Zero-Shot to Full-Resource: Cross-lingual Transfer Strategies for Aspect-Based Sentiment Analysis

Jakob Fehle[1], **Nils Constantin Hellwig**[1], **Udo Kruschwitz**[2], **Christian Wolff**[1]
[1]Media Informatics Group, University of Regensburg, Regensburg, Germany
[2]Information Science Group, University of Regensburg, Regensburg, Germany
jakob.fehle@ur.de, nils-constantin.hellwig@ur.de, udo.kruschwitz@ur.de, christian.wolff@ur.de

Abstract

Aspect-based Sentiment Analysis (ABSA) extracts fine-grained opinions toward specific aspects within text but remains largely English-focused despite major advances in transformer-based and instruction-tuned models. This work presents a multilingual evaluation of state-of-the-art ABSA approaches across seven languages (English, German, French, Dutch, Russian, Spanish, and Czech) and four subtasks (ACD, ACSA, TASD, ASQP). We systematically compare different transformer architectures under zero-resource, data-only, and full-resource settings, using cross-lingual transfer, code-switching and machine translation. Fine-tuned Large Language Models (LLMs) achieve the highest overall scores, particularly in complex generative tasks, while few-shot counterparts approach this performance in simpler setups, where smaller encoder models also remain competitive. Cross-lingual training on multiple non-target languages yields the strongest transfer for fine-tuned LLMs, while smaller encoder or seq-to-seq models benefit most from code-switching, highlighting architecture-specific strategies for multilingual ABSA. We further contribute two new German datasets, an adapted GERestaurant and the first German ASQP dataset (GERest), to encourage multilingual ABSA research beyond English.

Keywords: Aspect-based Sentiment Analysis, Cross-Lingual, Resources, Large Language Models

1. Introduction

Aspect-based Sentiment Analysis (ABSA) has become a central task for mining fine-grained opinions, aiming to detect sentiment toward specific aspects within text. Despite substantial methodological advances, from transfer learning-based classifiers (Cai et al., 2020; Cui et al., 2024) to instruction-tuned large language models (LLMs) (Scaria et al., 2024; Šmíd et al., 2024a), research has remained largely English-focussed, driven by the abundance of English benchmark datasets (Chebolu et al., 2023).

Šmíd and Král (2025) emphasize that ABSA generalization to non-English languages remains challenging due to limited multilingual data diversity, high sensitivity to translation quality, and structural divergence between languages, and they further emphasize that cross-lingual transfer remains inconsistent across languages and tasks. Cross-lingual studies (Lin et al., 2023; Zhang et al., 2025; Šmíd et al., 2025b) similarly report inconsistent transfer, with multilingual encoders such as mBERT or XLM-R showing substantial performance gaps across languages even under comparable supervision. While multilingual pre-trained language models such as mBERT (Devlin et al., 2019) and mT5 (Xue et al., 2021), or more recent LLMs including GPT-4 (OpenAI et al., 2024), LLaMA 3 (Dubey et al., 2024), and Gemma 3 (Gemma Team et al., 2025), have substantially improved zero-shot and cross-lingual transfer capabilities, systematic evaluations across diverse ABSA subtasks and multiple languages remain limited, particularly for complex generative tasks beyond End-to-End ABSA (E2E) or single-element extraction (Šmíd and Král, 2025).

We address this gap by conducting a comprehensive evaluation of state-of-the-art (SOTA) methods across four established ABSA sub-tasks: Aspect Category Detection (ACD), Aspect Category Sentiment Analysis (ACSA), Target Aspect Sentiment Detection (TASD), and Aspect Sentiment Quad Prediction (ASQP). To ensure cross-lingual comparability, we base our experiments on the SemEval-2016 restaurant datasets (Pontiki et al., 2016), the most widely adopted multilingual ABSA resource (Hua et al., 2024), covering six languages (English, French, Spanish, Dutch, Russian, Turkish). We extend this benchmark with German (Hellwig et al., 2024) and Czech (Šmíd et al., 2024b) datasets following the same schema, and **contribute the first German ASQP dataset**, *GERest*, to enable cross-lingual ASQP evaluation.

We systematically compare three modeling paradigms:

(a) **Encoder-only classification**, which conceptualize ABSA as a supervised multi-label classification problem, including BERT-based architectures and graph convolutional networks (GCNs).

(b) **Sequence-to-sequence text generation**, which reformulates ABSA as a structured text generation task using T5-based models with predefined templates for input and output.

(c) **Decoder-only LLMs**, encompassing both few-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7999–8013
11–16 May 2026.

shot in-context prompting and instruction fine-tuning for structured sentiment extraction.

This setup allows us to compare how different architectures and learning paradigms handle cross- and multilingual conditions under a unified experimental framework. Experiments are conducted under three resource conditions:

1. In the **zero-resource setting**, neither annotated data nor language-specific models are available; models must rely solely on cross-lingual transfer capabilities.

2. In the **data-only setting**, annotated training data in the target language is available, but no dedicated language-specific model exists, requiring multilingual models to adapt to the language.

3. In the **full-resource setting**, both annotated data and language-specific pre-trained models are available, allowing us to assess the performance ceiling for each language.

To enhance zero-resource settings, we apply code-switching and machine-translation augmentation (Zhang et al., 2021a) to generate pseudo-training data from English.

Our study provides a comprehensive empirical analysis of multilingual ABSA. We quantify the trade-offs between multilingual and language-specific models, assess the limits of transfer-based methods in low-resource conditions, and highlight practical implications for adapting SOTA ABSA techniques across linguistically diverse settings. All code and results are available at GitHub.[1]

2. Related Work

In recent years, ABSA has seen substantial progress through both classification-based and generative modeling approaches.

2.1. State-of-the-Art Modeling Approaches for ABSA

Recent advances in ABSA span a continuum from supervised classification to generative and instruction-based approaches. For simpler subtasks such as ACD and ACSA, transformer-based classifiers like BERT-CLF (Fehle et al., 2023; Hellwig et al., 2024), graph-based models such as HierGCN (Cai et al., 2020), and attention-augmented variants like ECAN (Cui et al., 2024) have established strong baselines. More complex tasks, including TASD and ASQP, are increasingly modeled as a text generation problem using sequence-to-sequence architectures such as T5 (Raffel et al., 2020), where predefined (Zhang et al., 2021a) or dynamically ordered templates (Hu et al., 2022; Gou et al., 2023) are used to structure the outputs.

LLMs have further advanced ABSA in both supervised and unsupervised settings. Instruction fine-tuning, which reformulates ABSA inputs as natural language prompts, achieves SOTA results across multiple benchmarks (Varia et al., 2023; Simmering and Huoviala, 2023; Šmíd et al., 2024a; Fehle et al., 2026). These studies show that instruction-tuned models like T5 or LLaMA outperform traditional fine-tuning and few-shot prompting, with prompt design playing a minor role once models are instruction-aligned (Simmering and Huoviala, 2023). Šmíd et al. (2024a) further demonstrate that fine-tuned open-source LLMs can outperform proprietary ones, though most of current evaluations remain limited to English.

Recent research also focuses on ABSA under low-resource conditions, which has shifted from manual corpus creation (Akhtar et al., 2016; Rani and Anwar, 2020) to generative and hybrid annotation methods (Hu et al., 2022; Gou et al., 2023; Hellwig et al., 2025b) that leverage template variation and synthetic data for improved robustness.

Overall, ABSA has evolved from static classification toward flexible generative and instruction-based modeling, with augmentation methods providing the groundwork for multilingual and cross-lingual evaluation.

2.2. Cross-Lingual ABSA

Cross-lingual ABSA research has been driven by the availability of multilingual benchmarks, most notably the SemEval-2016 Task 5 datasets (Pontiki et al., 2016), which provide manually annotated restaurant, hotel, and laptop reviews across eight languages under a unified schema. Their accessibility has made them the de facto standard for multilingual evaluation. Later resources such as MultiAspectEmo (Szolomicka and Kocon, 2022) and M-ABSA (Wu et al., 2025c) expanded language coverage through translation or alignment but often lack gold-standard quality.

Early cross-lingual ABSA approaches transferred knowledge via bilingual embeddings or machine translation (Barnes et al., 2016; Jebbara and Cimiano, 2019; García-Pablos et al., 2018), but transfer performance remained limited for low-resource languages. With the advent of multilingual pre-trained encoders such as mBERT and XLM-R, encoder-only architectures became the dominant paradigm (Phan et al., 2021; Van Thin et al., 2023). Later studies enhanced zero-shot transfer through augmentation techniques like aspect code-switching (Zhang et al., 2021b), contrastive align-

[1] GitHub: https://github.com/JakobFehle/Cross-lingual-Transfer-Strategies-for-ABSA

ment and distillation (Lin et al., 2023), and synthetic data generation or consistency regularization (Wu et al., 2025a; Šmíd et al., 2025a), showing that unlabeled or LLM-generated target-language data can partially compensate for missing annotations.

Šmíd and Král (2025) provide a comprehensive survey of cross-lingual ABSA and identify remaining challenges, including reliance on translation quality, limited language diversity of available benchmarks, and weak generalization of models to morphologically rich languages. They further highlight the need for unified cross-lingual evaluations across multiple ABSA subtasks.

Recent work extends this direction with prompting- and generation-based methods: constrained decoding with mT5 and LLaMA-3 improves zero-shot transfer (Šmíd et al., 2025b), while few in-language examples can already yield substantial gains (Šmíd et al., 2025c), which underlines the strong potential of minimal in-language supervision as a cost-efficient alternative to large-scale annotation. Similarly, Wu et al. (2024) report that GPT- and LLaMA-based prompting captures sentiment cross-lingually but still trails fine-tuned models.

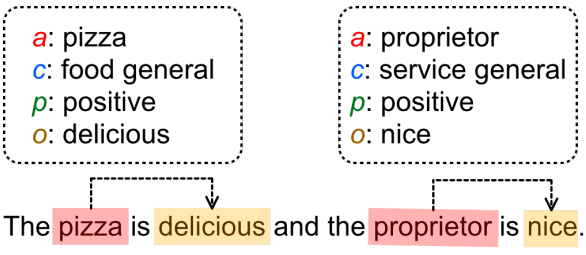
Overall, the literature reveals three main trends: (1) encoder-only models remain competitive for simpler classification-level tasks; (2) constrained or template-based generation performs best on complex subtasks such as TASD or ASQP; and (3) decoder-only LLMs generalize well but typically require adaptation or fine-tuning. Building on these insights, our work systematically compares all three paradigms, classification, sequence generation, and LLM-based methods, across multiple languages. In contrast to prior studies that rely on individual pretrained models or isolated techniques, we explore combinations of code-switching, machine translation, and multilingual training, in combination with several SOTA approaches.

3. Methodology

3.1. Tasks

ABSA comprises a number of subtasks that differ in the level of detail and the type of information they extract from the text. In this work, we focus on four common ABSA tasks with different levels of granularity that are supported by the structure and annotations of our multilingual datasets: Aspect Category Detection (ACD), Aspect Category Sentiment Analysis (ACSA), Target Aspect Sentiment Detection (TASD), and Aspect Sentiment Quad Prediction (ASQP) (Zhang et al., 2023). See Table 1 for examples illustrating the input and expected output of each task:

- ACD focuses on identifying all aspect categories mentioned or implied in a given input text.
- ACSA adds sentiment polarity classification (positive, neutral, negative) for each detected aspect.
- TASD further builds upon ACSA by requiring the detection of the exact text spans that represents the target of the expressed sentiment (aspect term). For implicit aspect terms, i.e., if no explicit phrase represents the aspect category in the text, the model is expected to return the value "NULL".
- ASQP aims to jointly extract all four components of a sentiment expression: the aspect term, its corresponding opinion term, the associated aspect category, and the expressed sentiment polarity, thus unifying extraction and sentiment classification in a single structured task.



Subtask	Output
Aspect Category Detection (ACD)	(c)
Aspect Category Sentiment Classification (ACSA)	(c, p)
Target Aspect Sentiment Detection (TASD)	(a, c, p)
Aspect Sentiment Quad Prediction (ASQP)	(a, c, o, p)

Table 1: Overview of ABSA subtasks used in this study with their expected outputs. Input for all tasks is the text sentence.

3.2. Datasets

To evaluate the multilingual applicability of SOTA methods for ABSA, we rely on a diverse set of datasets covering multiple languages and ABSA subtasks.

3.2.1. ACD, ACSA, and TASD

For ACD, ACSA, and TASD, we use the multilingual restaurant review datasets from SemEval-2016 Task 5 (Pontiki et al., 2016), employing the standardized training and test splits for five languages (English, Spanish, French, Dutch, and Russian) annotated with aspect terms, aspect categories, and sentiment polarity. All datasets share a unified 12-category schema, ensuring consistency across languages.

To expand the linguistic coverage, we added two further corpora: (1) GERestaurant (Hellwig et al., 2024), a German dataset we aligned to the SemEval schema through manual re-annotation of aspect categories, and (2) the Czech dataset by Šmíd et al. (2024b), which already adheres to the same guidelines. This extended collection allows evaluation of cross-lingual robustness beyond the original SemEval languages.

3.2.2. ASQP

Since the original SemEval 2016 Task 5 datasets do not contain annotations for explicit opinion phrases, they are not directly suitable for ASQP.

For English, we adopt the dataset by Zhang et al. (2021a), which extends SemEval-2016 with opinion-phrase labels and serves as the standard benchmark for generative ASQP models (Gou et al., 2023; Bai et al., 2024).

To extend the ASQP evaluation beyond English and enable a controlled cross-lingual comparison, we created a German counterpart (*GERest*). Starting from the existing GERestaurant dataset (Hellwig et al., 2024), we manually annotated opinion terms for a subset matching the size of the English ASQP-Rest16 corpus (1,264 train; 316 dev; 544 test samples), following the same annotation guidelines.

Further details on the dataset and its annotation process are provided in Appendix A. To prevent potential contamination of the GERest training and test sets in future language models, the dataset is not publicly released but is available upon reasonable request from the authors.

3.3. Preprocessing

To ensure that cross-lingual comparisons are not affected by differences in class distribution, we constructed balanced subsets for each language that share a similar distribution across the 12 predefined aspect categories. Instead of preserving the original imbalanced distributions, we aimed to approximate a unified target distribution applicable to all datasets.

We first computed the global frequency of each category across languages and determined the maximum common support per class, which is the largest number of samples consistently available in all datasets. These values served as sampling targets, allowing for minor flexibility. We then randomly sampled instances per language to match this shared distribution as closely as possible, resulting in seven balanced training sets of 1,162 examples each. Figure 1 illustrates the resulting category alignment and dataset sizes across all languages.

3.4. Evaluated Approaches

To ensure comparability across languages, we implement representative SOTA ABSA methods originally developed for the English SemEval-2016 restaurant dataset. Since most of these methods have not been systematically applied to other languages, we adopt their original hyperparameters and configurations with minimal adjustments. Methods were selected based on reported performance and the availability of reproducible implementations.

To assess the cross-lingual portability of SOTA ABSA methods, we experiment with three complementary paradigms: (1) encoder-based classification, (2) sequence-to-sequence generation, and (3) decoder-only LLMs with prompting or instruction fine-tuning.

3.4.1. Encoder-Only Classification

- **BERT-CLF:** A multi-label transformer classifier predicting aspect categories (e.g., `food`) or aspect-sentiment pairs (e.g., `food:positive`) per sentence, following Fehle et al. (2023) and Hellwig et al. (2024).
- **Hier-GCN:** Extends BERT with hierarchical graph convolutional layers to capture dependencies between aspects and sentiments (Cai et al., 2020).

Models: Multilingual *mBERT*[2] (Devlin et al., 2019) and language-specific variants, such as *ruBERT*[3] (Kuratov and Arkhipov, 2019) for Russian.

3.4.2. Seq-2-Seq Text Generation

- **DLO:** Dynamic Label Ordering (Hu et al., 2022) reformulates ABSA as a generative task by dynamically augmenting and reordering output tuples (e.g., for TASD, ASQP), improving alignment between input and structured outputs.

Models: Multilingual *mT5*[4] (Xue et al., 2021) and, where available, monolingual T5 variants, e.g., *ruT5*[5] (Zmitrovich et al., 2023).

3.4.3. Decoder-Only LLMs

- **Zero-/Few-Shot Prompting:** This approach uses instruction- and few-shot-based in-context learning (ICL) on a LLM with either

[2] https://huggingface.co/google-bert/bert-base-multilingual-cased

[3] https://huggingface.co/DeepPavlov/rubert-base-cased

[4] https://huggingface.co/google/mt5-base

[5] ai-forever/ruT5-base

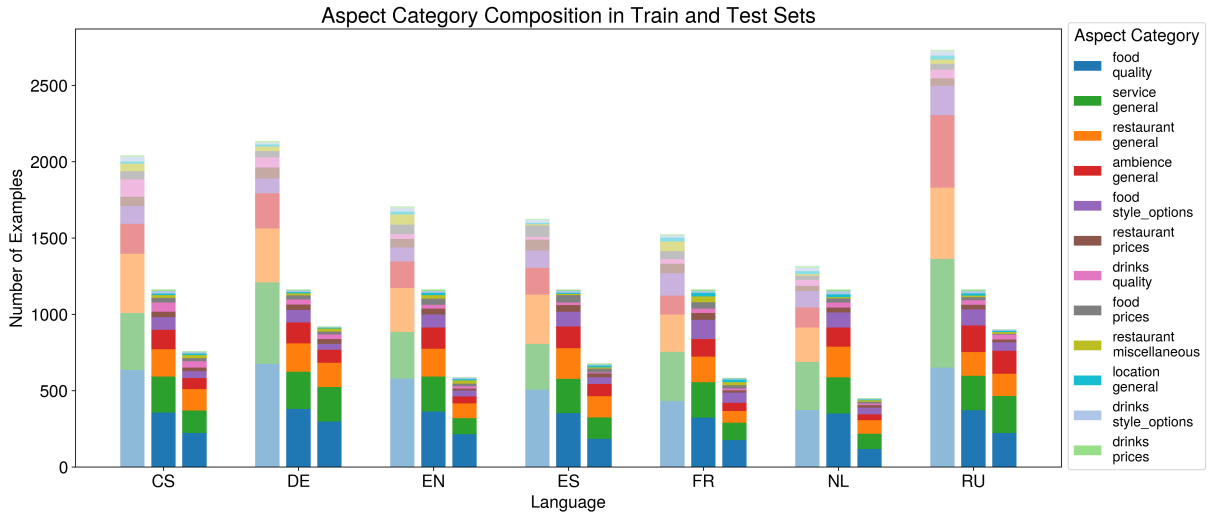


Figure 1: The diagram illustrates both the absolute dataset sizes and aspect category distributions across languages and splits. For each language, the three grouped bars represent (from left to right): original train, balanced train, and original test set. Each bar is stacked by relative aspect category distribution. Lighter colors indicate the original train split, while darker bars represent the balanced train and original test splits used in our experiments.

zero or a predefined set of annotated ABSA examples embedded in prompts.

- **Instruction Fine-Tuning:** Supervised fine-tuning of LLMs on task-specific instructions, updating model weights for ABSA generation. We use the same prompt structure as in the zero-/few-shot setting to maintain consistency across methods.

Models and Implementation: We use the same multilingual checkpoints for all languages: Gemma 3 27B[6] (Gemma Team et al., 2025) with up to 50 annotated ABSA examples for zero-/few-shot prompting and LLaMA 3.1 8B[7] (Dubey et al., 2024) for instruction fine-tuning. The selection of LLMs and amount of few-shots is based on results achieved by Hellwig et al. (2025a). Prompt templates are based on Gou et al. (2023), adapted to the specific subtasks and translated into each target language to ensure structural and linguistic consistency. Prompt examples for English are provided in Appendix B.

Fine-tuning utilizes Parameter-Efficient Fine-Tuning (PEFT) (Mangrulkar et al., 2022) and is performed with Quantized Low-Rank Adaption (QLoRA) (Dettmers et al., 2023) using the *unsloth*[8] framework.

[6] https://huggingface.co/google/gemma-3-27b-it

[7] https://huggingface.co/meta-llama/Llama-3.1-8B

[8] https://github.com/unslothai/unsloth

3.5. Systematic Evaluation and Hyperparameter Calibration

To ensure fair cross-lingual comparisons, we adopt a two-stage evaluation strategy for each model and subtask. First, we determine the optimal number of training epochs via an 80/20 split of the source training data, keeping all other hyperparameters (learning rate, batch size, optimizer settings) constant across languages and tasks. We explore predefined epoch ranges depending on model type: 20 – 50 for BERT-based classifiers, 15 – 30 for sequence-to-sequence models, and 5 – 20 for fine-tuned LLMs. These ranges are proportionally reduced for zero-resource setups that rely on larger pseudo-training sets (e.g., cross-lingual transfer or code-switching) to prevent overfitting. Second, models are retrained on the complete training data using the best-performing epoch count and evaluated on the held-out test set. Each configuration is repeated with five random seeds, and final results are averaged to mitigate stochastic variation. By pairing a multilingual baseline with a monolingual upper bound, we derive both a realistic transfer estimate and a ceiling for in-language performance.

We test significance ($p_{adj} \leq 0.05$) using parametric or non-parametric tests (ANOVA/*t*-test or Friedman/Wilcoxon (Field et al., 2012)) with Bonferroni-Holm correction (Holm, 1979) to assess differences in performance between multilingual and language-specific models, language-specific variability in training outcomes, and variation in zero-resource transfer effectiveness across tasks and languages.

3.5.1. In-Language Supervised Settings

To ensure fair cross-lingual comparability and consistent label distributions across languages, we employ a balanced dataset configuration. This setup equalizes dataset sizes and aspect category distributions as described in Section 3.3. Multilingual models are fine-tuned to simulate scenarios where moderately sized target-language data are available but no language-specific model exists, while monolingual pretrained models serve as upper bounds for in-language performance.

3.5.2. Zero-Resource Settings

To evaluate the transfer capabilities of supervised multilingual models in the absence of target-language training data, we utilize three zero-resource ABSA strategies:

- **Cross-Lingual Transfer (CLT):** The model is trained on all balanced training sets except for the target language.
- **Code-Switching (CS):** Augments each English training sentence with several variants, original, translated, and mixed, by replacing key lexical items (e.g., aspect and sentiment terms) with their target-language counterparts using LLMs. This hybrid data provides weak cross-lingual supervision and exposes the model to bilingual lexical patterns while maintaining grammatical structure.
- **Machine Translation (MT):** The English training data is automatically translated into the target language using LLMs.

To strengthen the generalizability of our work and based on the assumption that optimized monolingual models are generally unavailable under such conditions, all zero-resource experiments are conducted using multilingual models only.

4. Results

4.1. Results for Monolingual Training

In the monolingual setting, where training and testing use the same language, we compare two configurations on the balanced datasets (see Table 2): Multi = a multilingual model (e.g., *mT5-base*) fine-tuned per language, and Spec = a language-specific model (e.g., *ruT5-base* for Russian) fine-tuned on the same data. This allows a direct comparison between multilingual and monolingual pretraining. We make four main observations:

Inter-language variability Performance under the **Multi** configuration varies considerably across languages and tasks. In both ACD and ACSA, multilingual models achieve their strongest results for German and Spanish, while a language such as French shows consistently weaker performance. The same trend appears in TASD, where the disparity between high- and low-performing languages widens further. However, these cross-linguistic differences are only statistically significant in a few isolated cases. Interestingly, this pattern holds across architectures: encoder-only classifiers, graph-based models, and fine-tuned LLMs show similar cross-lingual ranking orders, suggesting that the limitations stem less from model design and more from the representational mismatch between multilingual embeddings and the typological diversity of target languages.

(a) ACD

Method	Setting	CS	DE	EN	ES	FR	NL	RU
BERT-CLF	Multi	77.18	78.19	75.42	76.62	71.48	73.69	76.89
	Spec	80.98	82.56	80.79	79.53	71.81	75.62	81.16
LLaMA 3.1 8B FT	Multi	75.64	78.40	83.82	80.33	76.84	80.27	79.40
Gemma 3 27B 50-shot	50-shot	74.52	77.38	77.38	75.16	70.06	78.63	75.23

(b) ACSA

Method	Setting	CS	DE	EN	ES	FR	NL	RU
BERT-CLF	Multi	61.37	64.73	57.15	63.92	52.37	54.24	57.56
	Spec	69.63	72.56	65.34	68.91	50.56	61.48	63.29
Hier-GCN	Multi	65.39	67.74	63.68	68.25	58.39	60.36	61.64
	Spec	70.66	76.09	70.22	72.46	60.97	67.11	67.56
LLaMA 3.1 8B FT	Multi	70.66	76.30	79.89	76.18	70.02	74.07	73.00
Gemma 3 27B	50-shot	70.12	76.51	75.68	72.49	66.78	75.06	73.68

(c) TASD

Method	Setting	CS	DE	EN	ES	FR	NL	RU
DLO	Multi	54.60	48.10	50.37	57.94	46.70	46.35	49.68
	Spec	–	59.59	70.55	58.74	50.99	59.12	59.87
LLaMA 3.1 8B FT	Multi	61.09	63.27	69.95	64.09	56.16	62.16	62.40
Gemma 3 27B	50-shot	60.58	62.94	68.53	59.29	55.14	58.40	58.90

Table 2: Supervised F1-Micro scores per language for each method and dataset/model resource setting, split by task. Abbr.: Spec = language-specific model; Multi = multilingual model. For TASD, Czech results are omitted ("–") as no language-specific T5 model exists.

Consistent monolingual gains when using language-specific models Across all languages and modeling paradigms, using language-specific pretrained models (**Spec**) consistently improves performance over their multilingual counterparts (**Multi**). The improvement is modest for simpler classification tasks such as ACD ($\approx$ 3.3 F1 points on average) and ACSA ($\approx$ 5.6 points), but becomes substantially larger for the more complex TASD task ($\approx$ 8.5 points). These gains are statistically significant across all tasks, confirming the robustness of

the monolingual advantage and consistent performance benefit that multilingual architectures have not yet fully bridged. Even advanced LLMs still have weaknesses in certain languages and tasks, suggesting that monolingual pretraining remains a key factor for achieving SOTA ABSA performance. Similar trends are observed in other NLP tasks, where monolingual models, such as FinBERT (Virtanen et al., 2019), CamemBERT (Martin et al., 2020), RobeCzech (Straka et al., 2021), and others (Ulčar et al., 2026), consistently outperform multilingual counterparts across multiple languages and benchmarks.

Superior performance of fine-tuned LLMs The fine-tuned LLaMA 3.1 8B achieves the strongest overall results in our supervised comparisons, consistently outperforming both multilingual and monolingual encoder-based models across most languages and tasks. Its advantage is most pronounced for complex generative subtasks like TASD, confirming the benefit of large-scale decoder architectures for structured sentiment extraction. However, this superiority is not universal. In some languages and simpler tasks (e.g., ACD or ACSA), fine-tuned BERT-based classifiers still match or even exceed multilingual LLM performance. These findings highlight a nuanced picture: fine-tuned LLMs excel in complex, structured ABSA subtasks and offer the highest performance for multilingual adaptation, but they do not yet replace specialized models as a universal solution across all languages and task granularities.

Notably, the few-shot Gemma 3 27B model shows strong robustness without fine-tuning, surpassing supervised baselines in ACSA for German, Dutch, and Russian, and performing on par with most language-specific SOTA models elsewhere. While it falls short in ACD and TASD, its consistent performance underscores the potential of few-shot prompting as a lightweight alternative to full fine-tuning.

First insights into multilingual ASQP experiments For the most complex ASQP task (available only in English and German), the overall impression remains (see Table 3d): both supervised methods, the generative DLO and the fine-tuned LLaMA 3.1 8B, achieve the highest scores (59.35 / 48.16 F1-micro for DLO vs. 57.93 / 53.44 F1-micro for LLaMA). In contrast, the few-shot Gemma 3 27B trails behind (51.10 / 41.47 F1-micro), despite remaining well above zero-shot levels. These results underline that the performance gap between few-shot prompting and fully supervised learning widens as task complexity increases.

(a) ACD

Method	Setting	CS	DE	ES	FR	NL	RU
BERT-CLF	Supervised	77.18	78.19	76.62	71.48	73.69	76.89
	CLT	61.51	71.83	67.46	67.74	64.86	67.48
	CS	69.05	71.41	69.42	68.36	67.31	69.76
	MT	66.82	70.21	67.66	66.34	66.02	67.81
LLaMA 3.1 8B FT	Supervised	83.31	84.52	84.80	81.65	85.56	87.28
	CLT	**82.27**	**84.51**	**81.90**	**81.49**	**83.64**	**86.45**
	CS	80.02	81.43	81.39	80.22	80.02	84.76
	MT	78.26	80.49	80.66	79.84	80.06	83.71
Gemma 3 27B	50-shot	74.52	77.38	75.16	70.06	78.63	75.23
	0-shot	66.11	68.42	66.35	59.10	69.50	73.28

(b) ACSA

Method	Setting	CS	DE	ES	FR	NL	RU
BERT-CLF	Supervised	61.37	64.73	63.92	52.37	54.24	57.56
	CLT	40.57	50.78	54.32	43.80	46.98	49.18
	CS	48.21	55.78	50.10	48.92	48.98	53.59
	MT	44.94	53.50	48.95	45.74	47.68	49.41
Hier-GCN	Supervised	65.39	67.74	68.25	58.39	60.36	61.64
	CLT	42.50	52.50	56.16	44.96	45.39	44.89
	CS	55.78	59.67	60.49	50.21	55.99	56.33
	MT	54.40	58.72	59.48	48.11	53.98	55.23
LLaMA 3.1 8B FT	Supervised	75.16	81.48	79.40	73.30	78.99	79.50
	CLT	**73.76**	**80.75**	**77.40**	**73.41**	**77.90**	**78.46**
	CS	71.42	77.41	76.20	70.36	74.63	75.51
	MT	70.03	76.61	74.38	70.64	73.86	74.96
Gemma 3 27B	50-shot	70.12	76.51	72.49	66.78	75.06	73.68
	0-shot	69.39	73.50	68.21	62.40	70.27	68.05

(c) TASD

Method	Setting	CS	DE	ES	FR	NL	RU
DLO	Supervised	54.60	48.10	57.94	46.70	46.35	49.68
	CLT	24.74	46.51	38.87	38.08	38.60	18.51
	CS	43.30	48.73	51.37	39.12	38.10	43.94
	MT	40.28	39.86	48.90	36.71	40.49	43.70
LLaMA 3.1 8B FT	Supervised	65.43	67.89	68.30	61.98	65.85	65.72
	CLT	53.54	**66.28**	60.18	**58.12**	**62.98**	57.61
	CS	**56.86**	62.86	**61.89**	55.99	59.27	**59.26**
	MT	55.27	61.21	60.73	53.99	59.08	57.75
Gemma 3 27B	50-shot	60.58	62.94	59.29	55.14	58.40	58.90
	0-shot	47.74	47.95	40.43	34.53	36.22	38.03

(d) ASQP

Method	Setting	EN	DE
DLO	Supervised	59.35	48.16
	CS	–	**36.92**
	MT	–	36.24
LLaMA 3.1 8B FT	Supervised	57.93	53.44
	CS	–	41.06
	MT	–	**42.95**
Gemma 3 27B	50-shot	51.10	41.47
	0-shot	28.96	17.90

Table 3: F1-micro scores per language and method under zero-resource conditions. CLT: trained on all languages except the target; CS: code-switched data; MT: machine-translated data. Supervised/50-shot scores are included for comparison. Bold values indicate the best-performing configuration.

4.2. Results for Multi-/Crosslingual Training

After establishing the supervised monolingual upper bound, we now assess ABSA portability under multi- and cross-lingual training, examining generalization from non-target languages or augmented data and how far multilingual pretraining substitutes in-language supervision. We compare the effects of task complexity, augmentation strategy, and model architecture on cross-lingual transfer.

Overall cross-lingual transfer performance and task complexity effects. Across all tasks and languages, a clear degradation is observed when moving from supervised to zero-resource conditions. Models trained on all non-target languages (CLT) consistently underperform their supervised counterparts, confirming the persistent gap between fine-tuning and actual transfer generalization. The magnitude of this gap, however, varies strongly with task complexity: for simpler classification tasks such as ACD and ACSA, multilingual fine-tuned models still provide competitive performance, often within 5 – 8 F1 points of the supervised baseline. In contrast, for more structurally complex tasks like TASD and ASQP, transfer performance drops substantially, with relative declines exceeding 20 points in several cases. This pattern underscores that while multilingual fine-tuning is sufficient for coarse-grained sentiment classification, the extraction of structured opinion relations (e.g., TASD triplets) still benefits from task- or language-specific adaptation. Nevertheless, all multilingual models achieve actionable results even without language-specific pretraining, enabling meaningful ABSA evaluation for low-resource or underrepresented languages where dedicated models and datasets are unavailable. Notably, the fine-tuned LLaMA 3.1 8B model demonstrates the most robust cross-lingual generalization, maintaining over 81 F1-micro in ACD and roughly 73 – 80 F1-micro in ACSA even under zero-resource conditions, whereas encoder-only architectures (BERT, Hier-GCN) suffer the largest degradation.

Impact of zero-resource strategies. We compare three zero-resource transfer strategies, CLT, CS, and MT, relative to supervised upper bounds. For encoder-based and seq-to-seq models, CS yields the strongest relative performance, exceeding CLT and MT by 3–6 F1 points on average and reducing the gap to supervised baselines. This supports findings by Zhang et al. (2021b) and Wu et al. (2025b), who attribute CS gains to increased lexical diversity, while Šmíd et al. (2025b) and Zhang et al. (2025) show that adding synonym replacement or distillation further stabilizes transfer, while outperforming direct translation-based approaches. In contrast, MT-based augmentation brings smaller improvements, suggesting that data quantity and variation matter more than grammatical accuracy, particularly for encoder-based models, which appear to benefit from the higher lexical diversity introduced by code-switching. Fine-tuned LLMs usually achieve their best results under CLT conditions, often matching or approaching supervised performance, indicating that large decoder-based models already internalize cross-lingual alignment without additional augmentation. Overall, while CS and MT are valuable for smaller architectures, CLT proves most effective for large fine-tuned LLMs in zero-resource evaluation.

Comparative robustness of fine-tuned vs. few-shot LLMs. The fine-tuned LLaMA 3.1 8B consistently achieves the highest zero-resource transfer scores across all subtasks, with at least one zero-resource configuration performing significantly better than all other approaches. However, the few-shot Gemma 3 27B model exhibits remarkable stability across languages without any fine-tuning: for ACD and ACSA, it almost matches the supervised encoder baselines, while for more complex tasks such as TASD and ASQP, the performance drop is more pronounced. This cross-lingual consistency underscores the utility of instruction models for low-resource scenarios where no language-specific supervision (neither datasets nor specialized models) is available. These observations align with the findings of Šmíd et al. (2025a), who show that LLM-based augmentation and in-context learning can approach supervised performance in low-resource settings, though performance gains diminish for larger model scales.

5. Conclusion & Future Work

This work presented a comprehensive multilingual evaluation of SOTA approaches for ABSA across seven languages and four subtasks. By comparing encoder-only, sequence-to-sequence, and decoder-only architectures under varying resource conditions, we analyzed how well current models generalize across languages and ABSA tasks.

Our results reveal a clear hierarchy: instruction fine-tuned LLMs achieve the highest overall scores, particularly in complex generative tasks such as TASD and ASQP. Smaller encoder-based models remain competitive for simpler classification tasks (ACD, ACSA), offering strong performance with lower computational costs.

Language-specific models still outperform multilingual ones, though this gap narrows with larger, more multilingual architectures. Code-switching yields the most consistent improvements in zero-resource settings, while cross-lingual training on

non-target languages allows fine-tuned LLMs to approach supervised performance. Gemma 3 27B achieves competitive zero-shot results on simpler tasks but declines on more complex ones such as ASQP; in few-shot mode, however, it stabilizes and approaches fine-tuned performance, making it a practical low-resource alternative.

Beyond empirical findings, we contribute two new German resources: an adapted *GERestaurant* dataset aligned with the SemEval aspect-category schema, and the first German ASQP dataset (*GERest*) for structured opinion extraction. These additions extend ABSA research beyond English and enable controlled cross-lingual ASQP evaluation.

Future research should extend multilingual ABSA evaluation to additional domains and typologically diverse languages to test the generalizability of current methods. Furthermore, while we cover representative SOTA models across major paradigms, additional approaches, such as hybrid syntactic LLM approaches (Negi et al., 2024) or dual-stream data synthesis frameworks (Xu et al., 2025), which combine structural linguistic information or synthetic data generation, could provide further insights. Finally, evaluating instruction-tuned LLMs in few-shot and semi-supervised scenarios across new domains will be key to understanding their practical potential for multilingual sentiment analysis at scale.

Limitations & Ethical Considerations

While this study provides a broad multilingual benchmark for ABSA, several limitations remain. As the benchmark datasets used here were released before the pretraining of the evaluated models, potential data contamination cannot be entirely ruled out, even though no direct overlap between models and datasets is documented. Nevertheless, these resources represent the de facto standard for multilingual and cross-lingual ABSA research and were therefore used to ensure comparability with prior work.

In addition, to ensure comparable label distributions across languages and experimental settings, we employed balanced versions of the datasets, which may differ from naturally occurring sentiment and aspect distributions in the original data and therefore are less representative of real-world scenarios.Moreover, only a few datasets offer a sufficiently broad and multilingual foundation to enable systematic cross-lingual evaluation at this scale.

To further validate and generalize our findings, future studies should extend cross-lingual ABSA to additional domains (e.g., product reviews or social media), such as the English OATS (Chebolu et al., 2024) or FlightABSA (Hellwig et al., 2025a) datasets.

Additionally, although we include seven languages, further evaluations on typologically diverse or low-resource languages are needed to better assess transfer robustness.

Moreover, while our experiments cover representative encoder-, sequence-to-seqence-, and decoder-based architectures, more recent transformer variants such as ModernBERT (Warner et al., 2024) could provide additional insights into cross-lingual generalization. We deliberately refrained from including these architectures in this study, as language-specific pretrained versions of such models (e.g., ModernGBERT (Ehrmanntraut et al., 2025)) are still scarce, limiting their comparability to existing baselines.

From an ethical standpoint, no new user data were collected for this study. The newly contributed German datasets (adapted *GERestaurant* and new *GERest*) were derived from existing, anonymized corpora and contain no personally identifiable information. We used Claude 4.0[9] for support in code optimization and linguistic editing; all methodological decisions, analyses, and reported results were manually created and verified by the authors to avoid automated bias or factual distortion.

6. Bibliographical References

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in Hindi: Resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2703–2709, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhao Bai, Zhixin Han, Yuhua Zhao, Hang Gao, Zhuowei Zhang, Xunzhi Wang, and Mengting Hu. 2024. Is compound aspect-based sentiment analysis addressed by LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7836–7861, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeremy Barnes, Patrik Lambert, and Toni Badia. 2016. Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623. The COLING 2016 Organizing Committee.

[9]Claude Sonnet: https://www.anthropic.com/claude/sonnet

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 833–843, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2023. A review of datasets for aspect-based sentiment analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. OATS: Opinion aspect target sentiment quadruple extraction dataset for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 12336–12347.

Jin Cui, Fumiyo Fukumoto, Xinfeng Wang, Yoshimi Suzuki, Jiyi Li, Noriko Tomuro, and Wanzeng Kong. 2024. Enhanced coherence-aware network with hierarchical disentanglement for aspect-category sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 5843–5855. ELRA and ICCL.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Anton Ehrmanntraut, Julia Wunderle, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. Moderngbert: German-only 1b encoder model trained from scratch. *arXiv [cs.CL]*.

Jakob Fehle, Udo Kruschwitz, Nils Constantin Hellwig, and Christian Wolff. 2026. Leveraging fine-tuning of large language models for aspect-based sentiment analysis in resource-scarce environments. *Knowl. Based Syst.*, 336(115277):115277.

Jakob Fehle, Leonie Münster, Thomas Schmidt, and Christian Wolff. 2023. Aspect-based sentiment analysis as a multi-label classification task on the domain of german hotel reviews. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 202–218. Association for Computational Linguistics.

Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. SAGE.

Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2VLDA: Almost unsupervised system for aspect based sentiment analysis. *Expert Syst. Appl.*, 91:127–137.

Gemma Team et al. 2025. Gemma 3 technical report. *arXiv [cs.AI]*.

Zhibin Gou, Qi Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397. Association for Computational Linguistics.

Nils Constantin Hellwig, Jakob Fehle, Markus Bink, and Christian Wolff. 2024. GERestaurant: A german dataset of annotated restaurant reviews for aspect-based sentiment analysis. volume Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), page 123–133. Association for Computational Linguistics.

Nils Constantin Hellwig, Jakob Fehle, Udo Kruschwitz, and Christian Wolff. 2025a. Do we still need human annotators? prompting large language models for aspect sentiment quad prediction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 153–172, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nils Constantin Hellwig, Jakob Fehle, and Christian Wolff. 2025b. Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Systems with Applications*, 261(125514):125514.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2486–2495, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv [cs.CL]*.

Nankai Lin, Yingwen Fu, Xiaotian Lin, Dong Zhou, Aimin Yang, and Shengyi Jiang. 2023. CL-XABSA: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2935–2946.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gaurav Negi, Rajdeep Sarkar, Omnia Zayed, and P Buitelaar. 2024. A hybrid approach to aspect based sentiment analysis using transfer learning. *LREC*, pages 647–658.

OpenAI et al. 2024. Gpt-4 technical report. *arXiv [cs.CL]*.

Khoa Thi-Kim Phan, Duong Ngoc Hao, Dang Van Thin, and Ngan Luu-Thuy Nguyen. 2021. Exploring zero-shot cross-lingual aspect-based sentiment analysis using pre-trained multilingual language models. In *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud Maria Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5 : aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Sadaf Rani and Muhammad Waqas Anwar. 2020. Resource creation and evaluation of aspect based sentiment analysis in urdu. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 79–84, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paul F Simmering and Paavo Huoviala. 2023. Large language models for aspect-based sentiment analysis. *arXiv [cs.CL]*.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue*, Lecture notes in computer science, pages 197–209. Springer International Publishing, Cham.

Joanna Szolomicka and Jan Kocon. 2022. Multi-AspectEmo: Multilingual and language-agnostic

aspect-based sentiment analysis. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 443–450. IEEE.

Matej Ulčar, Aleš Žagar, Carlos S Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja. 2026. Mono- and cross-lingual evaluation of representation language models on less-resourced languages. *Comput. Speech Lang.*, 95(101852):101852.

Dang Van Thin, Hung Quoc Ngo, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023. Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models. *J. Inf. Telecommun.*, 7(2):121–143.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. Instruction tuning for few-shot aspect-based sentiment analysis. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv [cs.CL]*.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9122–9129.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv [cs.CL]*.

Chengyan Wu, Bolei Ma, Ningyuan Deng, Yanqing He, and Yun Xue. 2025a. Multi-scale and multi-objective optimization for cross-lingual aspect-based sentiment analysis. *arXiv [cs.CL]*.

Chengyan Wu, Bolei Ma, Ningyuan Deng, Yanqing He, and Yun Xue. 2025b. Multi-scale and multi-objective optimization for cross-lingual aspect-based sentiment analysis. *arXiv [cs.CL]*.

Chengyan Wu, Bolei Ma, Yihong Liu, Zheyu Zhang, Ningyuan Deng, Yanshu Li, Baolan Chen, Yi Zhang, Barbara Plank, and Yun Xue. 2025c. M-ABSA: A multilingual dataset for aspect-based sentiment analysis. *arXiv [cs.CL]*.

Chengyan Wu, Bolei Ma, Zheyu Zhang, Ningyuan Deng, Yanqing He, and Yun Xue. 2024. Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models. *Int. J. Mach. Learn. Cybern.*

Hongling Xu, Yice Zhang, Qianlong Wang, and Ruifeng Xu. 2025. DS^2-ABSA: Dual-stream data synthesis with label refinement for few-shot aspect-based sentiment analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 15460–15478. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. MT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bolun Zhang, Yahui Zhao, Guozhe Jin, and Rongy Cui. 2025. Cross-lingual aspect-based sentiment analysis based on semi-supervised knowledge distillation. In *2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 57–61. IEEE.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021b. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021c. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 504–510, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, M Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, V Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian. *LREC*, pages 507–524.

Jakub Šmíd and Pavel Král. 2025. Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges. *Inf. Fusion*, 120(103073):103073.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2024a. LLaMA-based models for aspect-based sentiment analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2025a. LACA: Improving cross-lingual aspect-based sentiment analysis with LLM data augmentation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–853, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2025b. Advancing cross-lingual aspect-based sentiment analysis with LLMs and constrained decoding for sequence-to-sequence models. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, pages 757–766. SCITEPRESS - Science and Technology Publications.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2025c. Few-shot cross-lingual aspect-based sentiment analysis with sequence-to-sequence models. In *International Conference on Text, Speech, and Dialogue*, pages 27–38. Springer.

Jakub Šmíd, Pavel Přibáň, Ondřej Pražák, and Pavel Král. 2024b. Czech dataset for complex aspect-based sentiment analysis tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4299–4310. ELRA and ICCL.

A. GERest

GERest is derived from the TASD dataset GERestaurant, introduced by Hellwig et al. (2024). This dataset was prepared with the aim of mirroring the structure of ASQP-Rest16, ensuring comparable quantities of training, validation, and test examples. The original GERestaurant comprises a training set with 2,154 examples and a test set with 924 examples, but does not include a dedicated validation set. To address this issue, a subset of the training examples was allocated as a validation set for GERest. All examples were refined by introducing an additional opinion term. The final dataset distribution is as follows:

- **Training:** 1,264 examples derived from GERestaurant's training set.
- **Validation:** 316 examples derived from GERestaurant's training set.
- **Test:** 544 examples derived from GERestaurant's test set.

Moreover, the 13 aspect categories used in the ASQP dataset by Zhang et al. (2021a) were adopted for GERest. Annotators revised the examples from GERestaurant to comprise quadruples including one of the 13 aspect categories instead of the five aspect categories considered for GERestaurant.

The annotation process for GERest followed the ASQP annotation guidelines established by Zhang et al. (2021c) and Wan et al. (2020) for Rest15 and Rest16.

All label revisions were initially performed by a computer science bachelor's student (Annotator *A*). Subsequently, all examples were reviewed and refined by a PhD student (Annotator *B*) with prior experience in annotating ABSA datasets.

Among the 2,124 annotated sentences, annotator *B* proposed an alternative label to that proposed by annotator *A* in the case of 184 sentences. Of these 184 proposed changes, 179 were accepted by annotator *A*. For the remaining 5 cases, a joint decision was made: in 3 instances, the original annotation by annotator *A* was retained, while in 2 cases, annotator *B*'s label was adopted.

B. Prompt Examples for English

```
### Instruction:
According to the following sentiment elements definition:

- The 'aspect category' refers to the category that aspect belongs to, and the available categories includes:
'ambience general', 'drinks prices', 'drinks quality', 'drinks style_options', 'food prices', 'food quality',
'food style_options', 'location general', 'restaurant general', 'restaurant miscellaneous', 'restaurant
prices', 'service general'.

Recognize all sentiment elements with their corresponding aspect categories in the following text with the
format of ['aspect category', ...].

### Text:
Good Food, Great Service, Average Prices (For the Strip)

### Label:
['food quality', 'service general', 'restaurant prices']
```

Figure 2: Prompt example for the ACD task for the English-language SemEval 2016 restaurant dataset.

```
### Instruction:
According to the following sentiment elements definition:

- The 'aspect category' refers to the category that aspect belongs to, and the available categories includes:
'ambience general', 'drinks prices', 'drinks quality', 'drinks style_options', 'food prices', 'food quality',
'food style_options', 'location general', 'restaurant general', 'restaurant miscellaneous', 'restaurant
prices', 'service general'.
- The 'sentiment polarity' refers to the degree of positivity, negativity or neutrality expressed in the
opinion towards a particular aspect or feature of a product or service, and the available polarities include:
'positive', 'negative' and 'neutral'.

Recognize all sentiment elements with their corresponding aspect categories and sentiment polarity in the
following text with the format of [('aspect category', 'sentiment polarity'), ...].

### Text:
Good Food, Great Service, Average Prices (For the Strip)

### Label:
[('food quality', 'positive'), ('service general', 'positive'), ('restaurant prices', 'neutral')]
```

Figure 3: Prompt example for the ACSA task for the English-language SemEval 2016 restaurant dataset.

```
### Instruction:
According to the following sentiment elements definition:

- The 'aspect term' is the exact word or phrase in the text that represents a specific feature, attribute, or
aspect of a product or service that a user may express an opinion about, the aspect term might be 'NULL' for
implicit aspect.
- The 'aspect category' refers to the category that aspect belongs to, and the available categories includes:
'ambience general', 'drinks prices', 'drinks quality', 'drinks style_options', 'food prices', 'food quality',
'food style_options', 'location general', 'restaurant general', 'restaurant miscellaneous', 'restaurant
prices', 'service general'.
- The 'sentiment polarity' refers to the degree of positivity, negativity or neutrality expressed in the
opinion towards a particular aspect or feature of a product or service, and the available polarities include:
'positive', 'negative' and 'neutral'.

Recognize all sentiment elements with their corresponding aspect terms, aspect categories and sentiment
polarity in the following text with the format of [('aspect term', 'aspect category', 'sentiment
polarity'), ...].

### Text:
Good Food, Great Service, Average Prices (For the Strip)

### Label:
[('Food', 'food quality', 'positive'), ('Service', 'service general', 'positive'), ('NULL', 'restaurant
prices', 'neutral')]
```

Figure 4: Prompt example for the TASD task for the English-language SemEval 2016 restaurant dataset.