

BertKittens at SemEval-2026 Task 3: Multi-Domain Aspect Sentiment with BERT/DeBERTa Ensembles for VA Regression and Aspect–Opinion–VA Triplets

Arseny Sukhodolsky
Individual researcher
ars1suhod@gmail.com

Tatyana Yanshina
Individual researcher
t.ianshina99@gmail.com

Ruslan Salimgareev
Individual researcher
russal2010@gmail.com

Abstract

We describe our submission to SemEval-2026 Task 3 (Track A) (Yu et al., 2026), which focuses on quantitative tone assessment and aspect-based opinion extraction.

Our system is built on transformer encoders (BERT and DeBERTa) fine-tuned in a multi-task learning framework. For the regression subtask (evaluated with RMSE), we jointly predict Valence–Arousal (VA) scores and token-level opinion spans using a shared encoder with task-specific output heads. This formulation introduces auxiliary supervision at the token level, which stabilizes training and improves regression accuracy compared to single-task optimization.

When gold abstracts and opinion annotations are provided, our models achieve strong performance. However, in fully end-to-end settings requiring automatic span extraction, performance degrades substantially due to error propagation from token-level predictions.

Our findings highlight the benefits of joint affective regression and span modeling, while exposing the limitations of transformer-based sequence labeling under strict end-to-end evaluation constraints.

1 Introduction

In traditional psychology, sentiment is interpreted as continuous, fine-grained dimensions: valence (spanning from negative to positive) and arousal (ranging from low-energy or sluggish to highly excited states) (Russell, 2003). However, most of the modern approaches are based on categorical classification of sentiment (Zhang et al., 2022). Quantitative modeling of tone and aspect salience remains a challenging problem in contemporary NLP. However, dimensional sentiment analysis was successfully applied to multiple tasks recently (Yu et al., 2026). SemEval-2026 Task 3 addresses this challenge by requiring systems to jointly assess affective intensity and extract aspect–opinion structures.

Our modeling strategy varies across subtasks. For the regression setting where predefined aspects are provided, we adopt a joint learning approach that simultaneously predicts VA scores and token-level opinion spans using a shared encoder. This auxiliary span supervision improves regression stability and accuracy.

For subtasks without gold aspect and opinion annotations, we employ a sequential pipeline architecture. First, an ensemble of models predicts candidate aspects. Conditioned on these predictions, a second model extracts corresponding opinion spans. Finally, a regression model estimates VA scores based on the predicted aspect–opinion pairs. For configurations requiring entity–attribute prediction, we train additional classifiers operating on the review text and extracted spans.

On the primary regression subtask, our system achieves strong performance on English test data, ranking in the top 3 and top 2 for the laptop and restaurant domains, respectively. Although absolute RMSE values for Japanese were competitive, our ranking was lower, suggesting stronger competition in that setting. Notably, performance was substantially higher in the hotel domain, where gold opinion annotations were available, indicating that span supervision plays a critical role.

In contrast, results on fully end-to-end subtasks were considerably weaker. Although our systems generally outperformed the official baseline, they ranked near the bottom of the leaderboard. The primary limitation stems from error propagation in the sequential pipeline: aspect and opinion extraction achieved an macro-F1 score of approximately 0.85 for English, and inaccuracies at early stages directly degraded downstream VA predictions. These findings highlight the sensitivity of affective regression to upstream span extraction quality in end-to-end scenarios.

2 Background

The organizers provide datasets for six languages: English (eng), Japanese (jpn), Russian (rus), Ukrainian (ukr), Tatar (tat) and Chinese (zho) 1 (Lee et al., 2026). For most languages, more than one topic was presented: reviews of computer equipment, financial services, hotels, and restaurants were presented in various variations. For most datasets in the training section, aspects, opinions corresponding to these aspects, and categories were also proposed.

For most languages, multiple domains are included, such as reviews of computer equipment, financial services, hotels, and restaurants. Training datasets generally provide annotations for aspects, corresponding opinions, and categories.

SemEval-2026 Task 3 consists of three subtasks:

1. Predicting Valence and Arousal (VA) for a given set of aspects,
2. Identifying aspects and their opinions and predicting VA,
3. Predicting quadruplets of aspect–opinion–VA–Entity&Attribute (EA).

In our work, we focus on subtasks 1 and 2 for the English, Japanese, Russian, and Tatar datasets.

3 System Overview

Across all subtasks, we employed BERT- and DeBERTa-based encoders. Here, we describe in detail the predicted outputs and the architectural design of task-specific heads.

3.1 Subtask 1: VA Prediction for Given Aspects

Subtask 1 was treated as a regression problem (evaluated using RMSE) for English, Japanese, Russian, and Tatar. We adopted a joint multi-task learning setup, where, given an input sentence and a predefined aspect, the model simultaneously predicts:

1. token-level opinion spans corresponding to the given aspect, and
2. continuous Valence and Arousal (VA) scores.

A shared transformer encoder produces contextualized token representations, which are then fed into two task-specific heads:

- a token classification head for opinion span prediction, and

- a regression head operating on the pooled sentence representation for VA estimation.

This formulation allows opinion span supervision to act as an auxiliary signal, guiding the encoder to attend to aspect-relevant sentiment tokens and improving VA prediction accuracy.

3.2 Subtask 2: Aspect–Opinion–VA Extraction

Subtask 2 requires fully automatic extraction of aspects, associated opinions, and subsequent VA prediction from raw reviews. We formulate this subtask as a staged, multi-component pipeline:

1. Aspect Extraction: a token classification model identifies candidate aspect spans from the input sentence.
2. Opinion Extraction: a separate model predicts opinion spans conditioned on the sentence and the candidate aspect, which is also provided as input to the model.
3. VA Regression: a regression model predicts Valence and Arousal scores conditioned on the sentence, the extracted aspect, and the corresponding opinion span.

Each component employs an independently fine-tuned encoder with a task-specific head, forming the pipeline:

Aspect \rightarrow Opinion \rightarrow VA

This sequential design enables aspect-aware opinion extraction and accurate downstream VA prediction while maintaining modularity across components.

Although we obtained some preliminary results for Subtask 3, due to data loss and time constraints, we do not report them here. Given the small size of the datasets and the large variety of categories, achieving meaningful results for this subtask would have required extensive data augmentation.

4 Experiments and Results

For English, we employed the [microsoft/deberta-v3-large](#) model (He et al., 2023), and for Japanese, the [globis-university/deberta-v3-japanese-large](#) model (Team, 2023). For Russian, we employed the DeepPavlov RuBERT model (Kuratov and Arhipov, 2019). For Tatar, we experimented

Dataset	Source(s)	Subtask	Train s/t	Dev s/t	Test s/t	Total s/t
eng-rest	ACOS Yelp Open Dataset	ST1 ST2—3	2284/3659	200/340 200/408	1000/1504 1000/2129	3484/5503 3484/6196
eng-lap	ACOS Amazon Reviews 2023	ST1 ST2—3	4076/5773	200/275 200/317	1000/1421 1000/1975	5276/7469 5276/8065
jpn-hot	Rakuten Travel	ST1 ST2—3	1600/2846	200/284 200/364	800/1092 800/1443	2600/4222 2600/4653
jpn-fin	chABSA; EDINET	ST1	1024/1672	200/319	800/1302	2024/3293
rus-rest	SemEval’16	ST1 ST2—3	1240/2487	56/81 48/102	1072/1637 630/1310	2368/4205 1918/3899
tat-rest	SemEval’16 (MT)	ST1 ST2—3	1240/2487	56/81 48/102	1072/1637 630/1310	2368/4205 1918/3899
ukr-rest	SemEval’16 (MT)	ST1 ST2—3	1240/2487	56/81 48/102	1072/1637 630/1310	2368/4205 1918/3899
zho-rest	SIGHAN-2024 Google Reviews; PTT	ST1 ST2—3	6050/8523	225/416 300/761	1000/1929 1000/2861	7275/10868 7350/12145
zho-lap	Mobile01	ST1 ST2—3	3490/6502	261/431 300/551	1000/2662 1000/2798	4751/9595 4790/9851
zho-fin	MOPS	ST1	1000/2633	200/563	842/2354	2042/5550

Table 1: Overview of the DimABSA datasets (s/t = sentences/tuples)

with `/hplt_bert_base_2_0_tat-Cyrl` and `bert-base-multilingual-cased` (Devlin et al., 2019); the final results were produced using the BERT model, which performed substantially better.

For all languages, training was performed on the union of the official train and development splits, with 20% of the resulting data reserved for validation. To enable prediction recovery in all subtasks, offset mapping was applied during tokenization. During a preliminary stage, we also evaluated base BERT models for English and Japanese, but they showed significantly worse performance.

4.1 Subtask 1

4.1.1 Experimental setup

For English, the training data was constructed by merging instances from all topic-specific datasets within the corresponding subtask. Preliminary experiments showed that joint training consistently outperformed topic-specific training. For Russian, back-translation augmentation using Google Translator was applied to 50% of the training data. For all subtasks, a dropout rate of 0.2 was applied.

4.1.2 Results

The final submission results of the ensemble are listed in Table 2. The main weakness of our approach was arousal prediction. We hypothesize that

valence is easier to predict because it is more lexically explicit. This hypothesis is supported by the finance dataset, which lacked opinion annotations; in this case, the model learned to predict only VA without benefiting from transfer learning.

The hyperparameters tuned included the number of unfrozen layers, the weight of the opinion loss, the learning rate, and the weight of non-opinion tokens. For the regression head, we used `MSELoss`, while the opinion head used `CrossEntropyLoss`. A limitation of our approach was that the best hyperparameters were selected based on $\sqrt{\text{MSE}}$, which slightly differs from RMSE, the final evaluation metric. Full grid search results are provided in the supplementary material, including additional experiments that were not included in this report.

Overall, the model captures the general distribution of VA scores well but tends to underestimate extreme values and the overall spread, particularly in multi-aspect sentences 8 1. Analysis of the distributions suggests that separate training per domain may yield more accurate results, as review structure varies significantly across domains. However, preliminary experiments on the task data indicate that domain-specific models did not improve RMSE, likely due to limited dataset size.

Language	Domain	Model	Rank	RMSE_VA	PCC_V	PCC_A
ENG	laptop	microsoft/deberta-v3-large	3	1.2769	0.8613	0.5335
	restaurant	microsoft/deberta-v3-large	2	1.1812	0.8926	0.6296
JPN	finance	globis-university/deberta-v3-japanese-large	14	0.9675	0.8074	0.1157
	hotel	globis-university/deberta-v3-japanese-large	12	0.7267	0.9278	0.6156
RUS	restaurant	DeepPavlov/rubert-base-cased-conversational	12	1.5828	0.8506	0.5777
TAT	restaurant	bert-base-multilingual-cased	18	2.1301	0.5622	0.2922

Table 2: Results by language and domain, Subtask 1

Table 3: Best hyperparameter configurations for Subtask 1 (RMSE)

Language / Model	lr	n_unfrozen_layers	lambda	zero_weight
English (VA)	5e-05	6	7.5	0.6
Japanese	0.0001	6	5.0	0.6
Russian	5e-05	6	3.0	0.3
Tatar	0.0001	4	7.5	0.5

4.2 Subtask 2

4.2.1 Experimental setup

For English, the training data was constructed by merging instances from all topic-specific datasets within the corresponding subtask. Preliminary experiments showed that joint training consistently outperformed topic-specific training. For Russian, back-translation augmentation using Google Translator was applied to 50% of the training data. For all subtasks, a dropout rate of 0.2 was applied.

Subtask 2 involved training three consecutive models for aspect recognition, opinion recognition, and VA regression. In principle, different models could be fine-tuned for each component. However, for English and Japanese, preliminary experiments did not identify models outperforming the DeBERTa models used in Subtask 1. For aspect extraction, bert-base-multilingual-uncased achieved similar performance to the Japanese DeBERTa model, suggesting substantial room for improvement in model selection and architecture.

4.2.2 Results

The final submission results of the ensemble are listed in Table 4.

The main limitation of our approach was the bottleneck effect. The macro-F1-score for aspect token prediction was approximately 0.85 for English 9 and 0.80 for Japanese. Since all subsequent predictions rely on the predicted aspects, errors at this stage are crucial. However, for some instances (e.g., the second example in Figure 2), subsequent errors are less critical.

The Japanese DeBERTa model achieved surprisingly good results for opinion tagging, with an F1-score around 0.87, slightly outperforming the English DeBERTa model. Nevertheless, it exhibited lower performance for aspect extraction. This discrepancy may be due to structural characteristics of the Japanese language: words describing object qualities tend to follow a consistent grammatical pattern, which facilitates opinion identification.

Overall, these results highlight the critical role of accurate aspect extraction for downstream VA prediction, as well as the potential for language-specific characteristics to influence opinion identification. Future work could explore alternative architectures, targeted data augmentation, and language-specific tokenization strategies to further improve both aspect and opinion extraction performance.

Language	Domain	Model	Rank	cF1
ENG	laptop	microsoft/deberta-v3-large	20	0.4469
	restaurant	microsoft/deberta-v3-large	17	0.5628
JPN	hotel	deberta-v3-japanese-large	11	0.4202
RUS	restaurant	DeepPavlov/rubert-base-cased-conversational	17	0.3137
TAT	restaurant	bert-base-multilingual-cased	16	0.1692

Table 4: DIMASTE (Triplet Extraction) - CPI scores by language and domain

Table 5: Best hyperparameter configurations for Subtask 2

Language	Model	lr	n_unfrozen_layers	lambda	zero_weight
English	Aspect	0.0001	6	—	0.6
English	Opinion	0.0001	6	—	0.6
English	VA (RMSE)	5e-05	6	—	0.6
Japanese	Aspect	0.0001	6	—	0.6
Japanese	Opinion	0.0001	6	—	0.6
Japanese	VA (RMSE)	0.0001	6	—	0.6
Russian	Aspect	3e-05	8	—	0.5
Russian	Opinion	1e-04	6	5.0	0.6
Russian	Merging	1e-04	6	5.0	0.5
Tatar	Aspect	5e-05	12	—	0.6
Tatar	Opinion	0.0001	8	—	0.6
Tatar	VA (RMSE)	5e-05	12	—	0.5

Acknowledgments

We are deeply grateful to Alexander Panchenko for proposing this task as part of the study course.

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of NAACL-HLT*.
- James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172. <https://doi.org/10.1037/0033-295x.110.1.145>.
- Wenxuan Zhang, Xin Li, Yang Deng, and Lidong Bing. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *arXiv preprint arXiv:2203.01054*. <https://arxiv.org/abs/2203.01054>.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *arXiv preprint arXiv:2111.09543*.
- Yuri Kuratov and Mikhail Arkhipov. 2019. [Adaptation of deep bidirectional multilingual transformers for russian language](#). In *Computational Linguistics and Intellectual Technologies*.
- 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.
- Globis University NLP Team. 2023. [Deberta-v3 japanese large](#). Hugging Face model repository.
- 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.

5 Supplementary

Table 6: Top-10 hyperparameter configurations from Subtask 1, jpn-hot

Rank	lr	n_unfrozen_layers	lambda	zero_weight	best_val_rmse
1	0.0001	6	5.0	0.6	0.7179
2	0.0001	8	2.5	0.6	0.7183
3	0.0001	6	2.5	0.6	0.7540
4	0.0001	6	2.5	0.5	0.7654
5	0.0001	8	5.0	0.4	0.7694
6	0.0001	8	2.5	0.4	0.7725
7	0.0001	6	2.5	0.4	0.7727
8	0.0001	6	7.5	0.6	0.7880
9	5e-05	6	2.5	0.6	0.7938
10	5e-05	8	2.5	0.6	0.8008

Table 7: Top-10 hyperparameter configurations from Subtask 1, eng

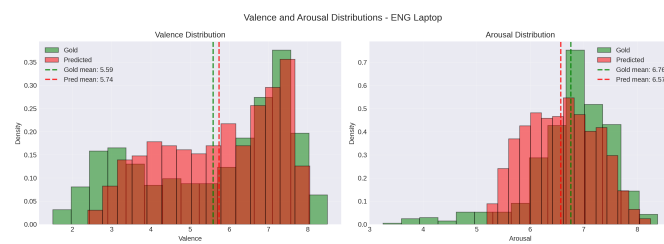
Rank	lr	n_unfrozen_layers	lambda	zero_weight	scaled_rmse
1	0.0001	6	7.5	0.6	1.0676
2	0.0001	6	2.5	0.6	1.0948
3	0.0001	8	2.5	0.5	1.0957
4	0.0001	8	7.5	0.5	1.0966
5	0.0001	6	5.0	0.5	1.0968
6	0.0001	6	2.5	0.5	1.1091
7	0.0001	8	7.5	0.6	1.1143
8	0.0001	6	7.5	0.5	1.1305
9	0.0001	8	5.0	0.5	1.1326
10	0.0001	8	2.5	0.6	1.1329

Language	Domain	Valence Gold	Valence Pred	Arousal Gold	Arousal Pred
ENG	laptop	5.585 \pm 1.901	5.738 \pm 1.480	6.759 \pm 0.798	6.572 \pm 0.656
	restaurant	6.355 \pm 1.873	6.380 \pm 1.502	7.040 \pm 0.763	6.864 \pm 0.781
JPN	finance	5.267 \pm 1.173	5.301 \pm 1.257	5.472 \pm 0.406	5.581 \pm 0.171
	hotel	5.706 \pm 1.474	5.825 \pm 1.525	6.309 \pm 0.532	6.394 \pm 0.360
RUS	restaurant	6.331 \pm 1.824	6.580 \pm 1.681	6.363 \pm 1.012	7.061 \pm 0.729
TAT	restaurant	6.406 \pm 1.836	6.760 \pm 1.519	6.385 \pm 0.997	7.119 \pm 0.509

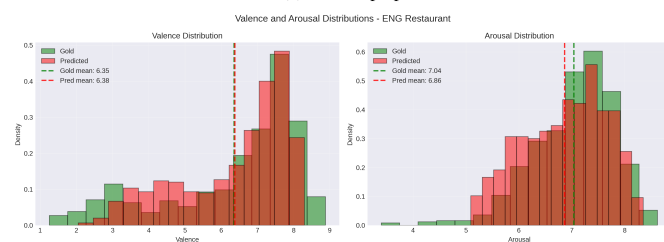
Table 8: Distribution statistics (mean \pm std) for Valence and Arousal across languages and domains

Table 9: Top hyperparameter configurations for Aspect F1: eng

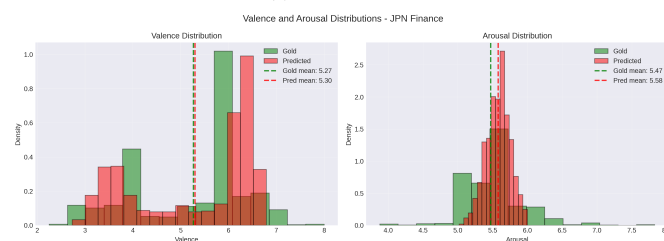
Rank	lr	n_unfrozen_layers	zero_weight	best_val_f1
1	0.0001	6	0.6	0.860962
2	0.0001	6	0.5	0.860836
3	5e-05	6	0.5	0.857975
4	5e-05	6	0.6	0.856777



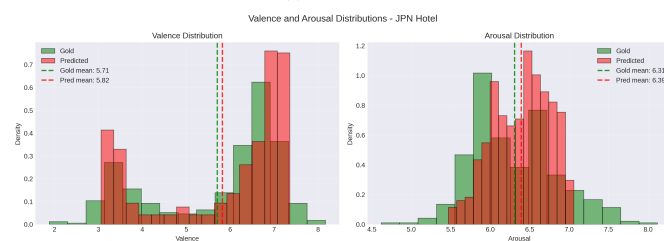
(a) ENG Laptop



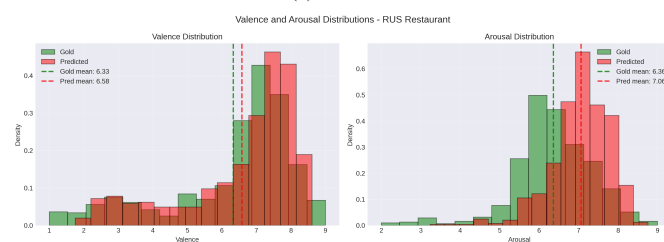
(b) ENG Restaurant



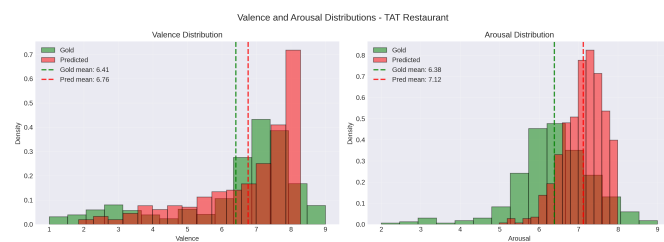
(c) JPN Finance



(d) JPN Hotel



(e) RUS Restaurant



(f) TAT Restaurant

Figure 1: Distribution of Valence and Arousal across all datasets in Subtask 1

Table 10: Top hyperparameter configurations for Opinion F1: eng

Rank	lr	n_unfrozen_layers	zero_weight	best_val_f1
1	0.0001	6	0.6	0.871769
2	5e-05	6	0.6	0.870638
3	0.0001	6	0.5	0.869604
4	5e-05	6	0.5	0.869598

Qualitative Error Analysis of Triplet Extraction

① **Negation Error (ID 179)**

As for food, I have never had a bad meal and pretty much everything on the menu.

GOLD

(food, never had a bad meal) [7.00#6.88]
(menu, pretty much everything) [7.75#7.50]

PRED

(food, bad) [6.70#7.01]
(meal, never had a bad) [7.61#7.63]

Model extracted 'bad' instead of full span.

② **Compound Aspect Split (ID 187)**

Chips & Salsa are always fresh and tasty.

GOLD

(Chips & Salsa, always fresh) [6.50#5.83]
(Chips & Salsa, tasty) [6.50#5.83]

PRED

(Chips, always fresh) [7.18#7.13]
(Salsa, tasty) [7.15#7.08]

Compound aspect split into two.

③ **Complete Aspect Loss (ID 192)**

From the friendly hostess to the waiter to the chef, everyone made us feel at home.

GOLD

(hostess, friendly) [7.33#6.83]
(waiter, friendly) [7.33#7.00]
(chef, personally came over) [7.50#7.00]

PRED

(NULL, NULL) [6.15#5.92]

All aspects lost in long sentence.

Figure 2: Examples of common errors in Subtask 2