# PUEB-DimASR at SemEval-2026 Task 3: Escaping the Mean Regression Trap with Graph-Enhanced Transformers for Dimensional Aspect-Based Sentiment Regression

**Oskar Riewe-Perła**
Poznań University of Economics and Business
Oskar.Riewe-Perla@ue.poznan.pl

**Agata Filipowska**
Poznań University of Economics and Business
Agata.Filipowska@ue.poznan.pl

## Abstract

The DimABSA shared task aims to combine dimensional analysis with aspect-based sentiment analysis (ABSA). It addresses the lack of continuous sentiment representation, as opposed to categorical labels (e.g., positive, negative, or neutral), and enriches it with an assessment of arousal. Our team's **PUEB-DimASR** investigates the "mean-regression trap" — the tendency of the MSE loss in high-dimensional sentiment tasks to over-predict values closer to the global mean. Within the paper we propose a two-step architecture. First, we enhance baseline transformers with graph convolutional networks (GCN) to capture syntactic aspect-sentiment dependencies. Second, we evaluate and recommend a hybrid loss function that combines mean squared error (MSE) and concordance correlation coefficient (CCC).

Our proposed GCN-deBERTa model consistently outperforms the baseline across six target languages. While MSE loss yields the best RMSE scores for English (0.876) and Chinese (0.546), it introduces significant variance collapse, which we successfully mitigated using the hybrid loss, achieving near-perfect distributional alignment (99.6%). Additionally, our model trained with the hybrid loss achieved the best RMSE scores for Russian (1.136), Tatar (1.207), and Ukrainian (1.178).

## 1 Introduction

Sentiment analysis (SA) is one of the most common tasks in natural language processing (NLP) that aims to assess the emotional tone of text. It emerged in the early 2000s alongside other text classification tasks. Compared to topic-based classification, SA has been described as more challenging due to the difficulty of understanding long contexts and identifying features related to specific aspects mentioned in the text (Pang et al., 2002).

Aspect-based sentiment analysis (ABSA) is expanding the traditional task by breaking down the full sentence into specific aspects and classifying them separately. The move toward finer granularity started as feature-based sentiment analysis (Hu and Liu, 2004) that later standardized as ABSA. This approach can be more accurate with longer texts that include multiple opinions on different topics. While it is common to use positive/negative/neutral labels describing the polarity of the text, dimensional sentiment analysis (DimSA) approaches the problem as a regression task providing a continuous numerical score. Additionally, it enriches the representation of valence by adding an arousal dimension. DimSA is aligning the methodology more closely with theories of psychology and affective science (Russell, 1980), which represent emotions as coordinates in a continuous space defined by valence and arousal.

Although ABSA and DimSA have existed since around two decades (Hua et al., 2024), their combination in aspect-level regression for valence (reflecting the polarity of emotion) and arousal (reflecting the emotional intensity) has not been explored equally by researchers.

Through our experimentation and methodology, we focused on addressing the "mean-regression trap", a problem where models trained using standard mean squared error (MSE) loss tend to predict values close to the global mean of the distribution, penalizing extreme predictions (Terven et al., 2025). This issue arises because MSE loss incentivizes the model to predict the expected value of the distribution, which can result in high overall scores but fails to capture the true variance of the dataset. This "flattening" effect diminishes the model's ability to capture nuances at the high and low ends of the scale. We observed this phenomenon in our experiments, where the model was skewed toward predicting values near the global mean.

In the field of DimASR research, our work explores the advantages of combining the standard transformer architecture with graph convolutional



*Proceedings of the The 20th International Workshop on Semantic Evaluation (2026)*, pages 361–366

networks (GCN) for syntactic structural modeling. A similar approach of using GCNs on dependency trees has been suggested in previous research on categorical SA (Zhang et al., 2019), we evaluate it in DimASR settings. Additionally, we propose modifying the loss function by integrating MSE and concordance correlation coefficient (CCC) to develop a nuance-aware sentiment regressor. Furthermore, we emphasize the importance of using multiple evaluation metrics (beyond RMSE) to assess model performance comprehensively.

This work is the outcome of participation of our PUEB-DimASR team in Task 3 (Yu et al., 2026), Track A, Subtask 1, for Dimensional Aspect Sentiment Regression (DimASR) at SemEval-2026. The goal of the task is to predict valence and arousal values that describe specific aspects mentioned in the text on a continuous scale from 1 to 9. Both the aspect and the text were used as input for the inference process. Separate datasets and evaluations were provided for six languages: English, Japanese, Russian, Tatar, Ukrainian, and Chinese. We trained and evaluated our model individually for each language. The proposed shared task, along with the annotated datasets (Lee et al., 2026) published by the organizers of SemEval task, represents a significant step towards standardizing DimABSA.

The code and further technical details of our experiments are available in a repository on GitHub: `https://github.com/OskarRiewe/SemEval2026_PUEB_DimASR`.

## 2 System Overview

### 2.1 Dataset

We conducted our experiment using the official SemEval-2026 Task 3 dataset (Lee et al., 2026), which encompasses six languages across diverse domains, including Laptop, Restaurant, Finance, and Hotel. As shown in Table 1, the dataset contains over 23,000 sentences and 39,000 individual aspect-level scores (as a single sentence may include multiple aspects described with different sentiments). The dataset sizes varied significantly between languages; however, since we evaluated each language individually, we did not apply any dataset balancing methods.

Each record in the dataset contains its id, text and a list of aspects. Example: {"ID": "R001", "Text": "average to good thai food, but terrible delivery." "Aspect": ["thai food", "delivery"]}

The expected output included a related id and a list of predicted scores for valence and arousal separated by "#" for each aspect mentioned in the text. Example: {"ID": "R001", "Aspect_VA":[ {"Aspect": "thai food", "VA": "6.75#6.38"}, {Aspect": "delivery", "VA": "2.88#6.62"}]}

| Language | Domains | Sentences | Aspects |
|---|---|---|---|
| English | 2 | 6,360 | 9,432 |
| Chinese | 3 | 10,540 | 17,658 |
| Japanese | 1 | 2,624 | 4,518 |
| Russian | 1 | 1,240 | 2,487 |
| Ukrainian | 1 | 1,240 | 2,487 |
| Tatar | 1 | 1,240 | 2,487 |
| **Total** | | **23,244** | **39,069** |

Table 1: Dataset Characteristics for SemEval-2026 Task 3: Distribution of Sentences and Aspect-Level Instances Across Six Languages.

### 2.2 Baseline: Transformer Regressor (deBERTa)

As a baseline, we utilize a pretrained multilingual version of the DeBERTa V3 encoder model, *mdeberta-v3-base*[1] (He et al., 2021). This model was selected due to its disentangled attention mechanism, which separates content information from positional information. This aligns well with our ABSA task, where the position of a word in a sentence is more crucial for its local semantic meaning than for understanding the full sentence. We extended the model with a stabilized regression head, and the entire architecture is fine-tuned on the DimASR task. The architecture consists of:

1. **Encoder:** the input sequence is processed by a pretrained mdeberta-v3-base to obtain contextualized token representation.
2. **Pooling:** the final hidden state of the [CLS] token is used as a sequence representation.
3. **Stabilization:** the pooled representation is normalized using the layer normalization.
4. **Projection:** the normalized representation is passed through a linear layer, followed by a sigmoid activation. We rescale the result to produce final predictions in the range of [1,9].

### 2.3 Proposed System: GCN-deBERTa

To capture the relationship between opinion words and target aspects, we propose a combination of the

[1]https://huggingface.co/microsoft/mdeberta-v3-base

previously described encoder model with a graph convolutional network (GCN) layer.

We expand the text representation using its dependency tree, which is produced by a language-specific spaCy (Honnibal and Montani, 2017) dependency parser. A dependency tree is a structural representation of a sentence where words are connected based on their grammatical relationships rather than their order in the sentence. We omit specific dependency labels (such as subject or object) and focus purely on binary information if the syntactic dependency exists. It creates a bi-directional graph structure that can be used as input to the GCN layer. As the transformer tokenizer decomposes words into subwords, we connect every subword of the head word with every subword of its dependent.

Since we use the specific [CLS] token as the final sentence representation, we bi-directionally connect it to every non-padding token in the graph. This allows it to act as a global aggregator of the GCN-processed features.

| Language | Code | spaCy Model Name |
|---|---|---|
| English | eng | en_core_web_sm |
| Chinese | zho | zh_core_web_sm |
| Japanese | jpn | ja_core_news_sm |
| Russian | rus | ru_core_news_sm |
| Ukrainian | ukr | uk_core_news_sm |
| Tatar | tat | xx_ent_wiki_sm* |

Table 2: Language-specific spaCy models used for dependency parsing. *For Tatar, a multilingual model was used due to the absence of a dedicated language core.

The proposed system extends the baseline architecture with three key components:

1. **Dependency tree (graph):** for each input, we construct a word-level dependency tree using language specific spaCy models. It represents a syntactic structure of a sentence by mapping grammatical relationships between words.

2. **Graph convolutional network:** we integrate two GCN layers on top of the deBERTA encoder to refine its representations with syntactic information from the dependency tree.

3. **Global-aware [CLS] representation:** since we rely on the [CLS] for final sequence representation, we connect it with every non-padding token in the graph.

4. **Unified regression:** the output of the final GCN layer is pooled at the [CLS] index and passed through the same regression head (LayerNorm + Projection) as used in our baseline, ensuring that any performance gains are the result of including the structural graph modeling.

### 2.4 Optimization Strategy: MSE and CCC loss

We center our experimentation on addressing a common issue known as the "mean-regression trap" which may lead the model to overlook extreme valence and arousal values in order to reduce risk.

To tackle this, we explore the use of the concordance correlation coefficient (CCC) loss (Lawrence and Lin, 1989) to ensure distributional alignment. CCC evaluates the agreement between the ground truth labels ($y$) and the predictions ($\hat{y}$) by combining Pearson's correlation coefficient ($\rho$) with the squared differences of their means and variances. The coefficient is defined as follows:

$$\rho_{ccc} = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}$$

Unlike MSE, CCC loss penalizes the lack of variance in the model's output, thereby preventing the prediction of only safe values around the global mean. The CCC ranges from -1 to 1, representing a scale from a perfect inverse correlation to perfect agreement between the true and predicted values.

We evaluate our model using both MSE and CCC metrics and propose a hybrid approach that combines the two. The purpose of this loss function is to merge the stability of MSE with the nuance-preserving qualities of CCC. The hybrid loss function is defined as follows:

$$\mathcal{L}_{Hybrid} = \mathcal{L}_{MSE} + (1 - \rho_{ccc})$$

## 3 Experimental Setup

We evaluate our models (baseline and GCN-deBERTa) using 5-fold cross-validation across all six languages and all versions of the loss function (MSE, CCC, Hybrid). To ensure reproducibility, we applied a fixed random seed (42). For each fold, we utilized the official training datasets provided for SemEval-2026 Task 3, Track A.

Our implementation is based on PyTorch and uses pretrained models from the Hugging Face Transformers library. We trained the models on two RTX A4500 GPUs and one RTX 2000 Ada GPU.

| Hyperparameter | Value |
|---|---|
| Backbone Model | mDeBERTa-v3-base |
| Optimizer | AdamW |
| Epochs | 5 |
| Batch size | 64 |
| LR (Backbone) | $1 \times 10^{-5}$ |
| LR (GCN/Head) | $2 \times 10^{-5}$ |
| Warmup ratio | 0.1 |
| Seed | 42 |

Table 3: Hyperparameter Configuration for System Stabilization and Cross-Lingual Training.

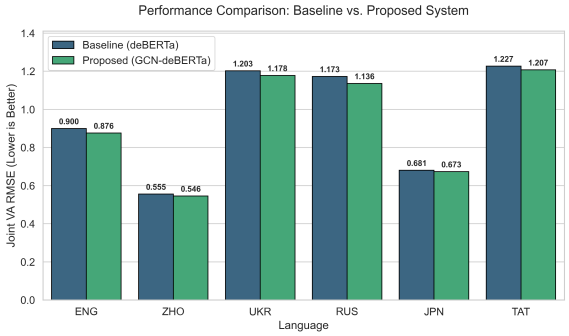


Figure 1: Performance comparison between the best performing baseline and proposed model. Evaluated on RMSE score.

### 3.1 Evaluation metrics

We evaluated our models using the root mean square error (RMSE) metric, as suggested by the SemEval organizers. We present an overall RMSE metric as well as variance and arousal specific scores. Additionally, we evaluated the models on the diversity ratio (Div), separately for valence and arousal. Div is calculated as a ratio of the standard deviation of the model's predictions to the standard deviation of the golden standard labels. Diversity close to 100% ensures that the model captures full range of sentiment intensity, while lower values indicate that it is being conservative and predicts values closer to the global average and fails to capture the full spectrum of emotions.

## 4 Results

Table 5 presents the results of each model for all languages, across all loss functions. Figure 1 visualizes the comparison between the baseline and the model with the best performance. The presented results differ from those submitted for the official task evaluation due to minor hyperparameter refinements, including extended training. We include the results presented in the official leader board in Table 4.

### 4.1 Transformer only vs GCN-enhanced model

The GCN-deBERTa model outperforms the baseline across all target languages. While the baseline achieves competitive results, the GCN enhancement enriches text representation by incorporating grammatical structure, significantly improving the model's ability to predict valence and arousal.

The GCN enhancement proves to be most effective in complex languages such as Russian (RUS) and Ukrainian (UKR), achieving scores of 1.136 and 1.178, respectively, using the Hybrid loss. Notably, these languages also had the smallest dataset sizes, making the results particularly noteworthy.

### 4.2 Mean-Regression Trap

The "mean-regression trap" is evidenced by the diversity (Div) metric. While models trained with MSE loss exhibit often competitive or even better RMSE results, they also show a significant variance collapse, particularly in the arousal dimension, where it can even drop below 50% (for example: 41.6% for TAT and 43.6% for RUS).

### 4.3 Density-Based Evaluation

The results for the Chinese language are particularly intriguing when evaluating the model solely based on the RMSE metric. Figure 2 illustrates the density specific to Chinese, representing the probability of the model selecting a particular value within the range [1,9]. This visualization offers insights into how realistic the model's worldview is for the Chinese language.

We observe that, while the model trained on MSE loss achieved the best result based on the RMSE metric, its density is significantly misaligned with the ground truth. In contrast, the Hybrid loss almost perfectly aligns with the curve presenting golden standard, especially for arousal.

This visually demonstrates that, although MSE achieves closer predictions to individual points, it fails to accurately recreate the true distribution of human emotions.

### 4.4 Valence vs. Arousal

We observe that arousal is consistently more difficult to predict (exhibiting lower diversity) than valence across all languages. However, the hybrid loss helped narrow this gap by incorporating information about variance that MSE overlooked.

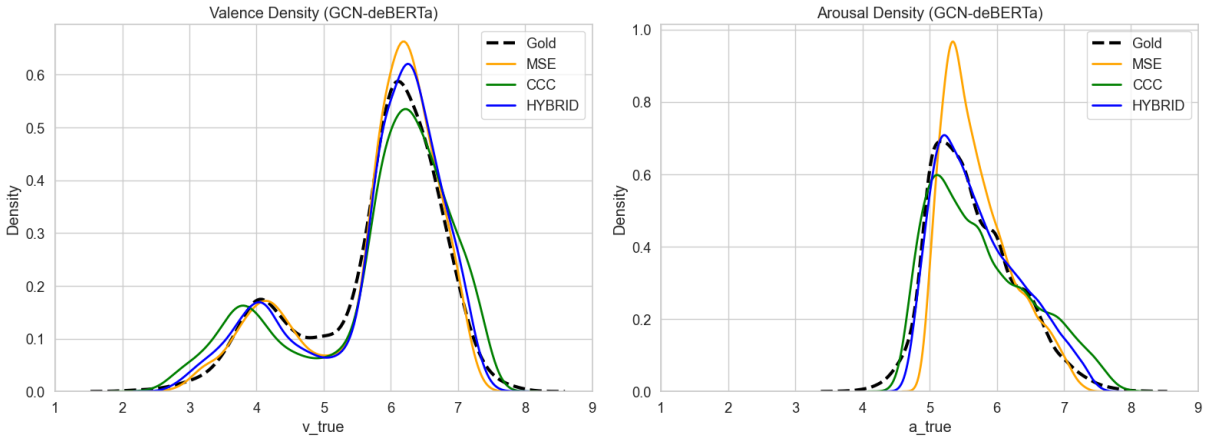


Figure 2: Chinese density of valence and arousal for GCN-deBERTa model trained on selected loss functions compared to the golden standard density.

## 5 Conclusion

Our evaluation revealed that MSE-optimized models tend to exhibit lower diversity, leading to predictions that are safer and closer to global mean values. In contrast, models with higher diversity excel at recognizing nuances in sentiment. This capability is particularly advantageous for real-life applications, where distinguishing between similar emotions, such as "slightly annoyed" and "irritated," is beneficial. In such cases, an MSE-optimized model might simply label both as 'unhappy.' A good balance can be achieved by employing hybrid loss functions, which leverage the strengths of both loss types. This approach has also been shown to deliver the best RMSE results for half of the target languages while simultaneously enhancing diversity.

Evaluating models solely based on RMSE scores may fail to capture the true distribution of emotions. For instance, in the case of the Chinese results, the RMSE metric favored the model trained on MSE loss, which tended to over-predict average values. While the hybrid loss model may have performed slightly worse on individual data points, it more accurately represented the true density of valence and arousal. The visual evidence presented in Figure 2 demonstrates that RMSE can be misleading when the goal is to capture the true variance of human emotions.

| Language | Domain | RMSE |
|---|---|---|
| ENG | laptop | 1.7587 |
| | restaurant | 1.7011 |
| JPN | finance | 1.4505 |
| | hotel | 1.2827 |
| RUS | restaurant | 2.2749 |
| TAT | restaurant | 2.3347 |
| UKR | restaurant | 2.2589 |
| ZHO | finance | 0.8179 |
| | laptop | 1.1343 |
| | restaurant | 1.2405 |

Table 4: Results of the previous version of the model (submitted for evaluation) from the official SemEval-2026 leaderboard (Yu et al., 2026)

## References

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Yan Cathy Hua, Paul Denny, Jörg Wicker, and Katerina Taskova. 2024. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57(11):296.

I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

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.

| Lang | Model | Loss | RMSE ↓ | V-RMSE | A-RMSE | V-Div ↑ | A-Div ↑ |
|---|---|---|---|---|---|---|---|
| ENG | deBERTa | MSE | 0.900 | 0.969 | 0.824 | 103.8% | 81.6% |
| | deBERTa | CCC | 0.990 | 1.058 | 0.917 | 113.6% | 104.7% |
| | deBERTa | HYBRID | 0.913 | 0.975 | 0.846 | 104.8% | 85.2% |
| | GCNdeBERTa v2 | MSE | **0.876** | 0.932 | 0.816 | 98.2% | 76.4% |
| | GCNdeBERTa v2 | CCC | 0.962 | 1.024 | 0.895 | 108.8% | 101.0% |
| | GCNdeBERTa v2 | HYBRID | 0.897 | 0.955 | 0.836 | 101.5% | 86.0% |
| JPN | deBERTa | MSE | 0.681 | 0.808 | 0.523 | 94.3% | 84.5% |
| | deBERTa | CCC | 0.781 | 0.920 | 0.612 | 114.1% | 119.4% |
| | deBERTa | HYBRID | 0.701 | 0.823 | 0.552 | 97.4% | 97.5% |
| | GCNdeBERTa v2 | MSE | **0.673** | 0.796 | 0.522 | 88.6% | 81.7% |
| | GCNdeBERTa v2 | CCC | 0.760 | 0.899 | 0.590 | 113.2% | 114.8% |
| | GCNdeBERTa v2 | HYBRID | 0.679 | 0.798 | 0.534 | 93.0% | 93.7% |
| RUS | deBERTa | MSE | 1.173 | 1.347 | 0.968 | 79.8% | 43.6% |
| | deBERTa | CCC | 1.354 | 1.571 | 1.096 | 96.2% | 88.2% |
| | deBERTa | HYBRID | 1.189 | 1.333 | 1.024 | 81.9% | 58.0% |
| | GCNdeBERTa v2 | MSE | 1.166 | 1.348 | 0.949 | 78.4% | 46.7% |
| | GCNdeBERTa v2 | CCC | 1.267 | 1.443 | 1.062 | 91.2% | 83.3% |
| | GCNdeBERTa v2 | HYBRID | **1.136** | 1.280 | 0.970 | 85.9% | 64.1% |
| TAT | deBERTa | MSE | 1.227 | 1.432 | 0.980 | 72.5% | 41.6% |
| | deBERTa | CCC | 1.387 | 1.596 | 1.139 | 96.7% | 97.2% |
| | deBERTa | HYBRID | 1.248 | 1.444 | 1.016 | 76.8% | 63.7% |
| | deBERTa | CCC | 1.387 | 1.596 | 1.139 | 96.7% | 97.2% |
| | GCNdeBERTa v2 | MSE | 1.220 | 1.429 | 0.967 | 68.5% | 49.9% |
| | GCNdeBERTa v2 | CCC | 1.319 | 1.533 | 1.063 | 90.7% | 87.5% |
| | GCNdeBERTa v2 | HYBRID | **1.207** | 1.401 | 0.976 | 75.5% | 62.1% |
| UKR | deBERTa | MSE | 1.210 | 1.395 | 0.993 | 77.7% | 44.2% |
| | deBERTa | CCC | 1.362 | 1.566 | 1.121 | 91.6% | 89.9% |
| | deBERTa | HYBRID | 1.203 | 1.361 | 1.021 | 78.2% | 62.9% |
| | GCNdeBERTa v2 | MSE | 1.197 | 1.390 | 0.966 | 70.9% | 45.7% |
| | GCNdeBERTa v2 | CCC | 1.242 | 1.432 | 1.017 | 85.8% | 83.1% |
| | GCNdeBERTa v2 | HYBRID | **1.178** | 1.361 | 0.960 | 75.9% | 59.4% |
| ZHO | deBERTa | MSE | 0.555 | 0.602 | 0.505 | 102.2% | 84.0% |
| | deBERTa | CCC | 0.641 | 0.680 | 0.599 | 117.2% | 120.2% |
| | deBERTa | HYBRID | 0.587 | 0.625 | 0.546 | 107.2% | 102.7% |
| | GCNdeBERTa v2 | MSE | **0.546** | 0.590 | 0.497 | 98.1% | 80.7% |
| | GCNdeBERTa v2 | CCC | 0.630 | 0.666 | 0.593 | 115.2% | 118.5% |
| | GCNdeBERTa v2 | HYBRID | 0.571 | 0.608 | 0.531 | 103.9% | 99.6% |

Table 5: Final experimental results.


Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

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

Juan Terven, Diana-Margarita Cordova-Esparza, Julio-Alejandro Romero-González, Alfonso Ramírez-Pedraza, and Edgar Chávez Urbiola. 2025. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58.

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.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4568–4578.