Figurative Language Understanding via Mixture-of-Adapters and Tensor-of-Cues

Ashkan Shafiei
Politecnico di Torino
s342583
s342583@studenti.polito.it

Amir Masoud Almasi
Politecnico di Torino
s337006
s337006@studenti.polito.it

Mehdi Nickzamir
Politecnico di Torino
s323959
s323959@studenti.polito.it

Balzhan Dosmukhametova
Politecnico di Torino
s343931
s343931@studenti.polito.it

***Abstract*—Modeling affective and figurative language requires balancing literal sentiment with pragmatic shifts like sarcasm, particularly under domain and English variety shifts. Using the BESSTIE benchmark (en-AU, en-IN, en-UK), we evaluate BERT, RoBERTa, and Mistral-7B-Instruct across in-variety, cross-variety, and cross-domain settings. We find that while sentiment remains stable, sarcasm detection degrades significantly under variety shifts (e.g., RoBERTa F1 drops to 0.49).**

To improve robustness, we propose two extensions: (1) a Mixture-of-Adapters (MoA) architecture for RoBERTa that utilizes instance-level routing, boosting cross-variety sarcasm F1 from 0.49 to 0.65 and cross-domain sentiment from 0.75 to 0.83; and (2) a Variety-Aware Tensor-of-Cues (ToC) strategy for Mistral-7B-Instruct that models pragmatic signals like hyperbole. ToC improves zero-shot sarcasm F1 from 0.20 to 0.38 and fine-tuned cross-variety performance from 0.39 to 0.45 without compromising sentiment accuracy. Our results demonstrate that conditional parameterization and structured pragmatic modeling are essential for robust figurative language understanding. Code is available at https://github.com/emirmasood/FigLangUnderstanding.

***Index Terms*—Figurative Language Understanding, Sarcasm Detection, Sentiment Analysis**

I. Introduction

Sentiment analysis and sarcasm detection are central to understanding affective intent in text. Sentiment classification often correlates with surface lexical polarity, whereas sarcasm is frequently expressed through pragmatic inversion, exaggeration, and contextual incongruity, where the intended meaning can oppose the literal polarity [2]–[4]. As a result, sarcasm can systematically distort sentiment cues and remains notably fragile under distribution shift across domains and language varieties [5].

The BESSTIE benchmark provides a controlled setting to study this challenge across three English varieties (en-AU, en-IN, en-UK) and two platforms (Google and Reddit). Prior findings show that Transformer-based models can achieve strong in-domain sentiment performance, yet suffer substantial degradation under cross-variety and cross-domain transfer, with the largest drops typically observed for sarcasm. This suggests that monolithic models often under-capture dialectal variation and do not explicitly structure pragmatic cues, limiting robustness when the data distribution changes.

In this work, we investigate generalization under variety and platform shift through (i) *conditional adaptation* and (ii) *structured cue modeling*. We propose a RoBERTa-based Mixture-of-Adapters (MoA) that enables instance-level, variety-sensitive parameterization, and a Variety-Aware Tensor of Cues (ToC) prompting strategy that organizes LLM predictions around interpretable pragmatic signals. Experiments on BESSTIE show that these mechanisms improve sarcasm robustness under shift while maintaining competitive sentiment performance.

II. Methodology

Our pipeline consists of dataset preprocessing, baseline fine-tuning, and training/evaluation of the proposed MoA and ToC variants. All model families share identical split logic and evaluation metrics to ensure fair comparisons.

A. Dataset and Preprocessing

We use the BESSTIE dataset [1], which contains 17,760 training and 2,428 validation samples in its primary release. Each instance includes five features: the raw *text*, a binary *label* (sentiment or sarcasm), the linguistic *variety* (en-AU, en-IN, or en-UK), the *source* platform (Google or Reddit), and the *task* (sentiment analysis or sarcasm detection). While we use samples from both sources for sentiment, we restrict sarcasm-focused training and evaluation to Reddit instances since Google-derived sarcasm samples are negligible.

To minimize semantic distortion, we apply minimal text normalization. URLs, numbers, and user mentions are replaced with placeholders, while punctuation and capitalization are retained. This choice is motivated by the importance of affective markers (e.g., emphasis, punctuation, casing), which can be predictive for sarcasm where subtle textual cues may alter intended meaning.

We derive our training set from the original training partition and construct a validation set by taking 20% of the original test set. To preserve subgroup proportions and reduce variance in low-resource settings, we apply adaptive stratification with a hierarchical rule: when multiple varieties are present, we stratify by *label* + *variety*; if a single variety exists but multiple sources are available, we stratify by *label* + *source*; otherwise, we stratify by *label* alone. The final distribution for both tasks, categorized by variety and source, is reported in Table I.

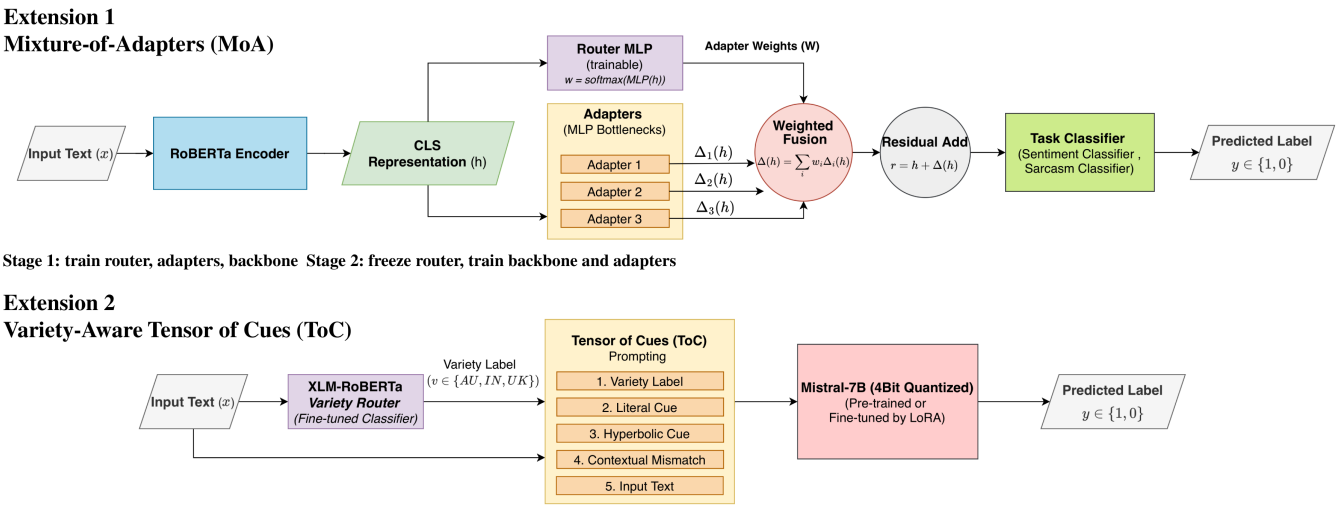

Fig. 1. Proposed architectural extensions: (Top) MoA with instance-level routing; (Bottom) ToC with structured pragmatic prompting.

TABLE I
DISTRIBUTION OF SAMPLES ACROSS DIFFERENT TRAINING SETTINGS AND DATASET SPLITS.

Task	Train Setting	Train	Val	Test
Sentiment	Google	3529	882	603
	Reddit	3564	891	609
	TRAIN_en-AU	2167	542	371
	TRAIN_en-IN	2667	666	455
	TRAIN_en-UK	2259	565	386
Sarcasm	Reddit	3585	895	612
	TRAIN_en-AU	1411	352	241
	TRAIN_en-IN	1349	337	230
	TRAIN_en-UK	825	206	141

TABLE II
MACRO-F1 BASELINE PERFORMANCE FOR ENCODER MODELS.

Task	Setting	BERT-base	RoBERTa-base
Sentiment	Cross-Domain	0.77	0.75
	Cross-Variety	0.87	0.89
	In-Domain	0.85	0.86
	In-Variety	0.90	0.91
Sarcasm	Cross-Variety	0.49	0.49
	In-Variety	0.64	0.69
	In-Domain	0.66	0.67

B. Baselines

1) Encoder-Based Models: To establish a baseline and select the backbone for our encoder extension, we fine-tune **BERT-base** [6] and **RoBERTa-base** [7] across multiple dataset splits for both sentiment and sarcasm detection. To address class imbalance, we use a class-weighted cross-entropy loss during full-parameter fine-tuning. Optimization is performed using AdamW [10] with learning rate 2×10^{-5} and batch size 16. We use a maximum sequence length of 256 tokens for 5 epochs, covering approximately 99% of the dataset.

As reported in Table II, RoBERTa-base consistently outperforms BERT-base, with the largest gains on sarcasm detection (e.g., in-variety Macro-F1 of 0.69 vs. 0.64). However, despite strong in-distribution performance, both encoders degrade substantially when generalizing to unseen English varieties or across platforms.

2) Decoder-Based Model: Due to hardware constraints, we use **Mistral-7B-Instruct-v0.3** as a computationally efficient decoder baseline [14]. To reduce memory usage, we employ 4-bit quantization during training and evaluation [13]. We evaluate two conditions: (i) a pretrained zero-shot baseline and (ii) supervised fine-tuned (SFT) variants.

For SFT, we fine-tune nine models (one per training split) using Low-Rank Adaptation (LoRA) [12] with rank $r = 16$ and scaling $\alpha = 32$. Training uses learning rate 2×10^{-4}, effective batch size 8 (per-device batch size 4 with 2 gradient accumulation steps), maximum sequence length 1024, and LoRA dropout 0.1. Although initial runs were configured for up to 30 epochs, validation performance typically plateaued within the first epoch; we therefore train for 5 epochs to reduce overfitting and computational overhead.

Both SFT and pretrained variants use identical prompts across all splits:

- **Sentiment:** "Generate the sentiment of the given text. 1 for positive sentiment, and 0 for negative sentiment. Do not give an explanation."
- **Sarcasm:** "Predict if the given text is sarcastic. 1 if the text is sarcastic, and 0 if the text is not sarcastic. Do not give an explanation."

C. Extension 1: Mixture-of-Adapters (MoA)

Given its stronger baseline performance, we adopt **RoBERTa-base** as the foundation for our Mixture-of-Adapters extension. As illustrated in Fig. 1, we attach a lightweight bank of bottleneck adapters to the final pooled representation, enabling instance-level conditional adaptation with limited additional parameters [8]. Routing follows the mixture-of-experts principle [9], but is applied at the representation level rather than within each transformer layer, which reduces overhead and isolates adaptation capacity.

Let $H \in \mathbb{R}^{T\times d}$ denote the final hidden states from RoBERTa-base. We take the classification token embedding $h = H_{[0]} \in \mathbb{R}^d$. A router network $g(\cdot)$ (two-layer MLP with $\tanh$ activation) maps h to a distribution over K adapters:

$$w = \mathrm{softmax}(g(h)), \quad w \in \mathbb{R}^K. \tag{1}$$

Each expert E_k computes a residual update $\Delta_k(h)$. The mixed representation is:

$$r = h + \sum_{k=1}^{K} w_k \, \Delta_k(h). \tag{2}$$

Finally, r is passed to a linear classifier to produce $\hat{y}$. We set $K = 3$ (AU, IN, UK); the router learns its policy implicitly without explicit variety supervision.

We adopt a two-stage training protocol to stabilize routing:

- **Stage 1 (Pooled Training):** Train on pooled data to expose the router to diverse examples. To reduce expert collapse, we optimize:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{lb}}\mathcal{L}_{\mathrm{lb}} - \lambda_{\mathrm{ent}}\mathcal{H}(w), \tag{3}$$

  where $\mathcal{L}_{\mathrm{lb}}$ encourages balanced utilization and $\mathcal{H}(w)$ discourages overly peaked routing. We use $\lambda_{\mathrm{lb}} = 0.02$ and $\lambda_{\mathrm{ent}} = 0.01$.
- **Stage 2 (Setting-Specific Adaptation):** Fine-tune separately per setting (variety- or domain-specific), initializing from Stage 1 and freezing the router. This keeps routing stable while adapters and backbone adapt to the target distribution.

D. Extension 2: Variety-Aware Tensor-of-Cues (VAToC)

We implement a two-stage pipeline that combines dialectal awareness with structured sarcasm detection.

a) Step 1: Variety Routing (XLM-RoBERTa).: We fine-tune **XLM-RoBERTa-Base** [11] as a variety classifier. To prevent leakage, we use a stratified 80/20 split of the original training set for fine-tuning, keeping the original test split intact. The model is trained for 20 epochs ($LR = 2 \times 10^{-5}$, batch size=32) and achieves **88% accuracy** on the held-out test set. The predicted variety conditions the downstream prompting to reflect regional pragmatic norms.

b) Step 2: Hyperbolic Calibration with Tensor of Cues (ToC): Alongside the predicted variety label, we apply the **Tensor of Cues (ToC)** framework, which decomposes sarcasm detection into three anchored cues:

1) **Literal Cues:** assess proportionality of phrasing to the described situation.
2) **Hyperbolic Cues:** detect extreme intensity/exaggeration (e.g., "absolute genius").
3) **Contextual Mismatch:** identify misalignment between high-intensity language and mundane/negative events.

This hybrid approach makes the decision process more explicit while accounting for variety-specific norms.

Beyond applying ToC to the pretrained model, we also perform SFT using the ToC prompt with LoRA ($r = 16, \alpha = 32$), matching Step 1 hyperparameters (20 epochs, $LR = 2\times10^{-5}$, batch size=32). We extend the context window to 2048 tokens to accommodate the structured prompt. The unified prompt template is provided in the accompanying GitHub repository.

III. Results and Analysis

A. Extension 1: Mixture-of-Adapters

Sarcasm: MoA achieves substantial cross-variety improvements, with the largest gain when training on en-AU and testing on en-UK (+75.7%: 0.37→0.65 F1). The smallest improvement occurs in in-domain settings (+1.5%: 0.67→0.68 F1).

Sentiment: MoA strengthens cross-domain robustness; for instance, when trained on *Reddit* and evaluated on *Google*, performance increases (+15.5%: 0.71→0.82 F1). In-variety performance remains stable, showing only a slight change (-1.1%: 0.91→0.90 F1).

B. Extension 2: Variant Aware Tensor-of-Cues

Sarcasm: ToC delivers transformative zero-shot improvements (+89.3%: 0.20→0.38 F1) and strong in-variety gains, particularly for en-UK (+50%: 0.44→0.66 F1). Cross-variety transfer shows moderate improvements (+16.8%: 0.39→0.45 F1), with the largest gain from en-UK→en-IN(30.4%: 0.46→0.60 F1).

Sentiment: ToC maintains stable performance across all settings, with modest average zero-shot gains (+5.3%: 0.81→0.85 F1) and approximately consistent cross-variety performance (0.91 F1 average).

In general, it can be seen that MoA excels at cross-variety sarcasm transfer through variety-aware routing, while ToC achieves breakthrough zero-shot sarcasm detection via structured pragmatic reasoning. Both extensions maintain stable sentiment performance, confirming that figurative language requires specialized architectural mechanisms beyond literal polarity detection.

IV. Discussion

Our results indicate that distribution shift affects sarcasm detection more severely than sentiment classification. While

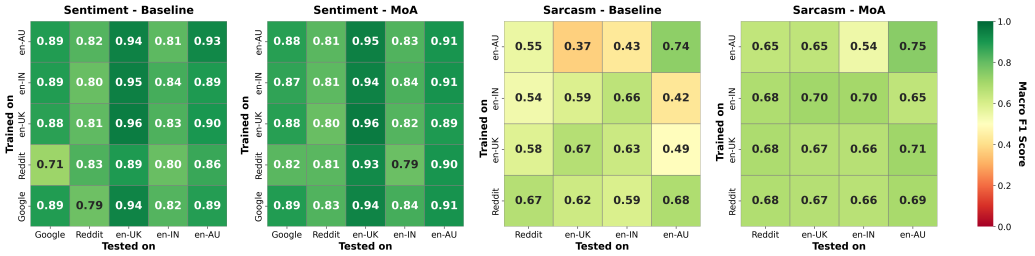

Fig. 2. Mixture-of-Adapters (MoA) extension: Macro-F1 performance across domain and variety shifts for sentiment and sarcasm detection.

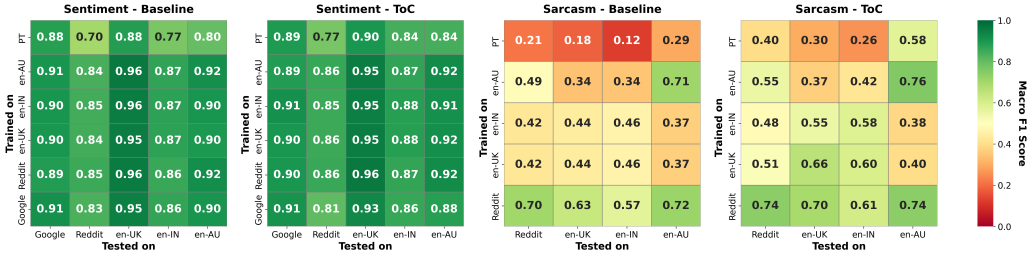

Fig. 3. Variety-Aware Tensor-of-Cues (VAToC) extension: Macro-F1 performance across domain and variety shifts for sentiment and sarcasm detection.

TABLE III
MACRO-F1 PERFORMANCE: EXTENSIONS VS. BASELINES

Task	Setting	Base*	MoA	Δ	Base†	ToC	Δ
Sentiment	In-Variety	0.91	0.90	−1.1%	0.81	0.92	**+12.9%**
	Cross-Variety.	0.89	0.89	**+0.5%**	0.91	0.91	+0.3%
	In-Domain	0.86	0.85	−0.5%	0.88	0.89	**+0.2%**
	Cross-Domain.	0.75	0.83	**+10.1%**	0.86	0.86	−0.8%
	Zero-shot	—	—	—	0.81	0.85	**+5.3%**
Sarcasm	In-Variety	0.69	0.71	+2.8%	0.54	0.67	**+24.0%**
	Cross-Variety.	0.49	0.65	**+33.7%**	0.39	0.45	+16.8%
	In-Domain	0.67	0.68	+1.1%	0.70	0.74	**+5.1%**
	Zero-shot	—	—	—	0.20	0.38	**+89.6%**

*Baseline Fine-tuned RoBERTa-base
†Baseline Fine-tuned Mistral-7B-Instruct

sentiment remains relatively stable across varieties and platforms, sarcasm suffers from significant cross-variety degradation, consistent with its reliance on culturally grounded pragmatic cues.

The Mixture-of-Adapters (MoA) mitigates this gap by enabling instance-level conditional adaptation. This approach improves cross-variety sarcasm from 0.49 to 0.65 Macro-F1 while maintaining competitive sentiment results and boosting cross-domain sentiment from 0.75 to 0.83 Macro-F1. These gains suggest that conditional routing can capture variety-specific signals without training separate models for each variety.

The Variety-Aware Tensor-of-Cues (VAToC) enhances Mistral-7B by making pragmatic evidence—such as hyperbole and contextual mismatch—explicit. This yields a substantial increase in zero-shot sarcasm detection (0.20 → 0.38 Macro-F1). Supervised variants also improve performance: 0.54 → 0.67 Macro-F1 in in-variety settings and 0.39 → 0.45 Macro-F1 in cross-variety settings. Overall, structured cue modeling appears complementary to supervised adaptation, particularly in low-resource or zero-shot scenarios.

V. CONCLUSION AND FUTURE WORK

We studied figurative language understanding under distribution shift on BESSTIE and introduced two mechanisms targeting robustness: (i) a RoBERTa-based Mixture-of-Adapters for instance-level conditional adaptation and (ii) a Variety-Aware Tensor-of-Cues (VAToC) strategy for structured pragmatic modeling with Mistral-7B.

Across variety and domain shifts, both approaches improve sarcasm robustness while preserving strong sentiment performance. MoA yields the largest gains in cross-variety sarcasm transfer (0.49 → 0.65 Macro-F1), while VAToC yields substantial improvements in zero-shot sarcasm detection (0.20 → 0.38 Macro-F1).

This work is limited to three English varieties and two sources, and we do not yet fully characterize router behavior or cue interactions. Future work will expand to more varieties and genres, analyze routing decisions and cue attribution for interpretability, and investigate joint training that integrates variety routing with cue-based prompting.

REFERENCES

[1] D. Srirag, A. Joshi, J. Painter, and D. Kanojia, "BESSTIE: A Benchmark for Sentiment and Sarcasm Classification for Varieties of English," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[3] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: A survey," *ACM Computing Surveys*, vol. 50, no. 5, pp. 1–22, 2017.

[4] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proc. ACL (Short Papers)*, 2015.

[5] H. Jang and D. Frassinelli, "Generalizable sarcasm detection is just around the corner, of course!," *arXiv:2404.06357*, 2024.
[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.
[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019.
[8] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Kelly, R. Jones *et al.*, "Parameter-efficient transfer learning for NLP," in *Proc. ICML*, 2019.
[9] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, J. Dean *et al.*, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. ICLR*, 2017.
[10] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. ICLR*, 2019.
[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proc. ACL*, 2020.
[12] 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," *arXiv:2106.09685*, 2021.
[13] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," *arXiv:2305.14314*, 2023.
[14] Mistral AI, "Mistral 7B," technical report / model documentation, 2023–2024.