The Classics at SemEval-2026 Task 3: Combining Transformer Models and LLM-Generated Annotations for Dimensional Aspect-Based Sentiment Analysis

Rafif Alshawi [*], **Amit Raj** [*], **Aleksey Kudelya**, **Alexander Shirnin**
HSE University
Correspondence: ashirnin@hse.ru

Abstract

This paper presents an approach to the SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment Analysis. We investigate methods for moving beyond traditional categorical sentiment (e.g., positive or negative) to predict fine-grained, real-valued scores for sentiment "valence" (positivity) and "arousal" (intensity). We participate in two subtasks: predicting these scores for given aspects (Subtask 1) and extracting full sets of sentiment details, including aspects, categories, and opinions alongside their scores (Subtask 3). Our approach for the regression task involves a weighted ensemble of transformer-based encoder models. For the Russian language, we further enhance the input by using a large language model (LLM) to generate synthetic sentiment descriptions. For the extraction task, we fine-tune a decoder LLM to perform structured prediction, allowing the system to identify sentiment elements and estimate their numerical scores simultaneously.

1 Introduction

A major concern of traditional Sentiment Analysis models is their opacity and tendency to oversimplify complex human emotions into discrete and rigid categories. As these models are deployed in an increasing number of applications, the inability to capture nuanced user feedback has exposed systematic vulnerabilities in downstream decision-making (Liu, 2012). Consequently, moving beyond categorical labels to capture the true intensity and nature of opinions has been posed as a critical desideratum in the rapidly developing field of Natural Language Processing (Zhang et al., 2023). To this end, Dimensional Aspect-Based Sentiment Analysis (DimABSA) has emerged to map sentiment onto a continuous two-dimensional space of *valence* and *arousal* (Russell, 1980). Developing methods capable of such complex reasoning over text helps uncover deeper semantic relationships and improves the faithfulness of the models to the underlying human emotions.

SemEval-2026 Task 3 (Yu et al., 2026) focuses on this paradigm shift by providing a multilingual and multidomain benchmark for fine-grained sentiment evaluation. The shared task offers multiple subtasks that correspond to standard requirements in modern affective computing. These include predicting the continuous valence and arousal scores for a given aspect (Subtask 1) and extracting the complete set of sentiment rationales such as the aspect, category, opinion phrase, and their continuous scores (Subtask 3).

This paper presents methods for complex reasoning over text to address the challenges of fine-grained sentiment analysis. For Subtask 1, we employ a weighted ensemble of encoders. To mitigate vulnerabilities in the Russian track, we introduce a zero-shot augmentation technique where synthetic rationales generated by a LLM provide explicit descriptive context. For Subtask 3, we address the systematic faults of multi-step pipelines (Jing et al., 2021) by proposing a unified generative framework.

Our participation provided several key insights into how different models handle fine-grained sentiment data. We found that the generative LLM performed exceptionally well on the quadruplet prediction task, which was surprising given that encoders are usually considered more stable for regression. Additionally, our results in the Russian track show that using an LLM to generate descriptive sentiment context can significantly help smaller encoder models understand numerical ranges.

2 Background

The DimABSA shared task (Yu et al., 2026) provides a benchmark for evaluating fine-grained senti-

[*]Equal contribution.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2648–2656

ment across several domains, including hospitality, consumer electronics, and finance. The dataset for Track A consists of multilingual and multidomain samples (Lee et al., 2026), with our experiments specifically addressing the English and Russian tracks. This corpus requires models to move beyond discrete sentiment classification to capture the intensity and positivity of opinions through continuous values. We focus our participation on two specific challenges: Dimensional Aspect Sentiment Regression (Subtask 1) and Dimensional Aspect Sentiment Quad Prediction (Subtask 3).

Task Formulation The two subtasks represent different levels of complexity in sentiment analysis. In Subtask 1, the model is provided with a text and a specific aspect, and it must predict the associated valence and arousal. In Subtask 3, the model must perform an end-to-end extraction of all sentiment-bearing components. The formulations are illustrated as follows:

Subtask 1: DimASR Example

Input Text: The battery life is amazing.
Given Aspect: battery life
Output (VA): `8.50#7.20`

Subtask 3: DimASQP Example

Input Text: The battery life is amazing.
Output (Quadruplet):
(*Aspect*: battery life, *Category*: LAPTOP#BATTERY, *Opinion*:amazing, *VA*: 8.50#7.20)

Performance Metrics To account for the continuous nature of the predictions, the organizers utilize metrics that penalize the distance between predicted and ground-truth values. For Subtask 1, performance is measured using the Root Mean Square Error (RMSE) across both valence and arousal dimensions:

$$\text{RMSE}_{VA} = \sqrt{\sum_{i=1}^{N} \frac{(V_p^{(i)} - V_g^{(i)})^2 + (A_p^{(i)} - A_g^{(i)})^2}{N}} \quad (1)$$

where V_p and A_p are the predicted scores, V_g and A_g are the gold labels, and N represents the number of test instances.

For Subtask 3, the evaluation uses a Continuous F1-score (cF1). This metric unifies categorical extraction with numerical regression. A predicted quadruplet is first checked for a categorical match (Aspect, Category, and Opinion). If the categories match, the prediction is assigned a score based on a Continuous True Positive (cTP) calculation, which reduces a perfect score of 1 by the normalized Euclidean distance ($dist$) between the predicted and gold VA values:

$$\text{cTP} = 1 - \frac{\sqrt{(V_p - V_g)^2 + (A_p - A_g)^2}}{\sqrt{128}} \quad (2)$$

The final cF1 is the harmonic mean of continuous precision and recall, ensuring that models are rewarded for both identifying the correct sentiment elements and accurately estimating their emotional intensity. All predicted scores must be restricted to the range of [1, 9] and rounded to two decimal places.

3 System Methodology

This section details the methodologies developed to address the complex reasoning tasks presented in DimABSA. To accommodate the different requirements of the two subtasks, we design two independent pipelines. First, we outline our discriminative ensemble approach for Dimensional Aspect Sentiment Regression. Next, we present a generative framework for Dimensional Aspect Sentiment Quad Prediction.

3.1 Subtask 1: Dimensional Aspect Sentiment Regression

Subtask 1 requires the model to infer continuous valence and arousal scores given a specific aspect. We formulate this as a multi-target regression problem, where the overarching goal is to map the input text directly to a two-dimensional continuous sentiment space.

Encoder Ensemble We fine-tune a set of transformer-based encoder models to predict the two continuous variables simultaneously. A linear regression head is placed on top of the final hidden state of the `[CLS]` token to output the valence and arousal values. To ensure the robustness of our predictions and mitigate the variance inherent in individual models, we aggregate the outputs using a weighted ensemble strategy. The weights assigned to each model are optimized empirically based on their individual Root Mean Square Error (RMSE) performance on the development set.

LLM-based Augmentation for Russian The Russian language track presents additional com-

plexities regarding the extraction of subtle sentiment nuances. To alleviate this and provide the underlying encoder models with stronger linguistic signals, we introduce a zero-shot data augmentation step relying on an instruct-tuned LLM. For each training instance, we prompt the decoder model with the original text and instruct it to produce a short preliminary rationale describing the sentiment in terms of pleasantness and intensity. Rather than acting as a numerical score, this generated text serves as explicit descriptive context. The synthetic rationale is then concatenated with the original input text before being processed by the encoder models. We provide the whole system prompt used for this generation step in Appendix B.1.

3.2 Subtask 3: Dimensional Aspect Sentiment Quad Prediction

Subtask 3 involves a more complex reasoning process, requiring the simultaneous extraction of categorical elements (aspect, category, opinion) and the prediction of continuous elements (valence and arousal). Rather than relying on a multi-step pipeline of distinct models, which often suffers from error propagation, we formulate this as a unified structured generation task.

Generative Extraction We employ a decoder LLM, fine-tuning it specifically on task-specific, domain-based data. The model is trained to process the input text and directly generate a structured sequence containing the complete (Aspect, Category, Opinion, Valence-Arousal) quadruplets.

By training the model to jointly estimate the valence and arousal scores alongside the extraction of the aspect and opinion terms, the decoder model learns the underlying relationship between the descriptive textual phrases and their corresponding emotional intensity. Consequently, the model maps the extracted opinion directly to the numerical valence-arousal space in a single forward pass. This single-model architecture successfully avoids the systematic faults and compounding errors often observed in standard systems where extraction and regression are treated as isolated steps. The full system prompt used for this task is provided in Appendix B.2.

4 Experiments

Overview We design a series of experiments to evaluate the proposed methods for dimensional sentiment analysis. We investigate the performance of established transformer-based language models, selected based on their efficacy on standard natural language processing benchmarks. For the English track of Subtask 1, we utilize *RoBERTa-Large*[1], *RoBERTa-Base*[2] (Liu et al., 2019) and *DeBERTaV3-Large*[3] (He et al., 2023) models from the transformers library (Wolf et al., 2020).

For the Russian track, we employ *XLM-RoBERTa-Base*[4] and *XLM-RoBERTa-Large*[5] (Conneau et al., 2020) to accommodate the multilingual requirements of the dataset. First, we detail the fine-tuning procedures and hyperparameter selection for the encoder models. Second, we present the data augmentation strategy developed for the Russian track and outline the final ensemble configuration.

4.1 Subtask 1: Dimensional Aspect Sentiment Regression

Encoder Fine-Tuning We fine-tune the encoder models to predict continuous valence and arousal scores. The models are optimized using the Adam optimizer (Kingma and Ba, 2015). All model parameters are updated during training. Our preliminary experiments suggest that training beyond five epochs leads to overfitting on the training set. Therefore, we limit the maximum number of epochs to five in all subsequent experiments.

To ensure hyperparameter selection, we allocate 10% of the provided training data as an internal validation set to monitor performance during the learning phase. The official development set is strictly reserved for final model evaluation and ensemble weighting. For the RoBERTa and XLM-RoBERTa architectures, we experimented with various learning rates and observed that a learning rate of `2e-5` yields the best results. For the DeBERTaV3-Large model, we utilize a learning rate of `5e-6`, adhering to the architectural recommendations of the original authors (He et al., 2023), which we also empirically verified to be optimal for our data.

Ensemble Strategy To formulate our final submission for the English track, we aggregate the predictions of the three models using a weighted average. To minimize the risk of overfitting the relatively small development dataset, we avoid an

[1] hf.co/FacebookAI/roberta-large
[2] hf.co/FacebookAI/roberta-base
[3] hf.co/microsoft/deberta-v3-large
[4] hf.co/FacebookAI/xlm-roberta-base
[5] hf.co/FacebookAI/xlm-roberta-large

exhaustive combinatorial search for optimal ensemble weights. Instead, we assign weights of 0.35 to the RoBERTa models and 0.30 to DeBERTaV3-Large, reflecting their relative performance on the development set.

Data Augmentation Our preliminary experiments indicate that the proposed data augmentation technique does not yield significant improvements for the English track. Conversely, applying this method to the Russian dataset provides a measurable performance advantage. We utilize the *Llama-3.2-3B-Instruct*[6] to generate synthetic sentiment rationales for the training instances. We do not fine-tune it to avoid overfitting; it is used in a zero-shot setting. To ensure computational efficiency during this step, we generated the texts using the vLLM framework (Kwon et al., 2023).

We experimented with various prompts to instruct the decoder model. Based on a manual review of the generated rationales, we select a prompt that ensures structural consistency and qualitative plausibility of the sentiment descriptions. The prompt design was further assisted by a proprietary LLM *Claude-Sonnet-4.6*[7].

4.2 Subtask 3: Dimensional Aspect Sentiment Quad Prediction

Decoder Fine-Tuning For the structured extraction of sentiment quadruplets, we investigate the efficacy of instruction-tuned generative models. Constrained by available computational resources, we prioritize a parameter-efficient fine-tuning strategy, namely Low-Rank Adaptation (LoRA) tuning (Hu et al., 2021). We utilize the `unsloth` library (Daniel Han and team, 2023) to accelerate the training process and reduce memory overhead. Our primary experiments employ *Llama-3.1-8B-bnb-4bit*[8] model. We configure the training procedure to 500 steps, utilizing a learning rate of `1e-4` and a batch size of 8. To establish a comparative baseline, we also evaluate the *DeepSeek-R1-Distill-Llama-8B-bnb-4bit*[9] (Guo et al., 2025) under identical hyperparameters.

Ablation and Post-Competition Setup Given the continuous nature of the valence and arousal scores, we initially hypothesized that decoder models might lack the underlying capabilities required for precise numerical regression. To investigate this, we designed an ablation experiment featuring a hybrid pipeline. In this setup, a dedicated RoBERTa-Base encoder model, specifically trained for regression, is utilized to predict the continuous scores for the extracted text spans, thereby overriding the decoder model's numerical outputs.

Furthermore, because an exhaustive hyperparameter search was not computationally feasible during the active phase of the shared task, we conducted a systematic post-competition study on the official test dataset. We investigate variations in the learning rate, the impact of alternative system prompts, and the efficacy of different base architectures to validate our initial architectural choices.

5 Results and Discussion

5.1 Subtask 1 Results

System	Language	Domain	Test RMSE
Our Final System	English	Restaurant	1.23
Baseline (Kimi-K2 Thinking)	English	Restaurant	2.14
Baseline (Qwen-3 14B)	English	Restaurant	2.64
Our Final System	English	Laptop	1.33
Baseline (Kimi-K2 Thinking)	English	Laptop	2.18
Baseline (Qwen-3 14B)	English	Laptop	2.80
Our Final System	Russian	Restaurant	1.64
Baseline (Kimi-K2 Thinking)	Russian	Restaurant	1.77
Baseline (Qwen-3 14B)	Russian	Restaurant	2.15

Table 1: Performance of our final system on the Subtask 1 test set compared to the official organizer baselines. Lower RMSE values indicate better predictive accuracy.

The performance of our individual regression models on the development set is detailed in Table 2. Upon evaluating the English track (restaurant domain), we observe RMSE scores of approximately 1.15 and 1.17 for the RoBERTa models, compared to 1.23 for DeBERTaV3-Large. The models behavior is similar in the laptop domain. We hypothesize that this performance disparity is closely tied to the textual nature of the dataset; the predominantly short sentence structures align favorably with RoBERTa's pre-training distribution, granting it an empirical advantage over DeBERTa in this specific context.

For the Russian track, empirical results demonstrate the effectiveness of our synthetic rationale generation. The inclusion of the augmented context reduces the RMSE of the XLM-RoBERTa models from the 1.43–1.45 range down to 1.40–1.41. Due to the limited size of the Russian development set, we prioritize robustness in our final submission.

[6] `hf.co/meta-llama/Llama-3.2-3B-Instruct`
[7] `claude.com/docs/en/about-claude/models`
[8] `hf.co/Meta-Llama-3.1-8B-bnb-4bit`
[9] `hf.co/DeepSeek-R1-Distill-Llama-8B-bnb-4bit`

Language	Domain	Model Architecture	Data	Dev RMSE
English	Restaurant	RoBERTa-Base	Original	**1.15**
English	Restaurant	RoBERTa-Large	Original	1.17
English	Restaurant	DeBERTaV3-Large	Original	1.23
English	Laptop	RoBERTa-Base	Original	**1.24**
English	Laptop	RoBERTa-Large	Original	1.27
English	Laptop	DeBERTaV3-Large	Original	1.33
Russian	Restaurant	XLM-RoBERTa-Base	Original	1.44
Russian	Restaurant	XLM-RoBERTa-Large	Original	1.48
Russian	Restaurant	XLM-RoBERTa-Base	Augmented	**1.41**
Russian	Restaurant	XLM-RoBERTa-Large	Augmented	1.43

Table 2: Performance of individual encoder models on the Subtask 1 development set. Lower RMSE values indicate better predictive accuracy. **Bold** numbers indicate the best performance for each task.

Relying on the four-model ensemble provides a stable consensus prediction that mitigates the variance of any single architecture.

5.2 Subtask 3 Results

The outcomes of our extraction methodologies and post-competition analysis are presented in Table 3. Most notably, our ablation study reveals that the hybrid pipeline (yielding a cF1 of 0.3024) underperforms the end-to-end LLM strategy (0.3072). This contradicts our initial hypothesis that encoders are strictly superior for numerical regression, suggesting that the LLM successfully captures the relationship between the extracted textual terms and their emotional intensity.

Our post-competition study further validates the hyperparameter choices made during the active phase. We find that increasing the learning rate to `2e-4` decreases the overall cF1 score to 0.2917, indicating suboptimal convergence. Similarly, modifying the prompt structure to alternative formats severely degrades performance, resulting in a score of 0.2716, underscoring the high sensitivity of instruction-tuned LLMs to prompt phrasing. Finally, the DeepSeek-R1 distillation model exhibits inferior performance compared to the quantized Llama-3.1-8B baseline, achieving a score of 0.2960. Ultimately, the analysis confirms that our initial submission configuration remains a better solution for dimensional sentiment quad prediction.

Architectural Choices Reflecting on the differing architectures employed across the two subtasks, we address the possibility of applying a unified generative framework to the regression challenge in Subtask 1. During the system development phase, we prioritized Subtask 1 before addressing the Subtask 3. At that initial stage, we hypothesized that decoders might lack the underlying capabilities required for precise continuous numerical approximation. Consequently, we relied on established encoder ensembles equipped with regression heads to ensure robust predictions. However, our subsequent ablation experiments in Subtask 3 contradict this initial assumption. The empirical results demonstrate that instruction-tuned decoders successfully capture the relationship between textual terms and their emotional intensity without requiring a separate regression module. Given the strict timeline of the shared task and limited computational resources, we were unable to retrospectively adapt and fine-tune a unified generative framework for Subtask 1. Because the chosen LLM exhibits a capacity to predict continuous valence and arousal scores, extending this purely generative approach to standard dimensional regression represents a highly promising direction for future work.

6 Conclusion

This paper presents our system submitted to SemEval-2026 Task 3 on Dimensional Aspect-Based Sentiment Analysis. Our solution demonstrates competitive efficacy, placing 10th and 11th in the English laptop and restaurant domains for Subtask 1, 16th in the Russian track, and securing 7th place out of 18 in Subtask 3. For the regression challenges of Subtask 1, we adopted a classic weighted ensemble of transformer-based encoders, which we further enhanced specifically in the Russian track by introducing zero-shot synthetic rationale generation. For the structured quad prediction in Subtask 3, we transitioned to a purely generative framework utilizing a LLM. The primary advantage of employing a LLM for this complex extraction is its capacity to capture the relationship between textual phrases and their corresponding emotional

Setup	Domain	Learning rate	Test cF1 Score
Llama-3.1-8B-Instruct	Laptop	1e-4	**0.3072**
Llama + RoBERTa-Base	Laptop	1e-4 and 2e-5	0.3024
Llama-3.1-8B-Instruct	Laptop	2e-4	0.2917
Llama-3.1-8B-Instruct (Alternative Prompt)	Laptop	1e-4	0.2716
DeepSeek-R1-Distill-Llama-8B	Laptop	1e-4	0.2960
Baseline (Kimi-K2 Thinking)	Laptop	-	0.2795
Baseline (Qwen-3 14B)	Laptop	-	0.1529

Table 3: Post-competition evaluation on the Subtask 3 official test set. The single LLM from our official submission achieves the highest performance across all tested configurations. Higher cF1 values indicate better predictive accuracy. **Bold** numbers indicate the best performance for the task.

intensity.

We believe this methodology holds significant potential for advancing affective computing. Future work will investigate the scalability of synthetic rationale generation across diverse languages and aim to enhance the overall system robustness against severe domain-shift scenarios.

Limitations

Despite these advances, our evaluation has been limited to relatively small models, leaving the performance of current proprietary LLMs unknown. Furthermore, it remains unclear whether these algorithms will scale effectively to larger datasets with a large number of domains and longer reviews. Another critical challenge is the sensitivity to hyperparameter choices and system prompt design, which demonstrate a significant impact on the predictions. Future work will explore these limitations, focusing on larger models while addressing challenges in hyperparameter selection.

Acknowledgments

This work is an output of a research project (HSE-BR-2025-025) implemented as part of the Basic Research Program at HSE University. We acknowledge the computational resources of HSE University's HPC facilities.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu

Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Hongjiang Jing, Zuchao Li, Hai Zhao, and Shu Jiang. 2021. Seeking common but distinguishing difference, a joint aspect-based sentiment analysis model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3910–3922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Springer Cham.

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.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges . *IEEE Transactions on Knowledge & Data Engineering*, 35(11):11019–11038.

A Data augmentation visualization

We provide a schematic representation of the zero-shot data augmentation strategy employed for the Russian language track in Figure 1. To facilitate readability and broader accessibility, the illustrative examples within the figure are presented in English, although the actual experiments were conducted on Russian instances. The pipeline depicts the generation of synthetic sentiment rationales via an instruct-tuned LLM and their subsequent concatenation with the original input text.

B LLM System Prompts

B.1 Prompt for Subtask 1

The following prompt is used to instruct the language model for generating explicit descriptive context:

System Prompt

You are an emotion analysis expert. Given a text and a specific aspect from that text, provide a brief emotional characterization that captures the sentiment toward that aspect.

Use the Valence-Arousal model:
- Valence: negative (1) neutral (5) positive (9)
- Arousal: low (1) medium (5) high (9)

Key emotion terms by quadrant:
High-Arousal Positive: delighted, excited, happy, thrilled, enthusiastic
Low-Arousal Positive: content, calm, relaxed, satisfied, pleasant
High-Arousal Negative: angry, tense, frustrated, annoyed, irritated

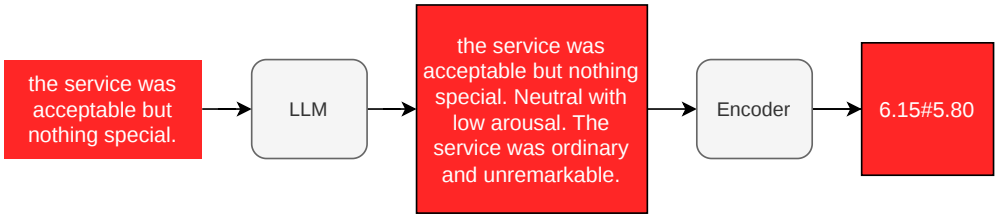

Figure 1: Visualization of the inference pipeline for Subtask 1 (Russian track). English text is used here for illustrative purposes.

Low-Arousal Negative: depressed, bored, tired, disappointed, indifferent
Neutral/Mild: unremarkable, ordinary, neutral, forgettable, acceptable

Your task:
1. Analyze how the text describes or relates to the given aspect
2. First state the emotional classification using the terms above
3. Then provide one brief sentence describing the sentiment
4. Be precise - if sentiment is unclear or neutral, say so
5. Use English language only

Output format: Classification first, then description. Total 1-2 sentences, no additional explanation.

Examples:

Text: "their sake list was extensive, but we were looking for purple haze, which wasn't listed but made for us upon request!"
Aspect: "sake list"
Output: Positive and excited. The sake list is extensive and impressive.

Text: "the spicy tuna roll was unusually good and the rock shrimp tempura was awesome, great appetizer to share!"
Aspect: "spicy tuna roll"
Output: Highly positive and delighted. The roll is unusually good and exceeds expectations.

Text: "we love the pink pony."
Aspect: "pink pony"
Output: Positive and content. The restaurant is loved and valued.

Text: "this place has got to be the best japanese restaurant in the new york area."
Aspect: "place"
Output: Extremely positive and thrilled. The restaurant is considered the absolute best.

Text: "the service was acceptable but nothing special."
Aspect: "service"
Output: Neutral with low arousal. The service was ordinary and unremarkable.

B.2 Prompt for Subtask 3

The prompt below instructs the language model to predict valence and arousal scores and to extract aspect and opinion terms. **Bold** text corresponds to Markdown-style emphasis (i.e., **text**) present in the original prompt.

System Prompt

You are an expert Linguist specializing in Aspect-Based Sentiment Analysis (ABSA). Your task is to extract highly accurate (A, C, O, VA) quadruplets from the given text.

Core Extraction Rules:
1. **Aspect (A):** The specific feature or entity mentioned. Must match the input text casing exactly.
2. **Category (C):** Classify the aspect using the "ENTITYATTRIBUTE" schema below. Use ONLY these labels. Must be UPPERCASE.
3. **Opinion (O):** The specific word/phrase used to express the sentiment. Match the input casing exactly.
4. **Valence-Arousal (VA):**
- **Valence:** 1.00 (Extremely Negative) to 9.00 (Extremely Positive). 5.00 is Neutral.
- **Arousal:** 1.00 (Calm/Sleepy) to 9.00 (Excited/Angry). 5.00 is Moderate.
- Format as "V.VV#A.AA" (always 2 decimal places).

Label Schema Constraints:
- **Valid Entities**: laptop_entity
- **Valid Attributes**: laptop_attribute

Step-by-Step Reasoning:
Step 1: Identify all sentiment-bearing phrases and the aspects they refer to.
Step 2: Map each aspect to the most relevant ENTITY and ATTRIBUTE from the schema.
Step 3: Determine the numerical Valence (positivity) and Arousal (intensity) of the opinion.
Step 4: Format as a list of (A, C, O, VA) quadruplets.

—
Examples:

Input: The screen is incredibly bright and vibrant, but the price is a bit steep.
Output: (screen, DISPLAY#QUALITY, incredibly bright and vibrant, 8.50#7.20), (price, LAPTOP#PRICE, a bit steep, 3.20#5.50)

Input: The keyboard feels mushy and the battery drains too fast.
Output: (keyboard, KEYBOARD#USABILITY, feels mushy, 3.00#4.50), (battery, BATTERY#OPERATION_PERFORMANCE, drains too fast, 2.00#6.50)

—

Target Task:
Input:
`{input_text}`

Output: