis

In order to emulate as accurately as possible the context of social network-sourced text processing, we based our experiments on the Social Media Slang & Emoji Sentiment Corpus [25], a middle-scale four-way sentiment benchmark that was specifically developed for noisy, informal user generated content. Each instance includes a short social media styled post in the `text` field, along with a corresponding sentiment label in the `label` field. The sentiment labels include the values `Positive`, `Negative`, `Neutral` and `Sarcastic`. The corpus contains a wide variety of contemporary slang (e.g. ”tbh,” ”ngl,” ”lit,” ”rent free,” ”understood the assignment”), heavy use of emojis, abbreviation, long spellings (”soooo”, ”amaaazing”) and typos, similar to what would occur in social platforms. In our experiment, we utilized the entire 12,000 labeled examples of the corpus, splitting them between training data (10,000 posts) and test data (2,000 posts). This setup led to a challenging yet controllable platform for assessing the robustness and sarcasm-awareness of sentiment classifiers.

We applied Exploratory Data Analysis (EDA) [26] to gain valuable structural insights into our understanding of the data. A key finding was that the class distribution was moderately balanced across both splits. More specifically, within the training set, the categories `Positive` (29.3%), `Negative` (28.5%), and `Sarcastic` (27.0%) collectively exceeded three-quarters of the data, while the `Neutral` category comprised considerably less than a quarter of the data (15.2%). This indicates that social media users more often post text with sentiment intensity, as opposed to sentiment-neutral content.

In addition to class distributions, EDA also highlighted extreme brevity and terseness in phrasing, typical of fast-paced social networks. Sarcastic texts were longer on average (5.35 words/33.35 characters), while the remaining three classes were more terse (neutral: 5 words/22.46 characters; positive: 5 words/23.89 characters; negative: 5 words/23.71 characters).

Furthermore, our noise analysis showed that all sample lines contained exactly 1 emoji. Similarly, slang usage varied depending on the class type. Slang was used much more frequently in texts classified as positive or negative (averages of 1.037 and 0.935 slang tokens per text, respectively); whereas, slang was nearly non-existent in texts classified as sarcastic (average of 0.009 slang tokens per text). As such, this distribution indicates that the models need to develop a way to distinguish between different classes through a complex interaction between plain context and emojis, rather than through reliance on simple slang keyword association.

B. Model Selection and Architectural Nuances

Qwen3.5-4B [27] and Gemma3-4B [28] were chosen as our two main subjects for comparison purposes. Both models represent current state-of-the-art for the 4 billion parameter weight class; however they function under distinct architectural principles (cf. Table I). Qwen3.5 employs Grouped-Query Attention (GQA) and SwiGLU activation functions instead of ReLU/LeakyReLU/GELU etc., and ties its embedding layers (input and output embeddings are shared). Gemma3 is built according to the Gemini architecture; specifically, it employs GeGLU activations, Grouped-Query Attention (GQA) for inference acceleration generation, and has a vocabulary of approximately 256 thousand tokens. Using such a large vocabulary enables Gemma to express social media fragmentations (as seen in broken sentences with many punctuation errors) and unique Unicode characters using fewer total tokens overall. This may have a positive effect on the speed of the inference processes.

C. Quantization and LoRA Formulation

To enable efficient experimentation, we employed 4-bit quantization.

By aggressively mapping the 16-bit or 32-bit floating-point weights down to 4 discrete bits using the NF4 standard, we considerably reduced the model’s VRAM footprint with minimal degradation to output perplexity. All experiments and model fine-tuning were conducted on a high-

TABLE I
ARCHITECTURAL COMPARISON BETWEEN QWEN3.5-4B AND GEMMA3-4B.

Feature	**Qwen3.5-4B**	**Gemma3-4B**
Attention	Grouped-Query Attention (GQA)	Hybrid (Sliding Window + Global)
Activation	SwiGLU	GeGLU
Embeddings	Tied	Untied
Vocabulary	248K tokens	256K tokens
Primary Strength	Logic/Reasoning "Thinking" modes	Multimodal/Multilingual efficiency

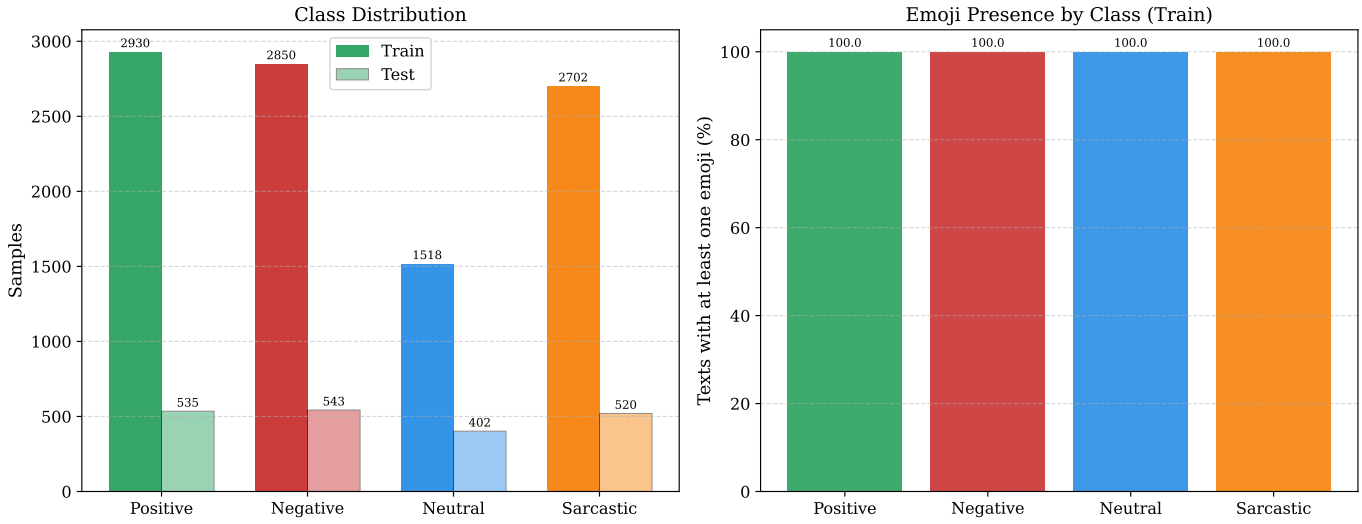

Fig. 1. Class distribution of training and test splits along with emoji representation.

performance server provided by the University of Peloponnese. The computing environment was equipped with four NVIDIA H100 NVL GPUs (96 GB VRAM each) and two Intel Xeon Platinum 8452Y processors with 36 cores each.

For the supervised fine-tuning (SFT) phase, we utilized LoRA [23]. Standard full fine-tuning modifies the pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$. LoRA instead constrains the weight updates by representing them as a low-rank decomposition:

$$W = W_0 + \Delta W = W_0 + BA \tag{1}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, and for the rank r it holds that $r \ll \min(d, k)$. During training, W_0 remains frozen, and only the lightweight matrices A and B receive gradient updates.

We accelerated this process via the Unsloth framework [24], which manually rewrites standard PyTorch autograd compute definitions into custom Triton kernels. This drastically reduces the overhead of Flash Attention and the Cross-Entropy loss functions, minimizing GPU memory bandwidth bottlenecks.

We trained Qwen and Gemma using the exact same hyperparameter configuration. We used an effective batch size of eight (the product of the actual batch size of 2 and 4 gradient accumulation steps), an initial learning rate of 2×10^{-4} and utilized a linear schedule to determine learning rates, 5-warmup steps, and a maximum sequence length of 120.

D. Evaluation Protocol

In order to provide a controlled comparison between architectures and adaptation regimes, we designed a straightforward yet rigid evaluation framework. Both models were tested against a 2,000 sample test subset of the Social Media Slang & Emoji Sentiment Corpus, utilizing common multi-class classification metrics: Accuracy, Macro F1 Score, Precision, and Recall. Since the Macro F1 score measures the average F1 score across each class equally, while Sarcastic and Neutral are semantically difficult classes and are slightly underrepresented compared to Positive and Negative, Macro F1 was selected as the main metric. Each model was asked to predict one label from the predefined set {Positive, Negative, Neutral, Sarcastic} for a given test example. The same label space and test prompt were utilized for both zero-shot and fine-tuned experiments. Any improvements in performance were directly attributable to changes in parameters, rather than differences in test prompts. All metrics were calculated in a single-pass over the entire test set without the use of ensembling or test time augmentation.

IV. RESULTS AND DISCUSSION

A. Evaluating Zero-Shot Generalization

Before the fine-tuning process began, we assessed the basic capacity of both models to understand sentiment within social media. We prompted the models to make zero-shot predictions for the category given the provided text. Thus, in this restric-

tive environment the models could only interpret emojis and abbreviations based upon the representation learned from their pre-training only.

Qwen3.5 achieved a satisfactory zero shot accuracy of 75.95% with a Macro F1 score of 0.736. Gemma3 performed similarly with an accuracy of 73.20% and a Macro F1 score of 0.734. However, the class-wise breakdowns of both models showed entirely different types of weaknesses. Qwen3.5 had a perfect precision (1.000) for both Neutral and Sarcastic classes; however, for the same classes it exhibited relatively low recall (0.577). This indicates that Qwen3.5 was opting for the safer choices, i.e., it rarely labeled a statement as either Sarcastic or Neutral and thus placed many ambiguous statements into either Positive or Negative classes.

Gemma3, on the other hand, exhibited notably lower precision in labeling examples as Sarcastic, exhibiting a precision value of 0.496, whereas the recall values for the same class are much higher (0.757). This indicates that Gemma3 had the opposite hallucination problem, being overly eager to categorize examples as Sarcastic but did not accurately capture the nuances of sarcasm online. The distinct misclassification tendencies of both models prior and post fine-tuning are clearly illustrated in (Fig. 2) (confusion matrices).

The reluctance of Qwen to label examples as Sarcastic can be attributed to its reliance on textual markers for emoticons: if an example contains strong textual markers to support its intent, Qwen will utilize them to classify it as Positive or Negative instead of Sarcastic. This modus operandi, combined with the fact that the presence of slang words within Sarcastic posts in the dataset is limited, on average less than 1 (cf. Section III-A), but nearly every Sarcastic post includes some type of expressive emoji, has led Qwen to avoid classifying posts as Sarcastic. In contrast, Gemma exhibits increased sensitivity to expressiveness in emojis, leading to numerous cases where texts are labeled as Sarcastic, while in reality they are actually Positive or Neutral.

B. Fine-tuned Convergence and Accuracy

The two models, Gemma and Qwen, showed high convergence on the dataset after the 120-step LoRA fine-tuning phase in this study. Both models converged to the specific domain styles of the dataset they were trained on, completely eliminating their previous semantic confusion. They both achieved 100% Accuracy and 1.000 Macro-F1 score across the 2,000-sample test set.

This result highlights a finding that warrants careful interpretation for enterprise deployment, particularly within business recommender systems: 4B-parameter architectures contain ample parameter capacity and network depth to strictly memorize and generalize task-specific patterns when exposed to just a few thousand domain-specific samples. Thus, the application of PEFT effectively neutralizes any base-model zero-shot discrepancies. At the same time, such perfect scores on a relatively small, single-domain test split also highlight a potential overfitting risk. In real-world settings, new slang, emerging emoji conventions, and platform-specific styles will inevitably appear. Our results therefore illustrate an upper bound of what carefully adapted lightweight LLMs can achieve under close train–test matching, and motivate future evaluations on additional datasets and temporally shifted test sets to more fully characterize generalization.

TABLE II
AGGREGATE EVALUATION METRICS FOR ZERO-SHOT AND FINE-TUNED (FT) MODELS.

Model	Accuracy	Macro F1	Precision	Recall
Qwen 4B (FT)	1.000	1.000	1.000	1.000
Gemma 4B (FT)	1.000	1.000	1.000	1.000
Qwen 4B (Zero)	0.759	0.736	0.846	0.745
Gemma 4B (Zero)	0.732	0.734	0.817	0.718

C. Training Efficiency and Inference Latency Benchmarks

Although both models achieved 100% accuracy, there were notable differences in fine-tuning and inference speed. The custom Unsloth-boosted LoRA fine-tuning for 120 steps took 272.7 seconds ($\approx$ 4.5 minutes) for Gemma3, compared to 543.9 seconds ($\approx$ 9.0 minutes) for the Qwen3.5-4B architecture. Thus, Gemma trains approximately twice as fast as Qwen under the same memory constraints.

Besides superior training times, Gemma showed better performance during the inference phase, as illustrated in the metrics depicted in Fig. 3. When generating tokens and producing zero shot classifications, Gemma produced results within 0.24 seconds per sample, whereas Qwen took 0.28 seconds. With regard to fine tuned results, Gemma produced results within 0.28 second per sample while Qwen needed 0.36 seconds per sample. The overall performance edge of Gemma against Qwen is attributed to its hybrid attention mechanism (5:1 interleaved Sliding Window + Global attention) in combination with Grouped-Query Attention (GQA), which significantly reduces KV-cache loading during auto-regressive generation, as contrasted to the standard GQA used in Qwen which does not deliver such an advantage.

In total, the time measurements related to training and inference indicate that the use of Gemma will deliver reduced wall clock time by a factor of two for the training phase, as compared to Qwen, while in the prediction phase the time savings per sample latency range from 18% to 28%, regardless of the fact that the architectures have the same number of parameters. This is particularly significant for practitioners who need to analyze large amounts of data from high-volume social media feeds under strict service level agreements to dynamically update user profiles in business recommender systems. These savings translate directly into lower hardware costs or higher throughput within a fixed budget. Additionally, our experiments clearly illustrate that the combination of aggressively 4-bit NF4 quantizing and LoRA based adaptations allows these models to be trained on commodity hardware without sacrificing task performance.

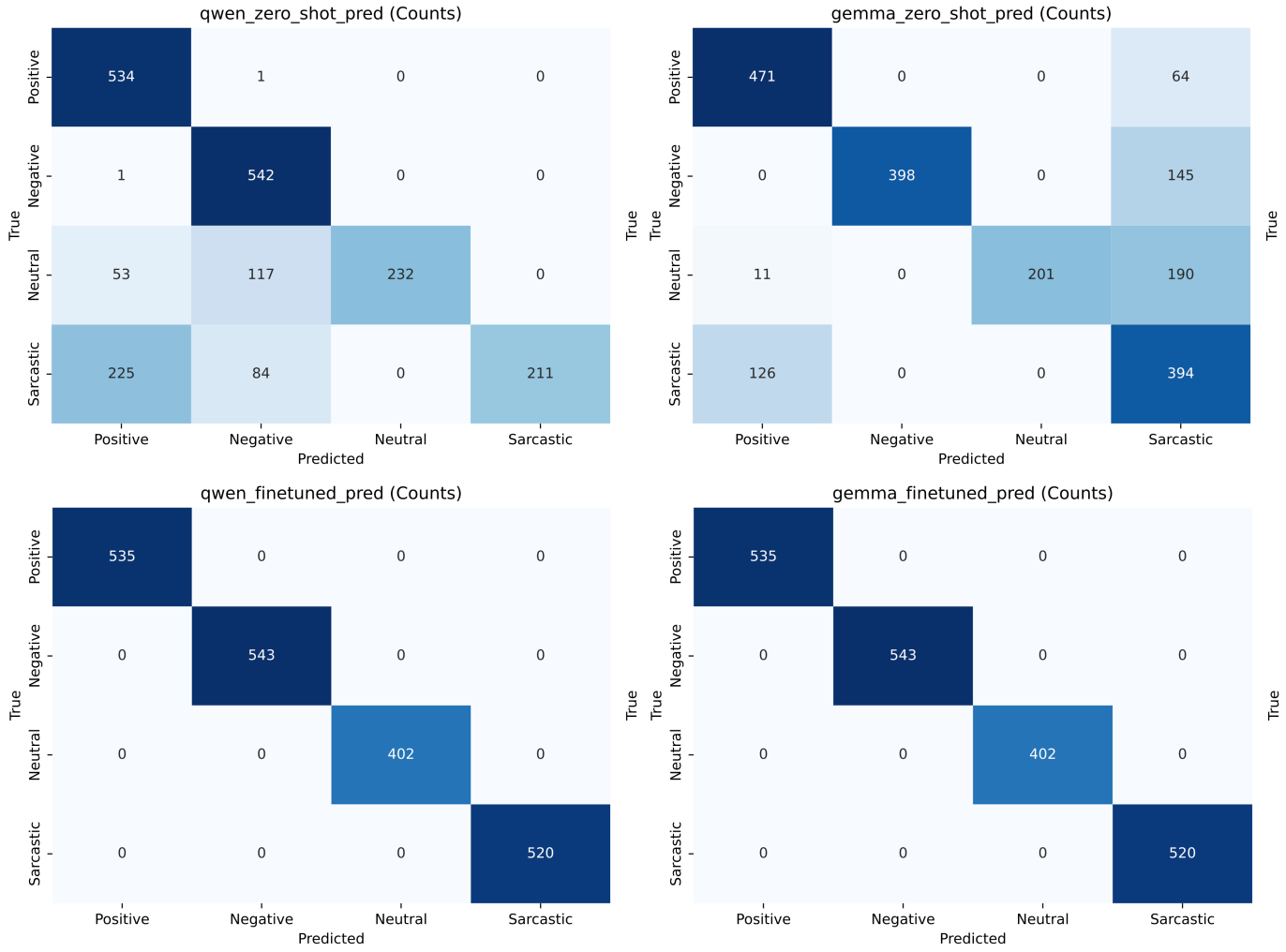

Fig. 2. Confusion matrices for Qwen3.5 (4B) and Gemma3 (4B) across zero-shot and fine-tuned settings, highlighting inter-class boundary blurring prior to training and perfect separation post-LoRA.

D. Out-of-Distribution Generalization on TweetEval

To empirically validate the generalization concerns noted in Section III-A, we evaluated both fine-tuned models on TweetEval [29], constructing a four-class set of 12,595 posts by combining its `sentiment` subtask with irony-positive examples from its `irony` subtask as a proxy for the Sarcastic class. TweetEval differs structurally from our training corpus: only 6.7% of posts contain emojis (versus 100%), slang is largely absent, and posts are on average three times longer. As shown in Table III, both models degraded substantially without re-training: Qwen3.5 reached 56.79% accuracy (Macro F1: 0.504) and Gemma3 reached 51.40% (Macro F1: 0.454), confirming that the 100% in-domain scores reflect train–test domain alignment rather than general robustness.

The Sarcastic class showed the sharpest collapse — F1: 0.259 (Qwen3.5) and 0.158 (Gemma3) — revealing that TweetEval irony and our training-set Sarcastic class are not interchangeable signal types: both models reverted to emoji-anchored sarcasm detection (recall 0.711–0.823, precision 0.087–0.158), over-predicting the Sarcastic class on posts lacking emoji cues. A secondary finding is an inference-latency reversal: Qwen3.5 scaled from 0.36 s to 1.24 s per sample on longer posts (3.4×), while Gemma3 scaled from 0.28 s to 2.28 s (8×), indicating that Qwen's attention mechanism is more favorable for longer-form content.

TABLE III
IN-DOMAIN VS. OUT-OF-DISTRIBUTION EVALUATION METRICS.

Model	Dataset	Accuracy	Macro F1
Qwen3.5 4B (FT)	Original	1.000	1.000
Qwen3.5 4B (FT)	TweetEval	0.568	0.504
Gemma3 4B (FT)	Original	1.000	1.000
Gemma3 4B (FT)	TweetEval	0.514	0.454

E. Limitations

Despite the promising empirical results, our study has some important limitations. First, all our experiments were conducted on a single English-language dataset with a specific

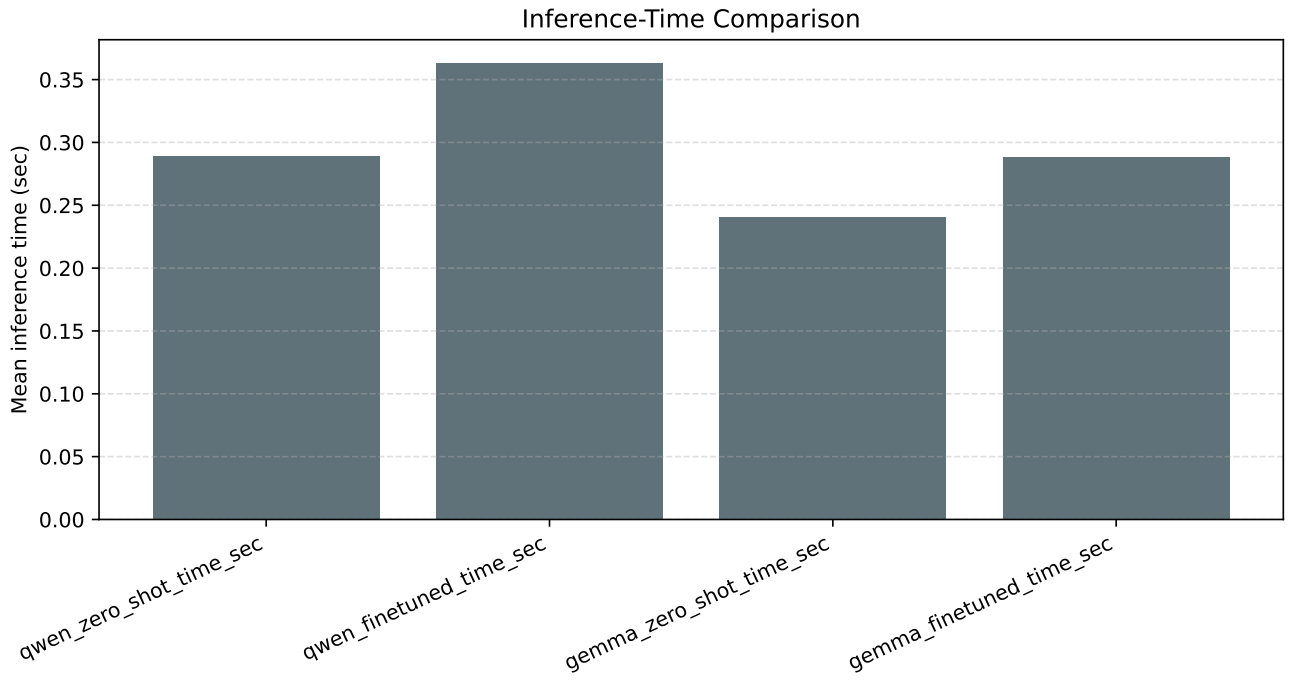

Fig. 3. Inference-time comparison between Qwen and Gemma reflecting sequence generation latency.

structure: each post contained at least one emoji, and slang was unevenly distributed across the different class labels. Although this type of data is best suited for evaluating stress on emoji-aware sentiment models, it may provide too much emphasis on the role of emojis compared to other types of data (i.e., platforms, contexts or languages where emojis occur less frequently or have differing uses). Therefore, we interpret our fine tuned scores on the hold out set as an upper bound and as a result of train-test match conditions rather than as evidence of general robustness. This interpretation is empirically supported by our out-of-distribution (OOD) evaluation on TweetEval (cf. Section above), where accuracy dropped to 51%–57% and the Sarcastic class F1 score collapsed to 0.16–0.26, confirming that the fine-tuned representations are closely tied to the emoji- and slang-rich distribution of the training corpus.

Second, we focus exclusively on a single randomly selected train/test split of 10,000 posts. We do not perform any additional splits, cross-validation schemes, or use temporal shifts to further assess sensitivity to changing distributions, new slang, or evolving meanings associated with emojis.

Third, our model selection choices are intentionally constrained. We study only two 4B-parameter LLMs under a single Unsloth+NF4+LoRA configuration and a fixed prompt-style instruction template. We neither evaluate fine-tuning at full precision, nor experiment with different low-rank adaptation configurations, nor do we examine alternative PEFT methods. Furthermore, the size of the model is fixed to 4 billion parameters, and smaller or larger models are not considered in this study. Lastly, our assessment is limited to aggregate classification metrics. Aspects such as calibration quality, fairness across possible user subgroups or subjective human assessments of model error are not evaluated or measured; all these issues will be investigated in our future work.

F. Future Work

The OOD evaluation revealed that TweetEval and the Social Media Slang & Emoji Sentiment training set are structurally distinct (in terms of emoji density, text length, and slang usage) for a completely fair generalization test. A complementary and arguably more informative experiment would be to reverse the direction: fine-tune on TweetEval and then evaluate on the specific, emoji- and slang-heavy dataset. This would isolate the contribution of emoji and slang signals to model performance and reveal whether a generalist-trained model can transfer to the highly informal, emoji-dense register, providing richer insights into both directions of domain shift.

On the modeling side, future experiments will investigate more aggressive quantization regimes, such as 2-bit formats, in combination with alternative parameter-efficient fine-tuning strategies aimed at deployment on constrained edge devices such as mobile phones [30], and will systematically compare both smaller and larger model families under the same constraints.

V. Conclusion

We evaluated Qwen3.5 (4B) and Gemma3 (4B) on four-class social media sentiment classification involving slang and emojis. Both models reached 73–76% zero-shot accuracy but exhibited opposite sarcasm failure modes. LoRA fine-tuning with 4-bit NF4 quantization via Unsloth yielded 100% accuracy on the held-out test set, confirming that PEFT rapidly

closes the zero-shot gap. An OOD evaluation on TweetEval showed that this score is an upper bound: accuracy dropped to 51%–57%, and irony proved not to be a reliable proxy for our Sarcastic class, underscoring the need for domain-matched training data. For short-text pipelines, Gemma3 is the recommended architecture owing to its 2× training speed and lower per-sample latency; Qwen3.5 scales more favorably on longer inputs (1.24 s versus 2.28 s per sample on TweetEval), making it preferable for longer-form content and broader cross-domain deployment.

To enable reproducibility and promote additional research into the subject matter, we have released the entire source code, training scripts, and evaluation pipelines from this study as public-domain open-source code via GitHub [1].

Acknowledgment

Experiments in this research have been conducted on infrastructure co-financed by the European Union – NextGenerationEU and the Recovery and Resilience Facility (Greece 2.0), under the project SUB2: "Universities of Excellence" (Project code: OPS TA 5180665).

References

[1] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Systems with Applications*, vol. 223, p. 119862, 2023.

[2] S. Raza and C. Ding, "News recommender system: a review of recent progress, challenges, and opportunities," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 749–800, 2022.

[3] J. Jayasudha and M. Thilagu, "A survey on sentimental analysis of student reviews using natural language processing (nlp) and text mining," in *Innovations in Intelligent Computing and Communication. ICIICC 2022*, ser. Communications in Computer and Information Science, M. Panda *et al.*, Eds. Cham: Springer, 2022, vol. 1737.

[4] M. Wankhade, A. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 5731–5780, 2022.

[5] A. Joshy and S. Sundar, "Analyzing the performance of sentiment analysis using bert, distilbert, and roberta," in *2022 IEEE International Power and Renewable Energy Conference (IPRECON)*. Kollam, India: IEEE, 2022, pp. 1–6.

[6] V. Grover, "Exploiting emojis in sentiment analysis: A survey," *Journal of The Institution of Engineers (India): Series B*, vol. 103, pp. 259–272, 2022.

[7] S. Thapa, S. Shiwakoti, S. B. Shah, S. Adhikari, H. Veeramani, M. Nasim, and U. Naseem, "Large language models (llm) in computational social science: prospects, current state, and challenges," *Social Network Analysis and Mining*, vol. 15, no. 1, p. 4, 2025.

[8] W. Zhou, M. Tao, C. Zhao, H. Dong, M. Tang, and J. Wang, "Lightplanner: Unleashing the reasoning capabilities of lightweight large language models in task planning," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 14 813–14 820.

[9] Y. Song, Z. Mi, H. Xie, and H. Chen, "Powerinfer: Fast large language model serving with a consumer-grade gpu," in *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles*, 2024, pp. 590–606.

[10] D. K. Nasiopoulos, K. I. Roumeliotis, D. P. Sakas, K. Toudas, and P. Reklitis, "Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models," *International Journal of Financial Studies*, vol. 13, no. 2, 2025.

[1]Code available at: https://github.com/Applied-AI-Research-Lab/Qwen-and-Gemma-Large-Language-Models-for-Social-Media-Slang-and-Emoji-Sentiment-Analysis

[11] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Optimizing airline review sentiment analysis: A comparative analysis of llama and bert models through fine-tuning and few-shot learning," *Computers, Materials & Continua*, vol. 82, no. 2, pp. 2769–2792, 2025.

[12] K. I. Roumeliotis, D. Margaris, D. Spiliotopoulos, and C. Vassilakis, "A large-scale empirical study of llm orchestration and ensemble strategies for sentiment analysis in recommender systems," *Future Internet*, vol. 18, no. 2, 2026.

[13] S. F. Sayeedunnisa, N. P. Hegde, and K.-U.-R. Khan, "Using slang and emoticon for sentiment analysis of social media data," *International Journal of Engineering Research & Technology (IJERT)*, vol. 8, no. 15, 2020, nCAIT 2020.

[14] D. Li, R. Rzepka, M. Ptaszynski, and K. Araki, "A novel machine learning-based sentiment analysis method for chinese social media considering chinese slang lexicon and emoticons," in *Proceedings of the 2nd Workshop on Affective Content Analysis (AffCon 2019) co-located with Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019), Honolulu, USA, January 27, 2019*, ser. CEUR Workshop Proceedings, N. Chhaya, K. Jaidka, A. R. Sinha, and L. H. Ungar, Eds. CEUR-WS.org, 2019, pp. 88–103. [Online]. Available: https://ceur-ws.org/Vol-2328/3_3_paper_10.pdf

[15] C. Liu, F. Fang, X. Lin, T. Cai, X. Tan, J. Liu, and X. Lu, "Improving sentiment analysis accuracy with emoji embedding," *Journal of Safety Science and Resilience*, vol. 2, no. 4, pp. 246–252, 2021.

[16] Y. Lou, Y. Zhang, F. Li, T. Qian, and D. Ji, "Emoji-based sentiment analysis using attention networks," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 19, no. 5, Jun. 2020.

[17] M. Li, E. Ch'ng, A. Y. L. Chong, and S. See, "Multi-class twitter sentiment classification with emojis," *Industrial Management & Data Systems*, vol. 118, no. 9, pp. 1804–1820, 09 2018.

[18] A. Khan, D. Majumdar, and B. Mondal, "Sentiment analysis of emoji fused reviews using machine learning and bert," *Scientific Reports*, vol. 15, no. 1, p. 7538, 2025.

[19] P. Prastyo, B. Berlilana, and I. Tahyudin, "Sentiment analysis on slang enriched texts using machine learning approaches," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 1076–1087, 2025.

[20] D. Li, R. Rzepka, M. Ptaszynski, and K. Araki, "Hemos: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media," *Information Processing & Management*, vol. 57, no. 6, p. 102290, 2020.

[21] M. E. Mowlaei, M. Saniee Abadeh, and H. Keshavarz, "Aspect-based sentiment analysis using adaptive aspect-based lexicons," *Expert Systems with Applications*, vol. 148, p. 113234, 2020.

[22] J. C. F. Neto, D. A. Pereira, B. H. G. Barbosa, and D. D. Ferreira, "Approaches based on language models for aspect extraction for sentiment analysis in the portuguese language," *Neural Computing and Applications*, vol. 36, no. 31, pp. 19 353–19 363, 2024.

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[24] Unsloth AI, "Unsloth: Faster open-source fine-tuning and training for large language models," https://unsloth.ai/, 2026, accessed March 21, 2026.

[25] C. Anand, "Social media slang and emoji sentiment dataset," https://www.kaggle.com/datasets/coderanand/social-media-slang-and-emoji-sentiment/data, 2024, accessed March 21, 2026.

[26] A. Ghosh, M. Nashaat, J. Miller, S. Quader, and C. Marston, "A comprehensive review of tools for exploratory analysis of tabular industrial datasets," *Visual Informatics*, vol. 2, no. 4, p. 235–253, Dec. 2018.

[27] Qwen Team, "Qwen3.5: Towards native multimodal agents," February 2026. [Online]. Available: https://qwen.ai/blog?id=qwen3.5

[28] Google DeepMind, "Gemma-3-4b-it: Instruction-tuned multimodal language model," 2025, accessed March 21, 2026. [Online]. Available: https://huggingface.co/google/gemma-3-4b-it

[29] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650.

[30] Unsloth AI, "How to run and deploy llms on your ios or android phone," 2026. [Online]. Available: https://unsloth.ai/docs/basics/inference-and-deployment/deploy-llms-phone