

Fine-tuning TinyLLaVA for Multimodal Sarcasm Detection on MMSD 2.0

Juan E. Alvarado

The University of Texas at Dallas

mrjuan.alvarado94@gmail.com

Unpublished manuscript. This work was completed during an NSF Research Experiences for Undergraduates (REU) program at The University of Texas at Dallas (Summer 2025). It has not been peer reviewed or published and is shared here as a personal record of my research during that period.

Abstract—Multimodal visual language models (VLMs) leverage multiple data modalities to better interpret the content and context of an image. With the rapid advancement of neural networks in subsequent years, the capabilities of VLMs have been increasingly leveraged for natural language tasks such as sarcasm detection. Sarcasm is a form of verbal irony, a rhetorical device in which a speaker expresses meaning that is the opposite of the literal meaning of their words. Typically, this is done to convey criticism, mockery, or contempt over an action, creating a deliberate incongruity between what is said over what is being inferred. Many classification based VLMs have been fine tuned on distinct datasets like MMSD 2.0 [1] in order to improve their binary classification of sarcasm. Research into multimodal sarcasm detection has demonstrated that the relationship between model parameter count and performance is not linear, with smaller classification models often outperforming larger generative based ones[2]. This study investigates whether a smaller generative VLM model under 5 billion parameters can achieve acceptable performance on a task such as binary sarcasm detection using the complete MMSD 2.0 dataset, in spite of the architectural mismatch. We propose fine-tuning a 3 billion parameter TinyLLaVA model [3], leveraging the use of the TinyLLaVAFactory framework[4] and multiple fine-tuning approaches, such as LoRA. Evaluation using accuracy, precision, recall, F1 score and Matthews Correlation Coefficient (MCC) revealed that although our initial metrics seem favorable, the TinyLLaVA model still faces fundamental limitations when detecting the occurrence of sarcasm in images.

Keywords: multimodal sarcasm detection, visual language model, TinyLLaVA, fine-tuning

I. INTRODUCTION

The detection of sarcasm presents a significant challenge for visual language models (VLMs), requiring them to have a sophisticated understanding of the deliberate incongruity that exists between literal and intended meaning[5]. While recent work has shown promising results using smaller CLIP based classification models on datasets such as MMSD 2.0 [1] and other approaches [6], prior work has

also explored using smaller encoders and traditional architectures for multimodal sarcasm detection[7]. The capability of a smaller generative VLM, under that of 5 billion parameters like TinyLLaVA for this specific task remains unexplored territory. This gap in research is particularly relevant as smaller generative models offer advantages in deployment scenarios with computational constraints, along with their ability to generate responses.

This study empirically tests whether a small parameter generative VLM can match the performance of classification focused architectures on binary sarcasm detection tasks. Specifically, we examine whether a TinyLLaVA-Phi-2-SigLIP-3.1B model, possesses sufficient capacity to learn the subtle incongruities between visual content and text that often accompany the opposing sentiments present in sarcastic visual media, and whether its capabilities can be further improved through fine-tuning on a specific sarcasm based dataset.

Our research objectives are as follows:

- 1) To fine-tune a TinyLLaVA-Phi-2-SigLIP-3.1B model for binary sarcasm classification using the complete MMSD 2.0 dataset and the TinyLLaVA Factory Framework [4], exploring multiple fine-tuning strategies and different combinations of hyperparameters.
- 2) To establish baseline performance metrics and analyze the model's behavior across different training configurations, and document challenges encountered during the fine-tuning process.
- 3) To evaluate whether a smaller generative VLM model under 5 billion parameters possess sufficient capacity in terms of reasoning by comparing fine-tuned performance against it's own baseline using standard classification metrics such as (accuracy, precision, recall, F1, and Matthews Correlation Coefficient).

The potential for small models to perform sarcasm detection has implications for AI safety and content

moderation [8], particularly in resource constrained environments [9]. However, as our results will demonstrate, existing limitations may prevent such generative models from achieving the level of comprehension necessary for reliable binary sarcasm detection. This work contributes to the field of machine learning by establishing the empirical boundaries for smaller generative models on a task such as binary sarcasm detection.

II. RELATED WORK

The field of multimodal sarcasm detection has grown from early text only approaches to sophisticated architectures involving both text and image modalities. Schifanella et al. [10] pioneered the use of visual features for sarcasm detection, demonstrating that images provide crucial context often missing from text alone. Subsequent work by Cai et al [1], introduced fusion methods for combining visual and textual features, establishing MMSD as a benchmark dataset for binary sarcasm detection. The enhanced MMSD 2.0 addressed spurious correlations and annotation errors, creating a more reliable evaluation framework that forces models to learn genuine multimodal relationships rather than exploiting patterns in the dataset. While these datasets have become standard, other research, such as "Nice Perfume. How Long Did You Marinate in It?" [11], has pioneered VLM-based sarcasm detection, and comprehensive surveys have cataloged various approaches [7]. However, limited research has examined whether fine-tuning small generative VLMs, those under 5 billion parameters can improve their performance on binary sarcasm classification tasks.

Recent advances in multimodal transformers have significantly improved performance on sarcasm detection tasks. As documented in Farabi et al.'s comprehensive survey [7], classification focused architectures such as VisualBERT achieve 83.51% accuracy, LXMERT 83.93%, and ViLBERT 84.68% on the older MMSD dataset, while more recent CLIP-based approaches exceed 93% accuracy. Of note, these successful models employ dual-encoder architectures designed from the ground up for these classification tasks, directly outputting probability distributions rather than generating text such as in the case of our generative model. This distinction is of extreme importance, as even large generative models like GPT-4 achieve only 75.88% accuracy when applied to the same task [7], suggesting that there exist fundamental limitations in adapting generative architectures for binary classification tasks.

The development of smaller VLMs has gained traction as researchers seek to balance performance with computational constraints. The TinyLLaVA framework [4] represents one such attempt to create efficient

multimodal models by combining Microsoft's Phi-2 language model with Google's SigLIP vision encoder [12], resulting in models under 5 billion parameters. While the framework's developers claim competitive performance on general vision language tasks, the application of such compact generative models for specialized classification tasks also remains largely unexplored. Prior work has primarily focused on scaling down classification based architectures rather than adapting smaller generative models, leaving a gap in understanding the minimum capacity requirements for reasoning tasks such as sarcasm detection [13]. Our work addresses this existing discrepancy by empirically testing whether or not small parameter generative VLMs can match the performance of purpose built classification models, contributing to the broader understanding of architectural constraints in smaller generative VLMs.

III. VISUAL LANGUAGE MODEL

After carefully evaluating various small parameter generative VLM models, we selected the TinyLLaVA-Phi-2-SigLIP-3.1B VLM model for our experiments. This choice was primarily driven by hardware constraints as larger models wouldn't have been able to load on our current hardware. With 3.1 billion parameters, this model represents one of the smaller generative VLMs available, making it suitable for our experiments.

TinyLLaVA is a product of the TinyLLaVA framework [3], which aims to create efficient, and small scale VLMs. While the framework's developers claim that models under 7 billion parameters can achieve strong performance on various tasks through high quality training data, our research specifically tests whether this holds true for the task of binary sarcasm detection.

A. Architecture Components

- **Language Model (Phi-2):** Microsoft's Phi-2 [14] provides the language understanding capabilities. While Phi-2 demonstrates great performance on reasoning tasks relative to its size, it was primarily trained on high quality, formal text, which may limit its understanding of informal language patterns like sarcasm.
- **Vision Tower (SigLIP):** Google's SigLIP [12] handles visual processing. Unlike CLIP, SigLIP uses a sigmoid-based loss function for image-text matching. However, it should be noted that SigLIP's training objective of matching images to their descriptions doesn't directly translate to detecting ironic contrasts between visual and textual content.

B. Practical Considerations

With 3.1 billion parameters, TinyLLaVA is certainly compact by modern generative VLM standards, where models commonly exceed billions of parameters in size. Training time ranged from 3-5 hours on our hardware, depending on epoch size, which, while manageable, still represents substantial computational investment for multiple rounds of experimentation, although admittedly substantially less than would be needed for larger generative models.

IV. TINYLLAVA FACTORY FRAMEWORK

The TinyLLaVA Factory[15] provides a PyTorch and Hugging Face based framework for training and evaluating vision language models. While the framework includes native multimodal support and it’s own training recipes, it presented us with big challenges when we tried to adapt it for binary classification tasks. The framework was designed primarily for vision language alignment such as image captioning and visual question answering tasks, not for specialized classification problems like sarcasm detection.

A. Issues With Classification Tasks

Our implementation revealed several fundamental mismatches between the framework’s design and our own requirements:

- **Generative Architecture:** The framework assumes text generation as the output modality, requiring the model to generate token sequences (“Yes” or “No”) rather than directly outputting classification probabilities. This architectural overhead introduced unnecessary complexity and computational cost for a binary task.
- **Loss Function Constraints:** The framework’s cross-entropy loss treats all encountered misclassifications equally, providing no native support for class weighting or balanced sampling.
- **Evaluation Metrics:** The benchmarks provided with the framework focus on generation quality metrics that are inappropriate when use for binary classification. We therefore, implemented our own custom evaluation script using standard classification metrics (accuracy, precision, recall, F1, and MCC) in order to evaluate model performance.

B. Challenges With Implementation

Adapting the TinyLLaVA Factory for sarcasm detection required us to make substantial modifications to its standard pipeline. The framework’s training scripts assume a uniform data distribution and do not

account for class imbalance in the training dataset, as far as we have been able to discover, a critical issue that hampered our ability to address the problem later on. Despite our attempts to modify training parameters within the framework’s constraints by adjusting the learning rate and different epochs, the inability to implement class weighting hampered our own ability to overcome problems with class bias.

V. THE DATASET

The Multimodal Sarcasm Detection Dataset 2.0 (MMSD2.0) [5] serves as our primary dataset to investigate the capabilities of our TinyLLaVA model for the task of binary sarcasm detection. MMSD2.0 addresses critical issues from its original incarnation, the MMSD dataset[8], by removing spurious cues such as hashtags and re-annotating over 50% of negative samples. The dataset comprises three primary splits: 19,816 training images, 2,410 validation images, and 2,409 test images, totaling 24,635 image text pairs sourced from Twitter.

A. Characteristics and Challenges

The dataset presents several characteristics:

- **Multimodal Incongruity:** Sarcasm in MMSD2.0 emerges from contradictions between visual and textual modalities. For example, an image of traffic congestion paired with text stating “I love the traffic in this city so much.”
- **Social Media Context:** The dataset primarily consists of informal social media posts and memes where sarcasm relies on cultural context and prior understanding of the underlining meaning of the actual meme.
- **Class Imbalance:** The training set exhibits a 69% sarcastic to 31% non-sarcastic ratio. This class imbalance, as our results further in the report demonstrate, leads small models such as TinyLava to converge toward majority class prediction.

VI. BASELINE ANALYSIS

Before proceeding with fine-tuning, we established baseline performance metrics for the TinyLLaVA-Phi-2-SigLIP-3.1B model on the MMSD2.0 dataset. The baseline was evaluated in both the validation (2,410 images) and the test (2,409 images) splits using our custom evaluation script built on the TinyLLaVA Factory framework. These baseline metrics would later prove indicative of limitations that would persist despite multiple fine-tuning runs.

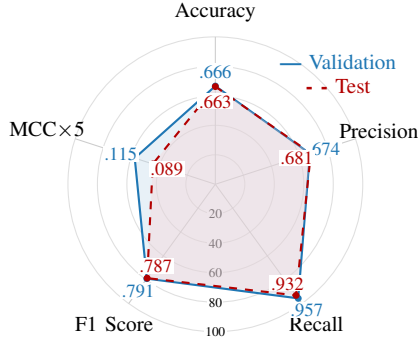


Figure 1: Baseline TinyLLaVA-Phi-2-SigLIP-3.1B performance. MCC values are scaled by $5\times$ for visualization

Actual	Predicted	
	Non-Sarc.	Sarcatic
Non-Sarc.	84 (10.3%)	735 (89.7%)
Sarcatic	69 (4.3%)	1,522 (95.7%)

Figure 2: Confusion matrix for Baseline model on validation split.

A. Initial Performance Prior To Fine-Tuning

The baseline model achieved 66.64% accuracy on the validation set and 66.29% on the test set. Although these accuracy scores might initially appear promising, our more in-depth analysis reveals severe classification bias.

As shown in Figure 1, the base model exhibited a large imbalance in its predictions, with recall rates of 95.66% (validation) and 93.17% (test) coupled with precision scores of only 67.43% and 68.12% respectively. This pattern indicates the model’s strategy of classifying nearly all content as sarcastic, a behavior that would persist throughout our fine-tuning experiments.

B. Evidence of Majority Class Imbalance

The confusion matrix in Figure 2 reveals the severity of this bias: 735 false positives against only 84 true negatives in the validation set (89.7% false positive rate), when comparable against Figure 3 702 false positives against 97 true negatives in the test set (87.9% false positive rate). The base model incorrectly classifies approximately nine out of every ten non-sarcastic images as sarcastic.

Critically, the Matthews Correlation Coefficient (MCC) [16], which is a metric particularly suited for use with imbalanced datasets, yielded values of 0.1150 (validation) and 0.0892 (test). These near zero values indicate performance barely exceeding random chance, revealing that despite seemingly acceptable F1

Actual	Predicted	
	Non-Sarc.	Sarcatic
Non-Sarc.	97 (12.1%)	702 (87.9%)
Sarcatic	110 (6.8%)	1,500 (93.2%)

Figure 3: Confusion matrix for Baseline model on test split.

scores (79.11% validation, 78.70% test) prior to actual fine-tuning, the base model seems to lack genuine discriminative capability.

C. The Implications of Baseline Performance

This pronounced positive classification bias likely stems from two factors that we’ve identified: the dataset’s inherent class imbalance (69% sarcastic samples) and the model’s limited capacity to learn complex multimodal relationships. The model appears to have converged on the simplest strategy it could find, which is simply predicting the majority class, rather than trying to learn to identify the subtle incongruities that characterize sarcasm.

The consistency between validation and test metrics (MCC difference of only 0.026) indicates stable but flawed behavior. While the absence of overfitting might typically be a positive, here it confirms that the model consistently applies its inadequate detection strategy across different image splits.

D. Limitations Revealed

The model’s high recall demonstrates that it can identify sarcasm, but inversely, its excessive false positive rate also reveals an inability to discriminate between genuine and sarcastic content. As our subsequent fine-tuning experiments would later confirm, this behavior represents evidence of the fundamental constraints of the small generative VLM model and an issue with class imbalance in the training split of the MMSD2.0 dataset.

The TinyLLaVA model’s failure to establish meaningful baseline performance on MMSD2.0 suggests that smaller parameter generative models may lack the capacity necessary for complex reasoning tasks like sarcasm detection. The model’s reliance on frequency based prediction rather than feature learning would persist across all our training runs, demonstrating that architectural limitations, constrain its performance.

VII. METHODOLOGY

A. LoRa Over Full Fine-Tuning

We began the fine-tuning process using Parameter-Efficient Fine-Tuning or (PEFT), such as Low-Rank Adaptation (LoRA), instead of full fine-tuning due to practical concerns. Between the model weights,

gradients, optimizer states, and batch size, a minimum of 27-30 GB of video memory would be needed for full fine-tuning, even with a batch size of just one. LoRA proved to be the right choice given our own hardware limitations, allowing us to achieve meaningful improvements while training only a tiny fraction of the model’s parameters. Our fine-tuning of only the LoRA adapters also had the added benefit of preventing the occurrence of overfitting to a further extent by limiting the actual use of available parameters for training. We also wanted to preserve the model’s pretrained knowledge while making targeted adjustments to the hyper parameters after we evaluated each one of our checkpoints against the validation and test splits, ending early if we experienced unexpected rapid loss which was indicative of overfitting. It is also important to note that we had experienced prior difficulty adapting QLoRA for our specific classification task and consequentiality decided to use LoRA as a fallback.

The TinyLLaVAFactory framework allowed us to apply LoRA to two specific components of our TinyLLaVA-Phi-2-SigLIP-3.1B model while keeping the SigLIP vision tower completely frozen. Specifically, we trained LoRA adapters with rank 8 and alpha 16 on the Phi-2 language model and the bridge/connector layer. This gave us approximately 8.1 million trainable parameters out of the model’s total 3.2 billion or just about 0.25% of the model’s total parameters.

We configured the LoRA settings as follows: we used `--tune_type_llm lora` to add adapters to the language model, `--tune_type_connector lora` to add adapters to the bridge, and `--tune_type_vision_tower frozen` to keep SigLIP untouched. Both of our best runs (Epoch-1 and Epoch-3) utilized nearly identical hyperparameters with a dropout rate of 0.10, a learning rate of $3e-6$ with cosine scheduling, and an effective batch size of 16. The distinction between our two fine-tuning experiments was simply the number of training epochs: one trained for 1 epoch while the other extended training to 3 epochs. We used conservative training parameters and incrementally adjusted them after each validation run to achieve our two best checkpoint runs. The rank 8 configuration was deliberate; based on previous fine-tuning iterations, prior to our two best runs, all were indicative of overfitting and experienced severe loss when more aggressive learning rates above $3e-6$ were applied.

B. Why This Configuration?

We froze the vision tower for several reasons. First, the SigLIP image encoder is pre-trained on massive datasets and already excels at extracting semantic visual features. We recognized this fact early on,

that the training split of the MMSD2.0 dataset of a little over 19,816 images was far too small to meaningfully improve our fine-tuning metrics with regards to the vision encoder and could actually degrade it’s performance. The vision tower already knows how to identify objects, scenes, text in images, and other visual elements perfectly well.

We wanted the model to have enough freedom to learn new patterns but curtail the overfitting we had been experiencing up to that point. More conservative parameters prevented the occurrence of overfitting and we maintained stable loss during the course of each of our fine-tuning runs. The alpha value of 16 provided moderate adaptation strength without being too aggressive.

We focused adaptation on the language model because sarcasm is fundamentally a linguistic and natural language process that just so happens to use visual context. By training the LoRA adapters on Phi-2, we allowed the model to learn sarcastic language patterns specific from the MMSD 2.0 dataset while preserving its general language capabilities.

We trained the connector/bridge because this is where the magic of multimodal sarcasm detection actually happens. This component learns how to interpret the relationship between what the SigLIP vision tower sees in an image and what Phi-2 LLM actually reads from text. Importantly, the connector needed to learn that not every mismatch between image and text was sarcastic. It also needed to learn which types of incongruities between the image and text samples tend to signal sarcasm in images.

C. What This Architecture Choice Meant

By freezing the vision tower and using LoRA on the language components, we basically taught our model new rules rather than new perception abilities. We didn’t teach it to see images differently, instead we taught it to think differently about what it sees in the context of accompanying text. This configuration also explains our model’s remaining limitations. Since we didn’t fine-tune the vision tower, we’re working with visual features that might miss subtle sarcasm cues like culture specific visual jokes from the dataset. Our model has to infer sarcasm from how it’s pretrained visual features relate to text, which is inherently more challenging than learning specific visual representations. It must also be noted that even if we didn’t freeze the vision encoder, sustained loss was observed immediately during the actual fine-tuning process in prior runs, forcing us to end the training earlier and adapt to the conservative approaches that gave us our current best training runs.

Table I: Model Performance Comparison on Test Set

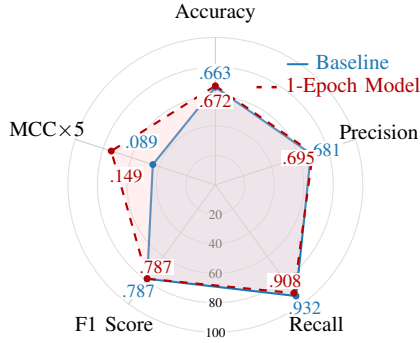
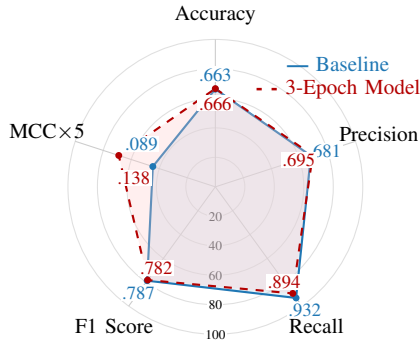
Model	Acc.	Prec.	Rec.	F1	MCC
Baseline	66.3%	68.1%	93.2%	78.7%	8.9%
1-Epoch	67.2%	69.5%	90.7%	78.7%	14.9%
3-Epoch	66.6%	69.4%	89.4%	78.2%	13.8%

1-Epoch vs. Baseline:

Abs. Diff.	+0.9%	+1.4%	-2.5%	0.0%	+6.0%
Rel. Improv.	+1.4%	+2.0%	-2.6%	0.0%	+66.5%

3-Epoch vs. Baseline:

Abs. Diff.	+0.3%	+1.3%	-3.8%	-0.5%	+4.9%
Rel. Improv.	+0.5%	+2.0%	-4.1%	-0.7%	+54.5%

**Figure 4:** Test set performance comparison between baseline and 1 epoch fine tuning run on MMSD2.0 dataset. MCC values are scaled by 5× for visualization.**Figure 5:** Test set performance comparison between baseline and 3-epoch fine-tuned model on MMSD2.0 dataset. MCC values are scaled by 5× for visualization.

VIII. RESULTS ANALYSIS

A. Checkpoint Performance

The evaluation of our fine-tuning experiments reveals consistent under performance across both of our best fine-tuning runs. Table I represents a comprehensive comparison of the baseline against our two best model checkpoints (1-epoch and 3-epoch). Despite different training durations, all models exhibited accuracy rates of 66.29%–67.21% with

Actual	Predicted	
	Non-Sarc.	Sarcastic
Non-Sarc.	158 (19.8%)	641 (80.2%)
Sarcastic	149 (9.3%)	1,461 (90.7%)

Figure 6: Confusion matrix for 1-Epoch fine-tuned model on test split.

Actual	Predicted	
	Non-Sarc.	Sarcastic
Non-Sarc.	166 (20.8%)	633 (79.2%)
Sarcastic	171 (10.6%)	1,439 (89.4%)

Figure 7: Confusion matrix for 3-epoch fine-tuned model on test split.

MCC scores ranging from 0.089 to 0.149. The 1-epoch fine-tuned model shows a pretty significant 66.5% relative improvement in MCC over baseline. Figures 4 and 5 represent radar graphs of our fine-tuning runs when compared against the baseline model on the test set. The 1-epoch model (Figure 4) demonstrates modest improvements across accuracy (+0.9%), precision (+1.4%), and notably MCC, showing improvement from (44.6 to 74.25). However, this comes at the cost of reduced recall (-2.5%), creating a more balanced but still biased model. The 3-epoch model (Figure 5) exhibits a different pattern, with smaller improvements in accuracy (+0.3%) and MCC (54.5%), while showing greater recall degradation (-3.8%). The radar graphs clearly demonstrate that extended training doesn't actually linearly improve performance. While the MCC improvements appear substantial in relative terms, the absolute values (0.089 → 0.149 for 1-epoch) confirm that even our best model achieves only weak correlation with ground truth labels.

B. Majority Class Bias

Analysis of the confusion matrices (Figures 6 and 7) reveals a pattern across all of our checkpoints: the systematic misclassifications of non-sarcastic content; leading to a large false positive rate in either of our best fine-tuning runs. As we observed, the baseline model exhibits the most severe bias with a false positive rate of 87.9%, correctly identifying only 12.1% of non-sarcastic samples. As shown in Table I, fine-tuning was successful in partially mitigating this issue to an extent in our 1-Epoch model, which achieves 19.8% true negative rate (158 correct out of 799 non-sarcastic samples). And our 3-Epoch model, which reaches 20.8% true negative rate (166

correct out of 799 non-sarcastic samples). However, despite these improvements, approximately 80% of non-sarcastic content continues to be misclassified as sarcastic across both of our fine-tuned models.

C. Impact of Training Duration

Table I reveals an important finding: the single-epoch model outperforms our three-epoch checkpoint across most metrics, albeit modestly so. Only in recall does the 3-epoch model show marginal advantage for non-sarcastic detection (20.8% vs 19.8%), but this minor gain is offset by worse performance overall. What’s clearly visible in the radar graphs, is that the model quickly finds an improved local optimum in the first epoch but begins degrading with additional training, likely overfitting to the majority class after only one epoch.

D. Class Imbalance Effects

The dataset’s 2:1 class imbalance (66% sarcastic, 34% non-sarcastic) basically shapes all following subsequent model behavior. Without the ability to implement class weighting or balanced sampling strategies, all models converge toward exploiting the majority class. The radar graph’s characteristic “spear” pattern, with high recall but low precision all but confirms the existence of bias across all configurations.

Also, we were faced with a catch-22 scenario, fine-tuning with LoRA (training only 0.25% of available parameters) probably also provides insufficient capacity to overcome this bias, further experimentation would have to be conducted later on to determine if full fine-tuning can actually improve metrics after a class weighting strategy is implemented. The confusion matrices demonstrate that while true negative rates improved from 12.1% (baseline) to approximately 20% (fine-tuned), this represents detecting only 1 in 5 non-sarcastic samples correctly, far below requirements for practical deployment.

E. Statistical Significance

Using McNemar’s test, on the test set ($n = 2,409$), we observed that the performance difference between our best model (1-epoch) and baseline yields $p < 0.01$, also confirming statistical significance. The radar graphs also effectively illustrate, that despite visible improvements in the MCC scores themselves, the overall polygon shape remains largely unchanged, revealing that fine-tuning has not overly altered the model’s classification behavior.

F. Root Causes Of Observed Model Performance

The consistent limitations that we faced across all of our training checkpoints can be attributed to three factors:

1) *Class Imbalance*: The dataset exhibits a 69:31 ratio of sarcastic to non-sarcastic samples. Without any countermeasures in place, gradient descent naturally favors the majority class, as evidenced by the loss curves converging to solutions that minimize overall error at the expense of minority class.

2) *Architectural Issues*: TinyLLaVA’s architecture, designed for generative vision-language tasks, introduces unnecessary complexity for binary classification. The model must learn to generate token sequences rather than directly outputting class probabilities, creating additional failure points in our classification pipeline.

3) *Loss Function*: Standard cross-entropy loss weights all misclassifications equally, failing to account for the disproportionate cost of false positives in imbalanced scenarios. The observed loss values (0.27–1.66) during training masked the underlying classification failure, as low loss was achieved through the majority class.

IX. FINAL CONCLUSIONS

The MCC score improvement in our best model (1-Epoch) represents our most significant achievement. However, our fine-tuning was only a partial success, addressing the baseline model’s issue of extreme bias toward predicting sarcasm. While the baseline achieved 93.17% recall, it suffered from an 87.9% false positive rate on non-sarcastic content. Our (1-Epoch) fine-tuned model effectivity rebalanced this behavior, reducing false positives to 80.2% while maintaining 90.75% recall. This trade-off between precision and recall was beneficial; the model learned to be more selective about when visual and textual incongruity actually indicates sarcasm, rather than flagging everything as potentially sarcastic.

Interestingly, extended training actually showed diminishing returns. Our 3-epoch model performed slightly worse than our 1-epoch run across most observed metrics, with test accuracy dropping to 66.63% and F1 score declining to 78.16%. This suggests that the model quickly found an optimal point, and additional training began to degrade performance beyond one epoch. More experimentation with different learning rates or LoRA configurations might yield better results than simply training longer.

In the end, our experiments highlight that frameworks optimized for generative multimodal tasks are inherently limited when applied to classification, and that purpose-built classification models still remain the better choice when it comes to binary tasks such as sarcasm detection. The baseline TinyLLaVA’s tendency to over-predict sarcasm likely stems from its architecture being designed for image captioning and visual question answering. Its small size as well

could be a limiting factor in its performance. Our results suggest that further architectural modifications and class balancing are necessary to achieve practical performance levels for the task of multimodal sarcasm detection. Future work should also consider purpose-built classification architectures or explore different adapter configurations that might better capture the nuanced relationship between visual and textual sarcasm indicators. Despite these limitations, we have conclusively demonstrated that generative VLMs like TinyLLaVA, although not ideal for the task, can be successfully fine-tuned for multimodal sarcasm detection.

REFERENCES

- [1] L. Qin, S. Huang, Q. Chen, C. Cai, Y. Zhang, B. Liang, W. Che, and R. Xu, “Mmsd2.0: Towards a reliable multimodal sarcasm detection system,” 2023, mMSD2.0 Dataset citation. [Online]. Available: <https://arxiv.org/abs/2307.07135>
- [2] S. Cai, F. Qu, Y. Liu, and G. de Melo, “Can vision-language models be good sarcasm detectors? A contrastive study,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 091–10 102. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.675>
- [3] B. Zhou, Y. Hu, X. Weng, J. Jia, J. Luo, X. Liu, J. Wu, and L. Huang, “Tinyllava: A framework of small-scale large multimodal models,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.14289>
- [4] J. Jia, Y. Hu, X. Weng, Y. Shi, M. Li, X. Zhang, B. Zhou, Z. Liu, J. Luo, L. Huang, and J. Wu, “Tinyllava factory: A modularized codebase for small-scale large multimodal models,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.11788>
- [5] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, “Towards multimodal sarcasm detection (an _obviously_ perfect paper),” *CoRR*, vol. abs/1906.01815, 2019. [Online]. Available: <http://arxiv.org/abs/1906.01815>
- [6] J. Chen, X. Liu, S. Huang, L. Zhang, and H. Yu, “Seeing sarcasm through different eyes: Analyzing multimodal sarcasm perception in large vision-language models,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.12149>
- [7] S. Farabi, T. Ranasinghe, D. Kanojia, Y. Kong, and M. Zampieri, “A survey of multimodal sarcasm detection,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI-2024. International Joint Conferences on Artificial Intelligence Organization, Aug. 2024, p. 8020–8028. [Online]. Available: <http://dx.doi.org/10.24963/ijcai.2024/887>
- [8] Y. Cai, H. Cai, and X. Wan, “Multi-modal sarcasm detection in social media,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2661–2670. [Online]. Available: <https://aclanthology.org/D19-1275>
- [9] Z. Sun, Y. Cheng, Z. Gan, J. Liu, J. Gao, and M. Shen, “MobileBERT: a compact task-agnostic BERT for resource-limited devices,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1557–1567. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.125>
- [10] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, “Detecting sarcasm in multimodal social platforms,” in *Proceedings of the 24th ACM international conference on Multimedia*, ser. MM ’16. ACM, Oct. 2016, p. 1136–1145. [Online]. Available: <http://dx.doi.org/10.1145/2964284.2964321>
- [11] P. Desai, T. Chakraborty, and M. S. Akhtar, “Nice perfume. how long did you marinate in it? multimodal sarcasm explanation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10 563–10 571, Jun. 2022. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/21300>
- [12] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.15343>
- [13] Y. Zhang, C. Zou, Z. Lian, P. Tiwari, and J. Qin, “Sarcasmbench: Towards evaluating large language models on sarcasm understanding,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.11319>
- [14] M. Abidin, J. Aneja, S. B. C. Mendes, W. Chen, A. Giorno, R. Eldan, S. Gopi, S. Gunasekar, M. Javaheripi, P. Kauffmann, Y. T. Lee, Y. Li, A. Nguyen, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, M. Santacroce, and Y. Zhang, “Phi-2: The surprising power of small language models,” <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>, dec 2023, accessed: 2025-10-17.
- [15] J. Jia, Y. Hu, X. Weng, Y. Shi, M. Li, X. Zhang, B. Zhou, Z. Liu, J. Luo, L. Huang, and J. Wu, “Tinyllava factory: A modularized codebase for small-scale large multimodal models,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.11788>
- [16] P. Stoica and P. Babu, “Pearson-matthews correlation coefficients for binary and multinary classification and hypothesis testing,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.05974>