

Pixel Phantoms at SemEval-2026 Task 3: Language-Specific Transformer Regression for Dimensional Aspect-Based Sentiment Analysis

Jithu Morrison S

Sri Sivasubramaniya Nadar College
of Engineering
jithumorrison2210564@ssn.edu.in

Abisha Rose S

Loyola-ICAM College of Engineering
and Technology
abisharoses20@gmail.com

Abstract

Aspect-Based Sentiment Analysis (ABSA) has traditionally relied on discrete polarity labels that fail to capture the continuous nature of human emotion. SemEval-2026 Task 3, Dimensional Aspect-Based Sentiment Analysis (DimABSA), addresses this limitation by requiring continuous Valence and Arousal prediction on a 1–9 scale for specific aspect terms across 15 language–domain combinations. The Pixel Phantoms system adopts a language-aware strategy, selecting dedicated pre-trained transformer models for each language and falling back to `xlm-roberta-base` for lower-resource cases such as Tatar and Ukrainian. All models share a common regression architecture with dual pooling, aspect-prompted input formatting, and a composite MSE + MAE loss. We participated in both Track A and Track B, with our strongest result in Japanese Hotel (rank 13/21, RMSE 0.7297) and competitive performance in Chinese restaurant (RMSE 0.9823 vs. Baseline Kimi-K2 Thinking 1.8959). Overall, language-specific encoders provide clear gains over generic multilingual baselines for dimensional sentiment regression.

1 Introduction

Sentiment analysis systems have long represented opinions as discrete categories. While practical, this approach is a lossy compression of the richness of human emotional expression. Two restaurant reviews stating “the food was acceptable” and “the food was absolutely incredible” both carry positive sentiment, yet differ profoundly in emotional intensity (arousal) and degree of positivity (valence). Dimensional sentiment analysis, grounded in the circumplex model of affect, captures these distinctions through two continuous axes: Valence and Arousal.

SemEval-2026 Task 3 (**Dimensional Aspect-Based Sentiment Analysis (DimABSA)**) is the first large-scale shared task to combine aspect-level

targeting with dimensional sentiment regression across diverse languages and domains (Yu et al., 2026). Subtask 1 (VA Regression) requires systems to predict Valence–Arousal scores for given aspect terms, while Subtask 2 (DimASTE) additionally requires extracting aspect and opinion spans. This setting is challenging because systems must resolve both which aspect is being evaluated and how strongly that aspect is expressed on a continuous emotional scale. We focused on **Subtask 1** and submitted predictions for all language–domain combinations in both Track A and Track B.

Our approach is centered on a **language-specific model selection strategy**: rather than applying a single multilingual encoder to all languages, we selected dedicated pre-trained models best suited to each language. Shared across all language models is a common regression head architecture using dual pooling (CLS + mean) and an aspect-prompted input format.

Key contributions of this work:

- A language-specific model selection strategy covering 7 languages and 15 language–domain combinations across two tracks
- An empirical comparison of language-specific versus generic multilingual encoders for dimensional sentiment regression
- Analysis of low-resource and domain-novel failure modes
- Documentation of four model variants developed during the competition

2 Background and Related Work

2.1 Aspect-Based Sentiment Analysis

ABSA has been a central task in opinion mining since the SemEval series introduced shared tasks on restaurant and laptop reviews (Pontiki et al.,

2014, 2016). These tasks established the convention of discrete polarity classification for aspect terms and strongly influenced later benchmark design. DimABSA extends this line by replacing discrete polarity with continuous Valence and Arousal scores, requiring regression rather than classification and enabling finer distinctions between mild and intense sentiment.

2.2 Dimensional Models of Affect

The psychological basis for dimensional sentiment analysis derives from the circumplex model of affect (Russell, 1980), which represents emotion in a two-dimensional space of Valence and Arousal. This model has informed affective computing and sentiment lexicon development (Mohammad, 2018; Warriner et al., 2013). However, its integration into multilingual ABSA has been limited, even though aspect-level sentiment is precisely the kind of task that benefits from capturing both polarity and intensity.

2.3 Transformer Models for Multilingual NLP

Pre-trained transformers have set state-of-the-art results across sentiment tasks. BERT (Devlin et al., 2019) and its variants, such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and XLM-RoBERTa (Conneau et al., 2020), provide strong contextual representations. However, language-specific models often outperform multilingual models in their target language, motivating our model selection strategy (Delobelle et al., 2020).

2.4 DeBERTa and Disentangled Attention

DeBERTa (He et al., 2021) improves upon BERT and RoBERTa by encoding token content and relative position separately using a disentangled attention mechanism. This is particularly beneficial for tasks requiring fine-grained span-level understanding, such as identifying sentiment toward a specific aspect term within a longer text. We use DeBERTa variants (deberta-v3-small and deberta-v3-base) as the primary backbone for English and German, and as our early experimental model for other languages.

2.5 Aspect-Prompted Encoding

Rather than encoding text and aspect as a two-sequence pair (standard in BERT-style sentence-pair tasks), aspect prompting prepends a labeled aspect marker to the input. This has been shown to improve aspect sensitivity in generative and

encoder-decoder models (Gao et al., 2022). In our setting, the prompt also serves as a lightweight way to inject task structure without modifying the encoder architecture itself. We extend this strategy to encoder-only regression models and adapt it for Japanese using a native-language aspect label.

3 System Overview

The Pixel Phantoms system is built around a shared regression architecture applied with **language-specific encoder backbones**. All model variants share the same output head design and training strategy; they differ in encoder selection and minor architectural refinements developed across four iterations.

3.1 Language-Specific Model Selection

A central design decision was to select the best available pre-trained model for each language, rather than using a one-size-fits-all multilingual encoder. Table 1 summarizes our assignments and rationale.

Models were selected based on availability of strong monolingual pretraining, prior NLP benchmarks, and tokenizer compatibility with each language.

Japanese tokenization note: Japanese requires morphological tokenization. We installed `fugashi` (MeCab Python binding) and `ipadic` (IPAdic morphological dictionary), which are required by `cl-tohoku/bert-base-japanese-v3`. We also used a Japanese-native aspect prompt: {aspect} [SEP] {text} (“aspect/viewpoint”) instead of the English-format Aspect: {aspect} [SEP] {text} used for all other languages. This was an empirical design choice motivated by better alignment with native syntax rather than established prior work.

3.2 Input Encoding: Aspect Prompting

All models (except Model V1) use an **aspect-prompted input format**:

```
Aspect: {aspect_term} [SEP] {review_text}
# English, Russian, Chinese, Tatar,
# Ukrainian, Swahili, German,
# Nigerian Pidgin, Japanese
```

This format highlights the aspect of interest before the full text and gives the model an explicit focus for prediction. In Model V1, text and aspect were encoded as a native sentence pair `tokenizer(text, aspect)`, which we found less

Language	Domain(s)	Track	Pre-trained Model
English	Laptop, Restaurant, Environment	A, B	microsoft/deberta-v3-base
Japanese	Finance, Hotel	A	cl-tohoku/bert-base-japanese-v3
Russian	Restaurant	A	DeepPavlov/rubert-base-cased
Chinese	Finance, Laptop, Restaurant, Env	A, B	bert-base-chinese
Tatar	Restaurant	A	xlm-roberta-base
Ukrainian	Restaurant	A	xlm-roberta-base
Swahili	Politics	B	Davlan/bert-base-multilingual-swa
German	Politics	B	microsoft/deberta-v3-base
Nigerian Pidgin	Politics	B	xlm-roberta-base

Table 1: Language-specific model assignments.

effective because the aspect did not receive the same prominence in the input sequence.

3.3 Model Architecture

The regression model consists of three components:

Encoder Backbone: Language-specific pre-trained transformer producing hidden states of dimension H (768 for base-size models).

Dual Pooling:

```
cls = last_hidden_state[:, 0]
mean = SUM(hidden_i * mask_i) / SUM(mask_i)
pooled = Dropout(p=0.2)(Concat[cls, mean])
```

Concatenating CLS (global sequence summary) and mean pooling (distributed representation) provides richer signal than either alone.

Regression Head:

```
hidden = GELU(Linear(2H->512)(pooled))
va = Sigmoid(Linear(512->2)(hidden))
      * scale + shift
```

Where scale and shift are **learnable scalar parameters** initialized to 8.0 and 1.0, mapping sigmoid output from (0,1) to approximately (1,9). Making these learnable allows the model to adaptively calibrate the output range per task/domain during training.

Sigmoid ensures bounded outputs, while GELU provides a smoother nonlinear transformation than ReLU in the regression head.

3.4 Training Objective

Primary loss used in all final submissions:

$$L = 0.7 * \text{MSE}(\hat{y}, y) + 0.3 * \text{MAE}(\hat{y}, y)$$

MSE penalizes large errors quadratically, while MAE provides robustness to outlier annotations.

This combination gave more stable training behavior than MSE alone in our later experiments. We also explored Huber loss in Model V3.

3.5 Model Evolution: Four Variants

We developed four architectural variants during the competition period.

Model V1: DeBERTa Sentence-Pair (Initial Baseline)

Component	Configuration
Encoder	microsoft/deberta-v3-small
Input	tokenizer(text, aspect)
Pooling	CLS token only
Head	Linear($H \rightarrow 2$)
Output scaling	$\text{sigmoid}(\text{logits}) * 8 + 1$
Loss	MSE only
Epochs	5
LR	$2e-5$
Batch size	16
Max length	128
Dropout	None
Early stopping	No

Table 2: Model V1 configuration.

This initial Track A implementation targeted `jpn_finance`. DeBERTa’s `token_type_ids` were removed because the model does not use them. It used no aspect prompting, dropout, or early stopping and mainly served as a proof of concept.

Model V2: DeBERTa-Base with Aspect Prompting and Layer Freezing

Key improvements were upgrading to `deberta-v3-base`, introducing aspect prompting, freezing the bottom two encoder layers, adding

Component	Configuration
Encoder	microsoft/ deberta-v3-base
Input	"aspect:{asp}[SEP]{text}"
Pooling	CLS token only
Head	Linear(H -> 2)
Output scaling	$\text{sigmoid}(\text{logits}) * 8 + 1$
Loss	MSE only
Epochs	Up to 8
LR	2e-5, decayed to 1e-5 (epoch 3)
Batch size	8
Max length	128
Dropout	None
Frozen layers	Bottom 2 DeBERTa layers
Early stopping	patience=1

Table 3: Model V2 configuration.

LR decay after epoch 3, and early stopping. The aggressive patience of 1 may have caused underfitting.

Model V3: DeBERTa with Dual Pooling and Huber Loss

Component	Configuration
Encoder	microsoft/deberta-v3-small
Input	"aspect:{asp}[SEP]{text}"
Pooling	CLS + Mean pooling -> concat
Head	Linear(2H -> 256) + GELU
Output scaling	$\text{sigmoid}(\text{logits}) * 8 + 1$
Loss	HuberLoss(delta=1.0)
Dropout	0.3
Epochs	Up to 10
LR	2e-5
Batch size	8
Max length	160
Frozen layers	Bottom 2 encoder layers
Weight decay	0.01
Mixed precision	Yes (torch.cuda.amp)
Early stopping	patience=2

Table 4: Model V3 configuration.

Key improvements were **dual pooling** (CLS + mean), a deeper MLP regression head, Huber loss, mixed-precision training, and dropout of 0.3.

Component	Configuration
Encoder	microsoft/ deberta-v3-base
Input	"Aspect:{asp} [SEP]{text}"
Pooling	CLS + Mean pooling -> concat
Head	Linear(2H -> 512) + GELU
Output scaling	$\text{sigmoid}(\text{logits}) * \text{scale}$
Loss	$0.7 * \text{MSE} + 0.3 * \text{MAE}$
Dropout	0.2
Epochs	Up to 12
LR	1e-5
Batch size	4
Max length	256
Weight decay	0.01
Mixed precision	Yes (torch.cuda.amp)
Gradient clipping	norm=1.0
Early stopping	patience=3

Table 5: Model V4 configuration.

Model V4: DeBERTa-Base with Learnable Scaling and MSE+MAE Loss (Final)

This is the final variant, applied to English (restaurant, laptop), German, and as the base architecture for all language-specific adaptations. Relative to V3, it upgrades to deberta-v3-base, expands sequence length to 256, introduces learnable scale/shift parameters, replaces Huber with weighted MSE+MAE, reduces LR to 1e-5, increases patience to 3, expands the hidden layer, and adds gradient clipping. These changes were intended to improve optimization stability while preserving the same overall regression formulation used across languages. We did not use gradient accumulation; small batch sizes were due to GPU constraints.

Language-Specific Model Adaptations (based on V4 architecture):

All language variants share AdamW, weight decay=0.01, FP16 mixed precision, gradient clipping, early stopping, composite MSE+MAE loss, dual CLS+mean pooling, and learnable scale/shift. Keeping these components fixed made it easier to attribute observed differences primarily to encoder choice and language-specific adaptation rather than to unrelated optimization changes.

Language	Encoder	Tokenization	Batch	LR/Epochs
English/German	deberta-v3-base	Standard	4	1e-5 / 12
Japanese	cl-tohoku/bert-base-japanese-v3	MeCab+IPAdic	8	2e-5 / 10
Russian	DeepPavlov/rubert-base-cased	SentencePiece	8	2e-5 / 10
Chinese	bert-base-chinese	WordPiece (char)	8	2e-5 / 10
Tatar	xlm-roberta-base	SentencePiece	8	2e-5 / 10
Ukrainian	xlm-roberta-base	SentencePiece	8	2e-5 / 10
Swahili	Davlan/bert-base-multilingual	WordPiece	8	2e-5 / 10
Nigerian Pidgin	xlm-roberta-base	SentencePiece	8	2e-5 / 10

Table 6: Language-specific training settings.

4 Experimental Setup

4.1 Dataset

We used the official DimABSA dataset (Lee et al., 2026; Becker et al., 2026), distributed in JSONL format. Training files provide quadruplet annotations {Text, Quadruplet: [Aspect, Opinion, VA]}, while test files provide {ID, Text, Aspect: [list]}. VA scores are formatted as "V#A".

Data loading decisions:

- Filtered Aspect == "NULL" entries (implicit aspects not useful for span-level regression)
- Validated VA format and range (0–9) with defensive assertions for Tatar and Nigerian Pidgin, which exhibited occasional format inconsistencies
- For Japanese finance, the training data used Aspect_VA format (subtask-1 specific labels) rather than the general Quadruplet format, requiring a separate loader function

Language	Domain	Rank
English	Laptop	19 / 33
English	Restaurant	24 / 37
Japanese	Finance	16 / 22
Japanese	Hotel	13 / 21
Russian	Restaurant	19 / 23
Tatar	Restaurant	14 / 21
Ukrainian	Restaurant	12 / 20
Chinese	Finance	17 / 21
Chinese	Laptop	14 / 24
Chinese	Restaurant	13 / 24

Table 7: Track A ranks (10 combinations).

4.2 Training Monitoring

Training was monitored using **training RMSE** computed after each epoch over the full training set

Language	Domain	Rank
German	Politics	8 / 12
English	Environment	13 / 15
Nigerian Pidgin	Politics	10 / 12
Swahili	Politics	9 / 12
Chinese	Environment	11 / 14

Table 8: Track B ranks (5 combinations).

(not a held-out validation set). The best checkpoint (lowest training RMSE with minimum improvement threshold of 0.003) was saved for final test inference. Early stopping was applied when no improvement was observed for patience consecutive epochs. Although this monitoring protocol is imperfect, it provided a consistent model-selection rule across all language-domain combinations during the shared task.

Due to limited dataset size and shared-task constraints, we prioritized maximizing training signal and relied on early stopping and regularization to mitigate overfitting. Empirically, models generalized reasonably across the official test sets.

4.3 Infrastructure

All experiments were conducted on **Google Colab with GPU acceleration (CUDA)**. Libraries used:

- transformers (HuggingFace) for model loading and tokenization
- torch, torch.cuda.amp for training and mixed-precision
- scikit-learn for RMSE computation during training
- fugashi, ipadic for Japanese morphological tokenization
- sentencepiece for XLM-RoBERTa and RuBERT subword tokenization

- tqdm for progress tracking

5 Results

The official evaluation metric is **RMSE** (Root Mean Squared Error) between predicted and gold Valence–Arousal score pairs, and lower is better. We report our RMSE and rank and compare against the official baselines.

Track A includes two official baselines provided by the task organizers: **Baseline (Kimi-K2 Thinking)** (a large LLM-based zero-shot baseline) and **Baseline (Qwen-3 14B)** (a zero-shot baseline using Qwen-3 14B). Track B provides **Baseline (Mistral-3 14B)** and **Baseline (mBERT)** as the official reference systems. Beating the baselines is the minimum threshold for a meaningful contribution.

5.1 Key Findings

Beating the baselines (general picture): Across both tracks, our system outperformed the official baselines in the majority of language–domain combinations. All 15 submissions beat the mBERT baseline. This confirms that fine-tuned language-specific encoders can outperform zero-shot LLM and generic multilingual baselines for dimensional VA regression.

Tatar (only case where we lost to a baseline): In the Tatar restaurant track, our RMSE (2.0729, rank 14) was worse than Baseline (Kimi-K2 Thinking) (RMSE 1.9380, rank 9). This is the single track where our fine-tuned approach failed to surpass even the official LLM-based baseline, highlighting the severe limitations of XLM-RoBERTa for an agglutinative Turkic language with minimal multilingual model coverage.

English Environment: Our English environment submission (RMSE 2.0893, rank 13) was outperformed by Baseline (Mistral-3 14B) (RMSE 1.6430, rank 7), likely because this domain involves abstract, policy-oriented sentiment. Such cases suggest that domain-specific reasoning demands may matter as much as language coverage for dimensional prediction.

Nigerian Pidgin: Our submission (RMSE 1.7878, rank 10) narrowly lost to Baseline (Mistral-3 14B) (RMSE 1.7390, rank 9) by 0.049, likely because zero-shot English LLMs transfer well to Pidgin.

Strong beats above baselines: Our strongest margins were in Chinese (restaurant: 0.9823 vs. Kimi-K2 1.8959), Japanese Hotel (0.7297 vs.

Kimi-K2 1.7553), and Chinese finance (0.7259 vs. Kimi-K2 1.9652), confirming that dedicated language-specific models can provide substantial gains in well-resourced Asian languages.

Track B mBERT baseline (always beaten): Our system outperformed the mBERT baseline across all 5 Track B combinations, with the largest margin in Nigerian Pidgin (1.7878 vs. 3.2152), showing that task-specific fine-tuning still often helps despite broad multilingual pretraining coverage, especially when baseline representations remain weak for the target language or domain.

We attribute this improvement to task-specific fine-tuning and explicit aspect prompting. The gains are largest where baseline coverage is weakest. Conversely, narrow margins or losses in low-resource settings suggest that encoder choice alone is not sufficient when pretraining coverage remains sparse.

5.2 Error Analysis

Low-resource language failures: For Tatar, Nigerian Pidgin, and Swahili, the substantial RMSE gap reflects the limitation of applying models with minimal coverage of these languages. Tatar is Turkic and agglutinative; Nigerian Pidgin is an English-based creole with distinct syntax; Swahili has complex noun class morphology. None are well represented in general multilingual pre-training corpora.

We did not explore Turkic transfer from related languages such as Turkish because of time constraints, but this is a promising direction for Tatar.

Domain mismatch: English environment (RMSE 2.0893) and German politics (RMSE 1.5509) show degradation likely from domain-specific vocabulary that differs from the product/hotel domains dominating pre-training. This suggests that domain adaptation remains important even when the base language model is strong.

Aspect representation limitation: All our models represent the aspect as a short string in the prompt. When the same aspect term appears multiple times in a text with differing sentiment contexts, the model cannot disambiguate between occurrences and may blend distinct sentiment signals into one prediction. This issue is especially noticeable in longer reviews with multiple clauses about the same target.

Arousal vs. Valence: Anecdotally, Arousal was harder to predict accurately than Valence. Arousal corresponds to emotional intensity and is more dependent on subtle linguistic cues, such as inten-

Language	Domain	Model	Our RMSE	Baseline (Kimi-K2) RMSE
English	Laptop	deberta-v3-base	1.4190	2.1893
English	Restaurant	deberta-v3-base	1.3656	2.1461
Japanese	Finance	bert-base-japanese-v3	1.0242	1.6396
Japanese	Hotel	bert-base-japanese-v3	0.7297	1.7553
Russian	Restaurant	rubert-base-cased	1.7686	1.7768
Tatar	Restaurant	xlm-roberta-base	2.0729	1.9380
Ukrainian	Restaurant	xlm-roberta-base	1.5937	1.7805
Chinese	Finance	bert-base-chinese	0.7259	1.9652
Chinese	Laptop	bert-base-chinese	0.7438	1.6440
Chinese	Restaurant	bert-base-chinese	0.9823	1.8959

Table 9: Track A results (baseline is Kimi-K2 Thinking).

Language	Domain	Model	Our RMSE	Baseline (Mistral-3 14B)
German	Politics	deberta-v3-base	1.5509	1.5910
English	Environment	deberta-v3-base	2.0893	1.6430
Nigerian Pidgin	Politics	xlm-roberta-base	1.7878	1.7390
Swahili	Politics	mBERT-swahili	2.2700	2.2990
Chinese	Environment	bert-base-chinese	0.7364	0.7400

Table 10: Track B results (baseline is Mistral-3 14B).

sifiers, punctuation, and discourse markers, that may not be prominently captured in CLS-based representations.

We did not experiment with specialized attention mechanisms for arousal, which remains future work. Another useful direction would be to analyze whether sentence-level or clause-level aggregation helps isolate intensity cues that are diluted when the entire review is compressed into a single pooled representation.

6 Conclusion

We presented the Pixel Phantoms system for SemEval-2026 Task 3, demonstrating a language-specific transformer selection strategy for dimensional aspect-based sentiment analysis. By deploying dedicated language models rather than a single generic multilingual encoder, we achieved competitive results in well-supported languages (Chinese restaurant: rank 13/24; Japanese Hotel: rank 13/21; German politics: rank 8/12).

Three takeaways emerge. First, language-specific models outperform generic multilingual fallbacks in dimensional VA regression, especially for typologically distant languages. Second, domain novelty remains challenging even for well-resourced languages. Third, low-resource languages with limited model coverage remain a major open problem, with RMSE gaps above 0.5 against

official baselines.

For future work, we plan to explore: cross-lingual transfer from related language families (e.g., Russian models for Ukrainian), ensemble methods combining language-specific and multilingual encoders, incorporation of affective lexicons as auxiliary features for Arousal prediction, and domain adaptation techniques for politics and environment domains. We also see value in better validation protocols for small multilingual datasets, including more systematic development splits and calibration analyses for the two affective dimensions. Such analyses could clarify when improvements come from stronger encoders, better output scaling, or better alignment between aspect representation and context. Another promising direction is lightweight parameter-efficient adaptation, which may allow stronger per-language specialization without the memory cost of full fine-tuning for every track. This could be especially useful when many language–domain combinations must be trained under limited shared-task compute budgets.

7 Ethical Considerations

Data biases: Dimensional sentiment annotations reflect the cultural and demographic background of annotators. Valence and Arousal norms differ across cultures; an aspect that evokes high arousal

in one linguistic community may not in another. Our models inherit any such biases present in the DimABSA dataset annotations.

Misuse potential: Dimensional sentiment systems could be repurposed for mass emotional profiling, manipulative political advertising, or automated identification of emotionally vulnerable individuals.

Low-resource language harms: Deploying inaccurate models for Tatar, Nigerian Pidgin, or Swahili risks amplifying errors in underrepresented communities. We caution against production use without domain-specific validation and community involvement. This is especially important when such systems may influence ranking, moderation, or access decisions. In such settings, even small systematic errors can disproportionately affect already underrepresented speakers and reinforce existing language-technology inequities.

Responsible use: We used only the officially distributed DimABSA dataset and collected no additional data. All experiments were conducted in accordance with the ACL Rolling Review ethics guidelines (<https://aclrollingreview.org/ethicsreviewertutorial>).

Acknowledgements

We thank the DimABSA task organizers for the dataset, evaluation infrastructure, and tutorial materials. We acknowledge Google Colab for providing GPU compute resources. We are grateful to the Tohoku NLP Group (cl-tohoku/bert-base-japanese-v3), DeepPavlov (rubert-base-cased), and Davlan (Swahili mBERT) for making language-specific models publicly available.

References

- 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.
- Alexis Conneau, Kartik 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 ACL*.
- Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: A dutch roberta-based language model. In *Findings of EMNLP*.
- 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*.
- Yue Gao and 1 others. 2022. Aspect-prompted encoding for aspect-based sentiment analysis. In *Proceedings of ACL*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *Proceedings of ICLR*.
- Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashchik, 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.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*. ArXiv:1907.11692.
- Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of ACL*.
- Maria Pontiki, Dimitris Galanis, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of SemEval*.
- Maria Pontiki, Dimitris Galanis, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of SemEval*.
- James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- Amy B. Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*.
- 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.