UNF-BMI at SemEval-2026 Task 3: Research Domain Criteria-Guided Large Language Models for Dimensional Aspect-Based Sentiment Analysis

Athlene V. Jones[*]
University of North Florida
School of Computing
Jacksonville, FL, USA
N00168025@unf.edu

Vishwaa Shah[*]
University of North Florida
School of Computing
Jacksonville, FL, USA
N01458714@unf.edu

Indika Kahanda
University of North Florida
School of Computing
Jacksonville, FL, USA
indika.kahanda@unf.edu

Abstract

We present UNF-BMI system for SemEval-2026 Task 3, Track A, Subtask 1 (Dimensional Aspect Sentiment Regression, DimASR), which focuses on predicting continuous Valence–Arousal (VA) scores for aspects in text. Our approach integrates psychologically grounded affective signals inspired by the Research Domain Criteria (RDoC) framework. We investigate two complementary methods: first, an in-context learning framework using Mistral-7B-Instruct with semantically retrieved few-shot examples augmented by lexicon-derived RDoC valence-arousal cues; second, a supervised multi-task learning model based on RoBERTa, where VA regression is the primary objective and RDoC-based positive/negative signal prediction serves as an auxiliary task to regularize shared representations. Experiments on English laptop and restaurant review datasets demonstrate that incorporating RDoC-inspired affetcive priors reduces RMSE, particularly in low-signal text where explicit sentiment cues are sparse.

1 Introduction

Sentiment analysis aims to quantify the affective content of text, either by assigning categorical labels (positive, negative, neutral) or by predicting continuous values along affective dimensions (Liu, 2012). A particularly challenging subtask is Aspect-Based Sentiment Analysis (ABSA), which detects sentiment expressed toward specific aspects of an entity, providing a more fine-grained understanding beyond overall document polarity (Nazir et al., 2020). Traditional ABSA approaches often frame this as a categorical classification problem, labeling sentiment as positive, negative, or neutral (Pontiki et al., 2016), but such discrete labels fail to capture the intensity and subtlety of emotions. Dimensional Aspect-Based Sentiment Analysis (DimABSA) addresses the limitations of categorical sentiment labels by predicting continuous valence-arousal (VA) scores for each aspect, offering a richer representation of affective states (Lee et al., 2026). In this work, we focus on English-language laptop and restaurant reviews from SemEval-2026 Task 3 (Lee et al., 2026), specifically Track A, Subtask 1 (DimASR – Dimensional Aspect Sentiment Regression) (Yu et al., 2026), where the objective is to predict real-valued VA scores for each aspect using both in-context learning with large language models (LLMs) and fine-tuned transformer models.

Dimensional approaches, such as the valence-arousal framework, provide a continuous representation of affective states, capturing both the positive-negative polarity (valence) and the intensity or activation level (arousal) of emotions (Russell, 1980; Mohammad et al., 2018). These approaches align naturally with psychological models of affect, including the Research Domain Criteria (RDoC) framework proposed by the National Institute of Mental Health, which structures affective processes into systems such as positive valence (reward) and negative valence (threat/loss) (Insel et al., 2010). We hypothesize that incorporating affective signals inspired by this framework into sentiment models can enhance interpretability and stabilize predictions, particularly for fine-grained or multi-aspect tasks (Mohammad et al., 2018).

Recent advancements in LLMs and transformer-based architectures have demonstrated strong capabilities for few-shot reasoning and supervised learning across diverse natural language tasks (Brown et al., 2020; Perikos and Diamantopoulos, 2024). Retrieval-based selection of semantically similar examples can further improve few-shot VA prediction (Wang et al., 2024), while multi-task learning has been shown to enhance sentiment models by incorporating auxiliary tasks such as negation detection (Barnes et al., 2021). Building on these

[*] Equal contribution.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2540–2547

insights, we propose leveraging affective signals inspired by the RDoC framework to guide both LLM-based in-context predictions and multi-task fine-tuning. These signals encode coarse valence-arousal cues that regularize predictions and promote psychologically consistent representations.

We explore this through two modeling tracks, fine-tuning, and prompting. In the prompting track, a Mistral-based large language model is augmented with RDoC-informed cues injected into the prompt to provide additional context. In the fine-tuning track, the RDoC signal derived from lexical features is incorporated as an auxiliary task within a multi-task RoBERTa architecture, where the model jointly learns valence-arousal prediction alongside RDoC-based sentiment indicators.

Our approach integrates semantic similarity-based example retrieval for LLMs and RDoC-guided multi-task learning for transformers. Experimental results on laptop and restaurant review datasets show that incorporating RDoC-inspired affective signals reduces RMSE for valence-arousal prediction compared to baselines, and ablation studies confirm the contribution of each component. By introducing psychologically informed auxiliary signals into dimensional ABSA, this work demonstrates a practical method for producing more stable, interpretable, and accurate aspect-level affect predictions. To our knowledge, this is the first attempt to explicitly incorporate RDoC-inspired affective signals into aspect-level valence–arousal regression for DimABSA.

2 Related Work

Dimensional affect modeling, grounded in the valence–arousal framework (Russell, 1980), has been applied to text-based emotion prediction using regression approaches over tweets and reviews (Mohammad et al., 2018). Large language models (LLMs) show that few-shot and in-context learning can generalize to new tasks with minimal examples (Brown et al., 2020). Instruction-tuned LLMs, such as Mistral-7B-Instruct, further improve performance by aligning outputs with natural language instructions (Jiang et al., 2023; Ouyang et al., 2022). Retrieval-based methods using lexical or embedding similarity outperform random example selection, motivating the use of semantic similarity and sentiment-guided weighting for aspect-aligned demonstrations (Wang et al., 2024).

Transformer-based models, such as RoBERTa, have been extensively fine-tuned for ABSA, demonstrating strong performance on aspect-level sentiment tasks (Perikos and Diamantopoulos, 2024). Explainability analyses reveal that these models can capture subtle interactions between words and aspects, motivating their use for nuanced VA regression. Multi-task learning has been shown to improve sentiment prediction by incorporating auxiliary tasks, such as negation detection, to regularize the main task (Barnes et al., 2021).

3 Methodology

We approach Dimensional Aspect-Based Sentiment Analysis (DimABSA) for Track A as a Valence–Arousal (VA) prediction task, guided by lightweight RDoC-inspired affective signals. Instead of implementing a full clinical RDoC model, we approximate its positive and negative valence systems via simple lexicon-based counts, capturing coarse valence polarity and arousal tendencies. These signals are integrated into both in-context learning with LLMs and supervised multi-task fine-tuned models. An overview of the full pipeline depicting both approaches is shown in Figure 1.

3.1 Datasets and Preprocessing

The experiments use English-language datasets from two domains: **laptop**: 4,076 training instances, 200 validation, and 1,000 test instances, **restaurant**: 2,284 training instances, 200 validation, and 1,000 test instances. Text and aspect terms were kept case-sensitive. No preprocessing beyond standard tokenization was applied. We did not use any additional labeled training data beyond the shared task data.

3.2 RDoC-Based Affective Signals

While the RDoC framework (Insel et al., 2010) comprises many domains and constructs, in this study, we use a minimal proxy for its Positive and Negative Valence systems, based on a small set of high-precision lexical indicators. Our intention is to explore the ability of these indicators to capture strongly polarized affective expressions and provide coarse-grained sentiment signals to support Valence-Arousal (VA) regression in cases with sparse or implicit sentiment cues. The keywords were selected as prototypical and unambiguous sentiment expressions commonly used in prior sentiment analysis literature (Hutto and Gilbert, 2014). The Positive Valence keywords are: love, excellent,

amazing, wonderful, great, best, delicious, fantastic, perfect. The Negative Valence keywords are: terrible, horrible, awful, worst, disappointed, poor, bad, disgusting. These signals are directly computed from the input text and are not learned or tuned during training.

3.3 In-Context Learning (ICL) Classifiers

Our main in-context learning approach (**ICL-FS-RDoC**), which is powered by *Mistral*, predicts using few-shot examples augmented with RDoC-inspired affective signals. Unlike fine-tuning approaches, this method does not update model weights; instead, it relies on prompt engineering and example retrieval to guide the LLM's predictions. RDoC cues are integrated into both example selection and prompt construction to improve prediction accuracy and interpretability. See Appendix A.1 and A.2 for an example prompt, and for the computational resources used, respectively.

Selection of Few-Shot Examples: For a target text and aspect, k semantically similar examples are retrieved using cosine similarity of embeddings: $\text{sim}(x, x') = \cos\big(f(x), f(x')\big)$, where x is the target input text, x' is a candidate training instance, and $f(\cdot)$ denotes the embedding function used to compute the representations of sentences. In addition to semantic similarity, the following RDoC-inspired affective cues are incorporated: valence counts (positive/negative keywords) and coarse arousal categories (high, low) derived from lexical cues. The specific arousal keywords are: **High-Arousal Keywords:** excited, thrilled, energetic, amazed, surprised, and **Low-Arousal Keywords:** bored, calm, tired, relaxed, sleepy. These signals are used to weight similarity scores, prioritizing few-shot examples whose affective signals align with the target instance. If fewer than k exact aspect matches exist, additional examples are randomly sampled from other aspects.

Prompt Construction and Reasoning: Prompt structure is: $P = D + \sum_{i=1}^{k} E_i + x$, where D is the instruction prefix, E_i are retrieved example, and x is the target input. Each example includes: Text and aspect, RDoC signals, Reasoning summary (e.g., "positive sentiment, high arousal"), and Gold VA scores. The model generates predicted VA values and optional reasoning text, which is parsed using regular expressions.

Parameters and Ensemble Predictions: Predictions are generated with a temperature of 0.1 and sampling enabled (`do_sample=True`). For each target input, three prompt instantiations are generated to form an ensemble. The final Valence–Arousal (VA) predictions are computed by averaging across these instantiations. Note that for our official leaderboard submission, we did not apply ensemble averaging; this procedure was used only for our re-run experiments to obtain more stable results.

Other ICL Variants: We implemented two ICL variants: **ICL-FS:** Standard few-shot in-context learning without RDoC cues, and **ICL-FS-CoT:** Few-shot prompting with chain-of-thought reasoning included in demonstrations but without explicit RDoC signals.

3.4 Fine-tuned (FT) Classifiers

Our main fine-tuned approach (**FT-MT-RDoC**) predicts Valence–Arousal (VA) scores using a multi-task learning framework with an auxiliary objective to encourage sentiment-aware representations. We used **RoBERTa-base** (Liu et al., 2019) as the foundational encoder, with 12 transformer layers, 768 hidden dimensions, and approximately 125M parameters.

Each input was formatted as (i.e., r_{pos} and r_{neg} mentioned above) and tokenized to a maximum length of 128 tokens. The model employs two prediction heads over the shared representation: a primary VA regression head and an auxiliary RDoC head, each consisting of a two-layer feedforward network with ReLU activation and dropout (0.1).

Multi-Task Loss Function: The total loss balances primary and auxiliary objectives: $L_{\text{total}} = L_{\text{VA}} + \lambda \times L_{\text{aux}}$ where, L_{VA}: Mean squared error for V/A predictions (primary task), L_{aux}: Mean squared error for auxiliary feature predictions, and λ: Task weighting hyperparameter (set to 0.3). A formal grid search was not performed due to computational constraints, which we acknowledge as a limitation. λ=0.3 was selected empirically following common practice in multi-task learning, where auxiliary weights in the range [0.1, 0.5] are standard. This weighting allows the model to benefit from auxiliary features without overwhelming the primary V/A objective, aligning with the shared task RMSE evaluation metric.

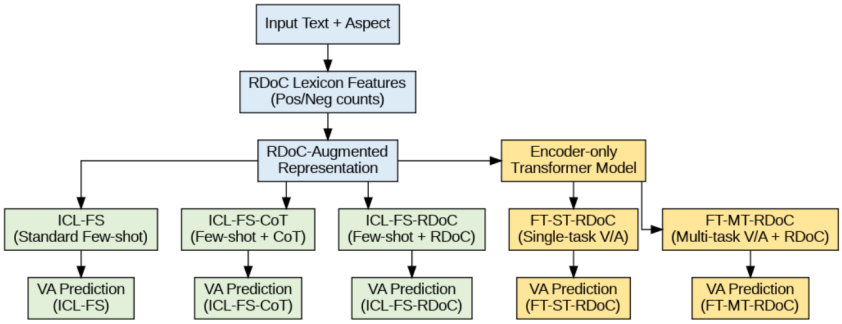

Figure 1: Overview of the proposed DimABSA methods. Input text and aspect terms are augmented with RDoC-based lexical features. The augmented representation is used in two modeling branches: (i) in-context learning (ICL) with few-shot prompting variants, and (ii) fine-tuned (FT) RoBERTa models with single-task and multi-task objectives. An Ensemble model, called ICL-FT-Ensemble, that combines ICL-FT-RDoC and FT-MT-RDoC was also developed (not shown on the figure).

3.4.1 Other FT Variant

We implemented one FT variant: **FT-ST-RDoc**: RoBERTa-base fine-tuned with a single regression head for V/A prediction. RDoC sentiment features (positive and negative keyword counts) were appended to the input as text tokens in the format of: `{The food was great. [SEP] food quality [SEP] RDoC: pos=1 neg=0}`, tokenized to a max length of 150 tokens. The model architecture consists of the RoBERTa encoder driven by a two-layer feedforward regression head with dropout of (0.1) and ReLU activation projecting from hidden size 768 to 256 to 2 output dimensions (valence, arousal). Training used the same optimizer and hyperparameters as FT-MT-RDoC, with MSE loss on V/A as the sole objective.

3.5 ICL-FT-Ensemble Model

To investigate whether combining our two modeling approaches, ICL and FT, could yield further gains, we built a weighted ensemble of the ICL-FS-RDoC and the FT-MT-RDoC models. We call it **ICL-FT-Ensemble**. For each test instance, the final predictions of valence-arousal were produced by blending the output of the two models using a scalar weight α, where a higher α gives more influence to the ICL-FS-RDoC model. The best value of α was obtained by minimizing the average RMSE across both domains on the validation set, subject to the constraint $0 \leq \alpha \leq 1$. The aspect terms were matched between the two prediction sets using normalized strings.

3.6 Experimental Setup

The training data was split 80/20 for internal training and validation. The official validation set was temporarily treated as a test set to evaluate early performance. Predictions were generated on the validation set for all five systems (**ICL-FS**, **ICL-FS-CoT**, **ICL-FS-RDoC**, **FT-ST-RDoc**, and **FT-MT-RDoC**) to identify the best-performing models. For the official leaderboard submissions, we selected ICL-FS-CoT due to its earlier availability in our experimental pipeline, alongside FT-MT-RDoC to represent the fine-tuned RDoC-enhanced approach. ICL-FS-RDoC was developed after the submission deadline and therefore was not included in the official leaderboard run, despite demonstrating stronger validation performance in our ablation studies. Upon the release of the full training and official test data, all five systems were retrained on the training set. The RMSE of the test is reported as the mean $\pm$ standard deviation over three independent runs to account for the variability in performance. See Appendix A.4 for more details.

4 Results

Unofficial Results (In-house): The results in Table 1 highlight the value of incorporating RDoC-inspired affective signals. ICL-FS-RDoC outperforms its baseline ICL counterparts (ICL-FS, ICL-FS-CoT), demonstrating that lightweight, psychologically motivated priors can meaningfully guide few-shot VA prediction. Similarly, FT-ST-RDoC, which directly incorporates RDoC valence and arousal signals as input features, achieves the lowest RMSE across both domains, confirming that affective feature augmentation enriches the model's input representations and improves stability.

The weaker performance of ICL-based models compared to fine-tuned models can be attributed to several factors. First, valence-arousal predic-

Model	laptop	restaurant
Baseline 1	2.19	2.15
Baseline 2	2.81	2.64
ICL-FS	2.20 ± 0.042	2.37 ± 0.035
ICL-FS-CoT	2.62 ± 0.065	2.68 ± 0.015
ICL-FS-RDoC	1.93 ± 0.047	2.01 ± 0.015
FT-ST-RDoC	1.43 ± 0.011	1.37 ± 0.013
FT-MT-RDoC	1.44 ± 0.042	1.39 ± 0.034
ICL-FT-Ensemble	1.42 ± 0.040	1.37 ± 0.033

Table 1: Unofficial performance of our systems on two English *test* datasets reported using RMSE (lower is better). Results are reported as mean $\pm$ standard deviation over three independent runs. Organizer Baseline 1 and 2 are provided by the DimABSA organizers. Baseline 1 is a closed-source LLM (Kimi K2 Thinking) with one-shot prompting; Baseline 2 is a fine-tuned LLM (Qwen3-14B) trained with QLoRA. Our ICL (In-context Learning) models include: ICL-FS (Few-Shot), ICL-FS-CoT (Few-Shot with Chain-of-Thought), and ICL-FS-RDoC (Few-Shot with RDoC). Our Fine-tuned (FT) models include: FT-ST-RDoc (single VA task) and FT-MT-RDoC (Multi-task with RDoC). ICL-FT-Ensemble combines ICL-FS-RDoC and FT-MT-RDoC.

tion is a continuous regression task, which is inherently more sensitive to output variability in large language models. Second, in-context learning is highly dependent on prompt formulation and example selection, making it less stable than gradient-based optimization. Finally, the absence of parameter updates limits the model's ability to fully adapt to domain-specific sentiment distributions.

As for the ICL-FT-Ensemble, across multiple runs, the optimizer consistently pushed α toward zero, settling around a ~8-13% ICL-FS-RDoC contribution. This strongly suggests that the two models make errors that are too similar for ensembling to provide any meaningful advantage. Rather than correcting each other's weaknesses, blending introduced noise from the weaker model into the stronger one's predictions. On the test set, the ICL-FT-Ensemble produced marginal improvements, achieving 1.42 ± 0.040 for laptop and 1.37 ± 0.033 for restaurant.

Official Results (Leaderboard): For the official leaderboard submissions, we selected the two representative ICL and FT models each from our initial in-house experiments prior to the DimABSA challenge deadline, ICL-FS-CoT and FT-MT-RDoC. On the leaderboard, ICL-FS-CoT scored 2.76 on laptop reviews and 2.74 on restaurant reviews. This performance is slightly worse than the results reported in Table 1, potentially due to stochasticity in few-shot generation and the absence of ensemble averaging in the original submission. FT-MT-RDoC, in contrast, achieved 1.43 on laptop reviews and 1.39 on restaurant reviews, more closely aligning with our results in Table 1 and highlighting the relatively higher stability of fine-tuned multi-task learning.

Ablation Studies and Error Analysis: To examine the contribution of RDoC-inspired affective signals, we analyze performance as a function of lexical signal strength (number of positive/negative RDoC keywords). For ICL-FS-RDoC, improvements are most pronounced in low-signal cases. In the restaurant domain, RMSE decreases from 1.95 to 1.68 (-0.27) when no RDoC keywords are present, while gains are smaller when keywords already appear (-0.07). A similar trend is observed for laptops (-0.13 in signal-absent instances), suggesting that RDoC cues act as compensatory priors when explicit sentiment markers are sparse.

For FT-MT-RDoC, signal strength correlates with lower overall error: RMSE decreases from 0.64 (no signal) to 0.48 (2 keywords). Notably, negative RDoC keywords nearly eliminate valence bias (from +0.20 to +0.00), though arousal error increases in these cases (RMSE 1.17), indicating sensitivity to intensity cues. Overall, it is evident that the RDoC features do not uniformly amplify strong sentiment, but instead stabilize predictions in low-cue settings and improve polarity grounding in multi-task learning.

5 Conclusion

We presented our UNF-BMI system for SemEval-2026 Task 3, Track A, Subtask 1 (DimASR), exploring whether psychologically grounded affective priors inspired by the Research Domain Criteria (RDoC) framework can improve aspect-level valence–arousal (VA) regression. Beyond performance gains, we provide evidence of the feasibility of integrating clinical affective theory with modern DimABSA systems. Future work will investigate richer and more scalable affective representations, including learned or lexicon-expanded RDoC signals, cross-domain and multilingual generalization, calibration of extreme valence–arousal predictions, and varied evaluation metrics to better capture alignment with human judgments.

6 Limitations

While our DimABSA approach achieves promising performance, several limitations should be noted. Our experiments are limited to English laptop and restaurant reviews, so generalization to other domains or languages remains untested. The RDoC-inspired lexicon is small and manually curated, which may miss domain-specific sentiment words, nuanced emotions, or culturally dependent expressions. In-context learning models are stochastic, leading to variability in predictions, whereas fine-tuned models are more stable. Mistral and RoBERTa are not representative of the capabilities of other LLMs and transformer models. Evaluation relies on RMSE, which is the official shared-task metric; while suitable for quantifying overall error, it does not fully capture correlations with human judgments, performance on extreme valence/arousal cases, or errors across different aspect types. Finally, although RDoC signals improve predictions, the auxiliary loss weight λ=0.3 was set heuristically rather than tuned via ablation, and their specific contribution relative to alternative affective priors or VA-specific pretrained models requires further investigation.

7 Ethical Considerations

Our work involves modeling human sentiment, which may reflect biases present in the data and in the English-language lexicons used. Dimensional sentiment predictions could be misused to influence consumer or public opinion, so we emphasize responsible research use. Despite improvements, models may still fail to capture context-dependent or mixed emotions, making human oversight essential when interpreting results.

Acknowledgments

This project was funded by the College of Computing, Engineering, and Construction, and the Graduate School of the University of North Florida. We would like to thank the DimABSA organizers for providing gold-standard data and a valuable platform for advancing this research.

References

Jeremy Barnes, Erik Velldal, and Lilja Øvrelid. 2021. Improving sentiment analysis with multi-task learning of negation. *Natural Language Engineering*, 27(2):249–269.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashevich, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad, and Saif M. Mohammad. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

Bing Liu. 2012. Sentiment analysis and opinion mining.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Isidoros Perikos and Athanasios Diamantopoulos. 2024. Explainable aspect-based sentiment analysis using

transformer models. *Big data and cognitive computing*, 8(11):141.

M Pontiki, D Galanis, H Papageorgiou, I Androutsopoulos, S Manandhar, M Al-Smadi, M Al-Ayyoub, Y Zhao, B Qin, O De Clercq, and 1 others. 2016. Semeval-2016 task 5: aspect based sentiment analysis (. pdf). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Liang-Chih Yu, Jonas Becker, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Lung-Hao Lee, Ying-Lung Lin, Jin Wang, Jan Philip Wahle, Terry Ruas, Alexander Panchenko, Ilseyar Alimova, Kai-Wei Chang, Lilian Wanzare, Nelson Odhiambo, Bela Gipp, and Saif M. Mohammad. 2026. SemEval-2026 task 3: Dimensional aspect-based sentiment analysis (DimABSA). In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

A Appendix

A.1 ICL Prompt Example

A concrete example of the few-shot in-context learning prompt used for predicting Valence–Arousal (VA) scores:

```
Predict Valence-Arousal for laptop reviews.
Format: V#A (e.g., 7.50#6.80)

Text: "the chromebook r 11 was hardly used
and only a few months old"
Aspect: chromebook r 11
VA: 4.88#4.88

Text: "the food options rule ."
Aspect: food
VA: 5.50#5.25

Text: "the laptop screen is bright and the
battery lasts all day"
Aspect: battery
VA:
```

Notes:

- The first two entries are retrieved few-shot examples from the training set, including their `Text`, `Aspect`, and gold `VA` scores.
- The last entry is the target input for which the model generates predicted VA values.
- The model uses RDoC-inspired signals for positive/negative valence and high/low arousal to select semantically similar examples.
- The predicted VA value is parsed from the generated text and clipped to the range [1.00, 9.00].

A.2 External Resources and Environment used for ICL models

We used the following for ICL models: **Pretrained Model:** *Mistral-7B-Instruct-v0.2* (Jiang et al., 2023), **Embedding Model:** *all-MiniLM-L6-v2* (Wang et al., 2020) for semantic similarity-based few-shot retrieval, **RDoC Lexicons:** Positive/negative valence keywords and high/low arousal keywords for affective priors, **Libraries:** `transformers`, `bitsandbytes`, `sentence-transformers`, `torch`, `numpy`, `pandas`, `Hugging Face accelerate`, **Hardware:** NVIDIA T4 GPU with 15GB VRAM as well as NVIDIA A100-SXM4 80GB GPU with 179.4 GB system RAM and CUDA 13.0, and **Software:** Python 3.12.12 on Linux 6.6.113+-x86_64-with-glibc2.35 and Google Colab Pro+ environment.

A.3 FT Model Training Configuration and Resources

Training used the AdamW optimizer with a learning rate of 2e-5, batch size 16, 500 warmup steps, and 5 epochs with MSE loss. The best checkpoint was selected based on validation loss and reloaded at the end of training. All experiments were conducted on an NVIDIA A100-SXM4 80GB GPU via Google Colab Pro+.

A.4 Implementation, Libraries and Availability

All experiments were conducted using Python 3.12.12, transformers v5.0.0, PyTorch v2.10.0 (CUDA 12.8), sentence-transformers v5.2.3, NumPy v2.0.2, and bitsandbytes v0.49.2. To ensure reproducibility, all experiments were run with a fixed random seed of 42 for Python, NumPy, and PyTorch, thereby controlling model initialization, data shuffling, and dropout. All code used in our experiments is publicly available at `https://github.com/VishwaaShah5/DimABSA_SemEval` to support reproducibility.