Team BOBW (Best Of Both Worlds) at SemEval-2026 Task 3: Modular Cross-Attention Encoders for Dimensional Aspect-Based Sentiment Analysis

Michal Rynowiecki
IT University of Copenhagen
miry@itu.dk

Rob van der Goot
IT University of Copenhagen
robv@itu.dk

Abstract

This paper presents our system for SemEval-2026 Task 3, which identifies four-part opinion tuples in product reviews. We used a sequence of pairs of BERT encoder models connected by cross-attention layers. The cross-attention mechanism provided marginally better results than a self-attention equivalent, failing to showcase a significant improvement. Error propagation through the pipeline hurt the correctness of the outputs, with certain stages collapsing the scores. The pipeline architecture's performance was largely independent of model size, suggesting that small modular encoders for downstream tasks are an efficient alternative to large decoder models. Our best model got a cF1 score of 0.53 on restaurant data and 0.26 on laptop data.

1 Introduction

Large decoder-based language models have become the standard tool for many NLP tasks (Vaswani et al., 2017; Radford et al., 2018). However, they have practical drawbacks. They are large and expensive to run and typically proprietary. For downstream tasks with structured outputs, encoder models (e.g. Devlin et al., 2018) offer a practical alternative.

SemEval-2026 Task 3 (DiMBaSA) (Yu et al., 2026) is a structured sentiment analysis task that requires identifying sentiment quadruplets in short reviews. Each quadruplet consists of: (1) an aspect term, (2) an opinion term, (3) a sentiment category from a fixed label set, and (4) continuous valence and arousal scores on a scale from 1 to 9. Valence captures positive versus negative sentiment, while arousal captures the intensity of the sentiment expressed. The following example from the laptop dataset:

> This **unit** is **pretty** and **stylish,** so my high school daughter was attracted to it for that reason.

Contains two quadruplets:

- (unit, pretty, LAPTOP#DESIGN_FEATURES, 7.12#7.12)
- (unit, stylish, LAPTOP#DESIGN_FEATURES, 7.12#7.12).

This task is challenging because it involves four interconnected sub-tasks. Solving each sub-task independently loses the relationships between them. For example, the sentiment category of an opinion depends on both the aspect it refers to and the specific opinion word used. Traditional graph-based neural approaches (e.g. Scarselli et al., 2009) can capture these relationships, but they become complicated when a single word participates in multiple quadruplets with different categories.

We instead propose a modular pipeline of encoder models, one per sub-task, connected by cross-attention. Each model in the pipeline can attend to the output of the previous model, passing useful contextual information forward. This is similar to how encoder-decoder models transfer information through cross-attention, but applied entirely within an encoder-only framework. Such an architecture could potentially offer an efficient and accessible approach to semantic analysis, combining the precision of task-specific models with the broader contextual awareness of a larger shared model, providing the best of both worlds.

This leads to our main question: *Can small, separate encoder models be linked with data pipelines and cross-attention to successfully share information across tasks?*

To answer this, our contributions are as follows:

(1) We built a step-by-step pipeline of small models linked by cross-attention to find sentiment quadruplets

(2) We tested if cross-attention actually helps by comparing it to a baseline system using only self-attention

(3) We analyzed how errors get passed down the steps of a pipeline and showed that small models can perform just as well as larger ones in this specific setup.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1385–1390

2 Background

2.1 Structured Sentiment Analysis

Sentiment analysis has evolved from coarse-grained document-level classification to fine-grained structured predictions. Early work focused on predicting overall sentiment polarity (Pang et al., 2002), followed by aspect-based sentiment analysis (ABSA), which identifies sentiment toward specific entities or attributes (Hu and Liu, 2004; Pontiki et al., 2014). ABSA tasks typically extract aspect terms, opinion terms, and the sentiment polarity expressed toward each aspect. This progression toward more nuanced representations of sentiment aligns with psychological models such as the one presented in (Russell, 1980), which conceptualizes emotions along continuous dimensions rather than discrete categories.

More recent benchmarks have introduced additional complexity. The SemEval-2022 Task 10 (Structured Sentiment Analysis) required extracting sentiment graphs where opinion tuples could span multiple tokens and overlap within sentences (Barnes et al., 2022). This task highlighted the need for models that can capture relationships between multiple overlapping extractions, a challenge that carries over to the DiMBaSA task.

2.2 Multi-Task Learning and Information Sharing

Multi-task learning (Caruana, 1997) provides a framework for sharing representations across related tasks. We will use multi-task learning to be able to share information across the tasks, specifically, we use **Cross-Attention Mechanisms:** Cross-attention was introduced in the Transformer architecture (Vaswani et al., 2017) to allow a decoder to attend to encoder representations. This mechanism has been extended to multi-encoder setups, such as cross-encoder models for sentence pair tasks (Reimers and Gurevych, 2019). Our work adapts cross-attention to connect multiple encoder models processing the same input, enabling information flow between sub-tasks while maintaining modular task-specific architectures.

2.3 Dataset

The dataset contains short reviews in multiple domains (Lee et al., 2026; Becker et al., 2026). We focus on the English datasets for the laptop and restaurant subdomains.

3 System overview

Our system follows a four-stage pipeline. Each stage corresponds to one subtask and is handled by one or two encoder models. Where two models are used, they share input representations through cross-attention layers. The full pipeline is: (1) aspect and opinion extraction, (2) binary pair validation, (3) category classification, and (4) valence and arousal regression. We use BERT as the base encoder throughout. For initial experiments, we use BERT-medium (Turc et al., 2019) (Bhargava et al., 2021). For final runs, we use BERT-large-cased. We also tested Snowflake Arctic Embed (Merrick et al., 2024) as an alternative encoder, due its focus on information retrieval during training, which we theorized could have been an advantage for aspect and opinion mining. This model was only used for the laptop dataset due to its initial bad performance for the task.

3.1 Cross-Attention Design

The core idea of our system is using cross-attention to link pairs of encoders within the pipeline. In standard encoder-decoder models, cross-attention lets the decoder look at the full encoder output at each step. We adapt this so two encoders working on the same task can attend to each other instead. Our goal was to remove the limitations of using either a single model for all tasks, or one model per task. By connecting encoders with cross-attention and placing them in a pipeline, we aim to get the best of both worlds: the simplicity of a single model and the precision of task-specific ones.

For each encoder pair (Encoder A and Encoder B), we add cross-attention layers where Encoder B uses the hidden states from Encoder A as key-value pairs, while its own hidden states serve as queries. This allows Encoder B to selectively attend to relevant parts of Encoder A's output. The mechanism is added at specific, pre-selected intermediate layers.

To be able to evaluate the effect of cross-attention, we compare against a baseline that uses the exact same setting, but without cross-attention. In this case, all the models use self-attention in place of cross-attention. For both scenarios, the loss is defined as the sum of the loss for both models.

3.2 Aspect and Opinion

The first stage identifies aspect and opinion spans in the input sentence. We treat this as a sequence

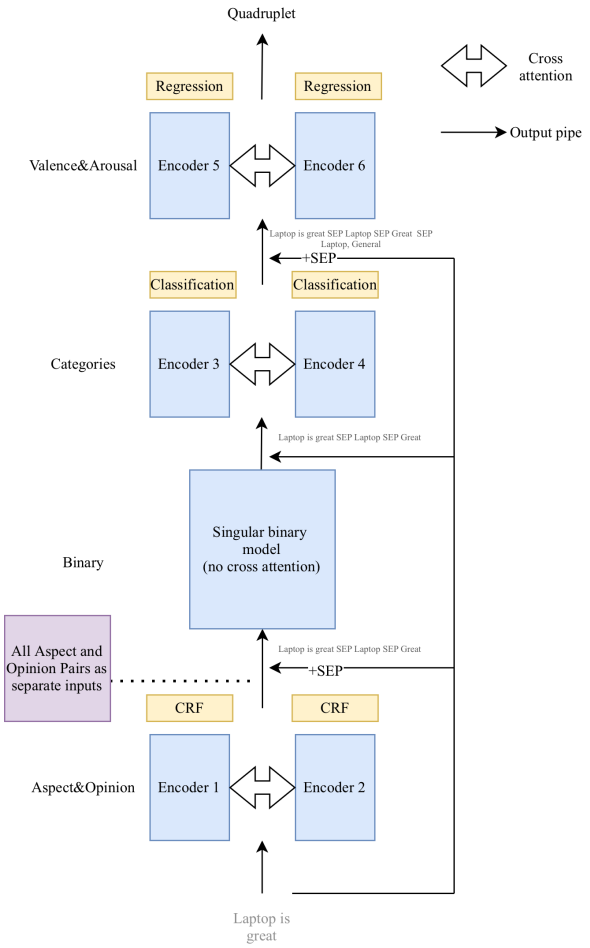

Figure 1: System overview

labeling task using BIO tagging. A conditional random field (CRF) (Lafferty et al., 2001) layer is added on top of the encoder to enforce label consistency across tokens. Two separate encoder models handle aspects and opinions respectively. Both models receive the same input sentence and are connected by cross-attention, so the aspect model can attend to opinion representations and vice versa. The output is a set of labeled token spans.

3.3 Combining Aspect and Opinion

The second stage determines which combinations of extracted aspects and opinions form valid pairs. For a sentence with X aspects and Y opinions, this generates up to X x Y candidates. The model receives each candidate pair as input and outputs a binary label (valid or not valid). For this stage, we use a single larger encoder without cross-attention, as the task is simpler and would not benefit as much from the dual-encoder design. The input includes the original sentence concatenated with the aspect and opinion spans, separated by a SEP token.

3.4 Category

The third stage assigns a sentiment category to each valid aspect-opinion pair. The category label set differs between domains: the restaurant domain has broader categories while the laptop domain has finer subcategories. We split this task across two encoders connected by cross-attention, each with a classification head equal in size to the number of categories. The input to each encoder includes the original sentence, the aspect, and the opinion, separated by SEP tokens.

3.5 Valence and Arousal

The fourth stage predicts continuous valence and arousal scores for each quadruplet. We again use two encoders with cross-attention, this time with a regression head on top. The input includes the original sentence together with the aspect, opinion, and predicted category, all separated by SEP tokens.

4 Experimental setup

We ran two sets of experiments. Small-model experiments used BERT-medium[1] and were ran on an M1 MacBook Pro with 16GB of RAM. Large-model experiments used BERT-large-cased and were ran on an RTX 6000 GPU with 48GB of VRAM. We also tested a Snowflake Arctic encoder as an alternative to BERT.

4.1 Training Procedure

All models were fine-tuned only on the task training set, without additional pretraining on related corpora. We trained for a fixed number of epochs. For the cross-attention experiments, we added cross-attention layers after the layers 3 and 7 for BERT-medium and arctic-embed, and layers 8, 14 and 23 for BERT-large-cased. For the self-attention baseline, we instead added equivalent self-attention layers to keep parameter counts comparable.

4.2 Hyperparameters

We selected a **learning rate** of 5e-5 for the BERT-medium model, and a learning rate of 1e-5 for the BERT-large-cased model by examining the following learning rates: 0.05, 0.01, 0.001, 0.0001, 5e-5, 1e5. The placement and number of **cross-attention layers** was selected through an empirical search due to the quadratic number of possible combinations, especially when tested against multiple models. All pairs of models were trained separately for

[1]https://huggingface.co/prajjwal1/bert-medium

3 **epochs**, i.e. Encoder 1 and 2 for three epochs, Encoder 3 and 4 for three epochs, Encoder 5 and 6 for three epochs, and a single epoch for the binary classification model. In our setup, cross-attention did not have a noticeable effect on training time, as can be seen on Figure 2.

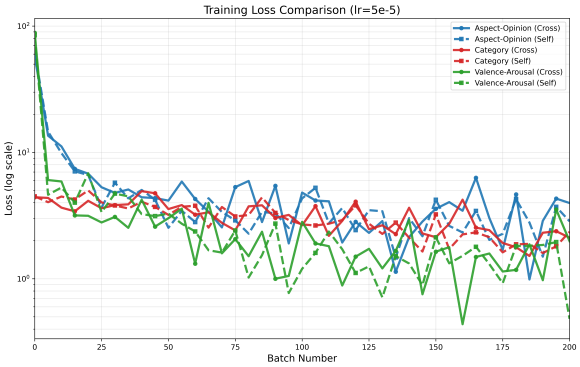

Figure 2: Loss over batches compared between self and cross attention for each sub-task over 1 epoch

4.3 Evaluation

We used a 0.75/0.25 split of the training set to obtain a dev set before the publication of the test gold labels, and the official dev set as a test set. The primary evaluation metric of the complete system is the combined quadruplet F1 score (cF1), which requires all four parts of a quadruplet to be correct for a prediction to count as a true positive. We also report component-level precision (cPrec) and recall (cRec). We also calculated precision and recall for each separate component of the pipeline, as well as per-class precision and recall for the classification task. We used MSE for the regression evaluation.

5 Results

5.1 Analysis

The results present a mixed picture, with no single model demonstrating a decisive advantage in terms of F1 score. Performance differences are more apparent when examining precision and recall independently. The larger BERT model consistently identifies the highest number of true positives across both datasets.

The results on the Laptop dataset are poor. Here, the self-attention model achieves a marginally higher F1 score than its cross-attention counterpart, despite the overall low performance (F1 scores around 0.25). This outcome is somewhat unexpected, as cross-attention was anticipated to provide a benefit.

A qualitative examination of the errors reveals that most mistakes are minor. When a prediction is incorrect, the extracted quadruplet is often almost correct, with only a single element being wrong. This highlights the issues of the pipeline architecture and error propagation.

For the self-attention model, errors more frequently involve aspects or opinions that span multiple words or contain negations with contractions. These specific errors however do not appear to heavily penalize the final F1 score compared to the other models.

One possible explanation for these results is that the direct information exchange provided by cross-attention layers may not be strictly necessary for the sub-tasks for which it was introduced. The sequential pipeline itself already connects the different stages, providing the later stages with the outputs of the earlier ones. This suggests that the representation passed between stages may be a more critical factor than the architectural mechanism used within a stage. In other words, the input to the model might be more effectively transformed into a performance gain by the pipeline itself than by interconnected weights.

Another consideration is that the cross-attention weights in these models are not pre-trained. It is possible that their potential would be more fully realized if the model were first trained with a language modeling objective after the cross-attention mechanism was added. However, this is not straightforward to implement, as it raises questions about the amount and domain of data required for effective pre-training.

Simply incorporating cross-attention does not guarantee a substantial increase in performance. While in both datasets either the small or large cross-attention model slightly outperforms the self-attention baseline, the gains are marginal.

5.2 Ablation experiments

Increasing model size improves performance for individual sub-tasks. For example, replacing BERT-tiny with DeBERTa-v3 Large for aspect and opinion detection increased precision from 0.65 to 0.84. However, using larger models throughout the entire pipeline does not significantly improve overall results, likely due to error propagation between stages.

The binary classification stage is the main bottleneck and is unreliable because even when we change the model, hyper-parameters, or adjust the

Restaurant						
Model	**cF1**	**cPrec**	**cRec**	**TP**	**FP**	**FN**
Bert-Medium-CA	0.53	0.62	0.46	1065	520	1064
Bert-Medium-SA	0.51	0.53	0.5	1081	806	916
Bert-Large-Cased	0.52	0.54	0.51	1084	810	928
Baseline (Kimi K2)	0.37	-	-	-	-	-
Baseline (Qwen-3 14B)	0.28	-	-	-	-	-

Laptop						
Model	**cF1**	**cPrec**	**cRec**	**TP**	**FP**	**FN**
Bert-Medium-CA	0.23	0.28	0.2	425	980	1550
Bert-Medium-SA	0.25	0.26	0.23	456	1235	1479
Snowflake	0.15	0.16	0.15	298	1529	1618
Bert-Large-Cased	0.26	0.29	0.24	464	1080	1471
Baseline (Kimi K2)	0.28	-	-	-	-	-
Baseline (Qwen-3 14B)	0.15	-	-	-	-	-

Table 1: Performance on Restaurant and Laptop test sets for Subtask 3 (DimASQP). cF1, cPrec, and cRec are the combined quadruplet scores (%). TP, FP, and FN refer to total true positives, false positives, and false negatives respectively. The results for Bert-Medium-CA are the results submitted for the official task ranking.

ratio of negative to positive examples, the rate of false positives and false negatives remains similar. This may occur because the model examines each aspect-opinion pair individually. We theorize that a method capable of classifying all pairs simultaneously could improve accuracy.

Classification performance depends heavily on the domain. The first part of the category is generally easy to detect, reaching approximately 0.80 accuracy for both laptops and restaurants. However, the second category component, particularly in the laptop dataset, causes the pipeline to fail, with accuracy dropping to 0.50. Adjusting loss weights or class weights based on frequency did not resolve this issue.

In contrast, predicting Valence-Arousal (VA) scores is straightforward. Qualitative analysis shows these scores are accurate and adapt correctly to the tone of the review, even when sarcasm is present.

6 Conclusion

Overall, the results suggest that simply connecting the different stages of the pipeline with cross-attention is not sufficient to guarantee better performance in dimensional semantic analysis. While the approach can lead to higher accuracy in some cases, the overall F1 scores remain similar across models and do not appear to be strongly tied to model size. It is worth noting, however, that the smaller BERT-Medium model performed reasonably well, which is somewhat promising and suggests that cross-attention or modular encoders may still have some utility even in resource-constrained settings.

As mentioned earlier, it may be worthwhile to explore whether pre-training with a language modeling objective before fine-tuning on the downstream task would lead to more substantial gains. Another direction for future work would be to modify the pipeline itself, for example by predicting all components of the quadruplet in a joint manner with full connectivity between stages - a setup where no additional tokens are passed between stages.

Acknowledgments

Computing resources were provided by ITU.

References

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Idris Abdulmumin, Lung-Hao Lee, Nelson Odhiambo, Lilian Wanzare, Wen-Ni Liu, Tzu-Mi Lin, Zhe-Yu Xu, Ying-Lung Lin, Jin Wang, Maryam Ibrahim Mukhtar, Bela Gipp, and Saif M. Mohammad. 2026. Dimstance: Multilingual datasets for dimensional stance analysis. *Preprint*, arXiv:2601.21483.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers.

2021. Generalization in nli: Ways (not) to go beyond simple heuristics. *Preprint*, arXiv:2110.01518.

Rich Caruana. 1997. Multitask learning. *Mach. Learn.*, 28(1):41–75.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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.

Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed: Scalable, efficient, and accurate text embedding models. *Preprint*, arXiv:2405.05374.

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.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

James A. Russell. 1980. A circumplex model of affect.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *Trans. Neur. Netw.*, 20(1):61–80.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.