German Aspect-based Sentiment Analysis in the Wild: B2B Dataset Creation and Cross-Domain Evaluation

Jakob Fehle[1], **Niklas Donhauser**[1], **Udo Kruschwitz**[2], **Nils Constantin Hellwig**[1], and **Christian Wolff**[1]

[1]Media Informatics, University of Regensburg, Germany,
[2]Information Science, University of Regensburg, Germany
Correspondence: Jakob.Fehle@ur.de

Abstract

Aspect-based sentiment analysis (ABSA) enables fine-grained sentiment extraction from user feedback but remains underexplored in many non-English languages and specialized application domains. In this study, we present insights from a multi-stage annotation of Business-to-Business (B2B) software reviews, highlighting key challenges such as domain-specific phrasing and implicit aspect terms. We document annotation practices and systematically benchmark state-of-the-art (SOTA) ABSA models on the three subtasks Aspect Category Detection (ACD), Aspect Category Sentiment Analysis (ACSA), and Target Aspect Sentiment Detection (TASD) using five German datasets. Results show that while simple classifiers remain strong baselines for category detection and fine-tuned Large Language Models (LLMs) excel in more structured tasks, performance varies notably across domains. Our findings emphasize that ABSA methods do not generalize uniformly, and that domain-sensitive annotation and evaluation strategies are essential for robust sentiment analysis.

1 Introduction

Aspect-based Sentiment Analysis (ABSA) is a core task in Natural Language Processing (NLP) that targets fine-grained sentiment classification by linking opinions to specific aspects mentioned in a text (Liu, 2022). This level of granularity enables applications to move beyond document-level sentiment and gain targeted insights, for example in product development, customer service, and social media analysis (Wankhade et al., 2022). By extracting structured opinion elements from unstructured text, ABSA helps organizations better understand user concerns and priorities.

In recent years, ABSA has gained traction in both research and industry, particularly for processing structured user feedback (Ligthart et al., 2021). While methodological progress, especially in English, has been driven by standardized benchmarks like SemEval 2016 (Pontiki et al., 2016), much research remains focused on narrow domains such as restaurants, hotels, or product reviews, limiting generalizability to more complex or diverse feedback settings (Hua et al., 2024).

At the same time, interest in applying NLP methods, including ABSA, to business-relevant contexts is growing. Recent work in information retrieval (IR) and applied NLP highlights the value of structured text analysis for enterprise use cases (Alonso and Baeza-Yates, 2024). While sentiment-bearing feedback plays a key role here, most ABSA research still targets consumer-oriented applications. Existing work in business and software domains (Alkalbani et al., 2016; Swillus and Zaidman, 2023; Rotovei and Negru, 2020) shows potential but often relies on simple polarity labels or app review analysis, offering only limited coverage of Business-to-Business (B2B) needs.

In contrast to English, ABSA resources for German remain limited. Domain-specific corpora exist, including hospitality (Hellwig et al., 2024) and public transportation (Wojatzki et al., 2017), and have driven research primarily within these domains (Chebolu et al., 2023a). However, exploring additional application areas, such as enterprise contexts, can help diversify evaluation environments and broaden understanding of model generalizability. In this paper, we investigate a new dataset consisting of B2B software feedback, a domain characterized by distinct linguistic patterns and specific user concerns.

We present an in-depth annotation study on user feedback from a B2B software application, representing a linguistically diverse and underexplored application context. While the dataset remains confidential, we include representative synthetic examples that closely reflect the content, structure, and

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 213–227, 2025

Sentence (de/en)	Triplets (Aspect Term, Category, Polarity)
die anwendung ist manchmal etwas langsam. (*The application is sometimes a little slow.*)	(anwendung, Technical Performance, Negative)
unkompliziert und übersichtlich (*uncomplicated and clear*)	(Null, Ease of Use, Positive), (Null, Interface Design, Positive)
Funktional auf das Wichtigste begrenzt. (*Functionally limited to the essentials.*)	(Null, Functional Scope, Neutral)

Table 1: Illustrative examples synthetically created to resemble the original, unpublished reviews.

style of the original data (Table 1). We also detail the annotation process, including quality assurance and lessons learned, to contribute to best practices in dataset creation.

Building on this resource, we systematically evaluate a range of state-of-the-art (SOTA) ABSA approaches across the three common subtasks, Aspect Category Detection (ACD), Aspect Category Sentiment Analysis (ACSA), and Target Aspect Sentiment Detection (TASD), and benchmark their performance not only on our B2B data but also across multiple established German-language datasets. Our results offer novel insights into annotation challenges, domain transferability, and the performance of modern ABSA approaches in enterprise-relevant settings.

2 Related Work

2.1 ABSA in Business and Software Contexts

While ABSA research has largely centered on domains like restaurants or consumer products, recent work increasingly explores business and software-related contexts, particularly in B2B environments (Davoodi et al., 2025; Rotovei and Negru, 2020). This reflects broader trends in applied NLP and IR, where methods like semantic enrichment and neural retrieval are used to structure enterprise data (Alonso and Baeza-Yates, 2024). Yet, most ABSA efforts still focus on consumer-facing software, offering only limited insights for enterprise use cases.

Early contributions by Atoum and Otoom (2016) and Alkalbani et al. (2016) addressed polarity classification in Software-as-a-Service reviews, while Rotovei and Negru (2020) applied ABSA to notes for customer relationship management in B2B contexts, integrating aspect-based insights into recommendation systems. Adjacent work on app and platform reviews, such as the AWARE corpus (Alturaief et al., 2021) and studies by Davoodi et al. (2025) and Alqaryouti et al. (2020), demonstrates the value of ABSA for structured feedback analysis, even if the reviewed platforms are not strictly enterprise-focused.

Further studies by Hegde and Seema S. (2017) and Swillus and Zaidman (2023) underscore the effectiveness of domain-adapted models and the richness of developer platforms as data sources. Beyond sentiment, related NLP techniques such as information extraction (Arslan and Cruz, 2024) and context-aware retrieval (Li et al., 2022) support enterprise applications by enabling structured access to unstructured business text.

Chebolu et al. (2023a) emphasize that most ABSA datasets still center on consumer domains, leaving enterprise-focused applications underrepresented. While those efforts demonstrate the feasibility and benefits of ABSA in software feedback, they fall short of capturing the full complexity of B2B scenarios.

2.2 Annotation Practices and Dataset Quality

Manual annotation remains a critical but resource-intensive step in building reliable NLP datasets, often considered a bottleneck in model development (Neves and Ševa, 2021). Activities such as schema definition, annotator training and experience, as well as quality control substantially influence the outcome.

Klie et al. (2024) conducted a large-scale analysis of annotation quality management practices in 591 NLP dataset papers. They outline four key dimensions: *stability* (consistency across annotators), *reproducibility* (consistent results under identical guidelines), *accuracy* (correctness and guideline adherence), and *unbiasedness* (absence of systematic bias). Their findings emphasize the importance of iterative refinement of guidelines and regular validation, such as expert reviews. Despite many positive examples, a significant number of studies still show deficits in quality assurance.

ABSA-specific challenges, such as implicit as-

pect mentions, irony, or sarcasm, can directly undermine these quality dimensions. While many datasets build on established annotation guidelines (e.g., from SemEval (Pontiki et al., 2014, 2016), GERestaurant (Hellwig et al., 2024), or Hotel Reviews (Fehle et al., 2023)), disagreements frequently arise from overlooked aspects or inconsistent span annotations. This underscores the ongoing importance of detailed instructions and systematic quality checks in ABSA annotation projects.

2.3 The State-of-the-Art for ABSA in English

Recent years have seen notable progress in ABSA across both classification- and generation-based approaches. For tasks like ACD and ACSA, robust baselines have been established using models such as BERT-CLF (Fehle et al., 2023; Hellwig et al., 2024), hierarchical architectures like Hier-GCN (Cai et al., 2020), and attention-enhanced variants like ECAN (Cui et al., 2024).

More complex generation tasks, including TASD and Aspect Sentiment Quadruple Prediction (ASQP), are typically addressed with sequence-to-sequence (seq-to-seq) models, often based on T5 (Raffel et al., 2020), that transform input into structured ABSA outputs. These include fixed-template approaches like Paraphrase (Zhang et al., 2021) and dynamic label ordering such as MvP (Gou et al., 2023) and DLO/ILO (Hu et al., 2022).

In parallel, instruction-tuned LLMs have shown promising results through fine-tuning (Šmíd et al., 2024), while few-shot prompting techniques (Hellwig et al., 2025; Simmering and Huoviala, 2023; Wang et al., 2023; Šmíd et al., 2024) increasingly demonstrate competitive performance in unsupervised or semi-supervised settings.

2.4 ABSA in German

Compared to English, ABSA in German remains considerably underrepresented. Although the availability of German datasets has improved, they are still rarely used or systematically evaluated.

Several corpora provide sentence-level ABSA annotations in specific domains, including *Hotel Reviews* (Fehle et al., 2023), *MobASA* (Gabryszak and Thomas, 2022), *Talk of Literature* (Greve et al., 2021), and *GERestaurant* (Hellwig et al., 2024), while *GermEval 2017* (Wojatzki et al., 2017) offers review-level labels. Earlier datasets like *SCARE* (Sänger et al., 2016), *Usage* (Klinger, 2014), and *MLSA* (Clematide et al., 2012) include sentiment or entity annotations but do not align with current ABSA definitions. Synthetic corpora such as *M-ABSA* (Wu et al., 2025) include German, but lack manually annotated ground truth.

Most evaluations remain tied to dataset releases and reuse English SOTA implementations. BERT-CLF (Fehle et al., 2023) serves as a common baseline for *Hotel Reviews* and *GERestaurant*, the latter also including an adaption of Paraphrase for German (Hellwig et al., 2024). For *GermEval*, systems range from BiLSTMs and LSTM-CRFs (Mishra et al., 2017; Lee et al., 2017) to CNN extractors (Schmitt et al., 2018) and rule-based approaches (Ruppert et al., 2017). More recent evaluations use transformer-based models such as BERT-CLF (Aßenmacher et al., 2021) and Pointer Networks (Wunderle et al., 2023).

However, many of these studies predate the rise of neural networks, transformers, and LLMs, which limits their performance and thus their relevance for current ABSA research. As modeling practices shift toward transformer-based and generative methods, there is a growing need for unified benchmarks and systematic re-evaluations using modern approaches.

3 Creation of a B2B Software Reviews Dataset

3.1 Data Collection

The dataset is based on proprietary user feedback collected from a B2B software platform. Users were able to voluntarily submit free-text comments during natural interactions with the application, typically in response to specific features or workflows.

To enable sentence-level ABSA, we sampled 1,500 reviews and segmented them into individual sentences using SpaCy (Honnibal et al., 2020), resulting in a total of 3,918 sentences.

3.2 Annotation Schema and Guidelines

The annotation guidelines[1] were adapted from GERestaurant (Hellwig et al., 2024) and SemEval 2016 (Pontiki et al., 2016), carrying over the general definitions, distinctions, and annotation principles for aspect categories and aspect terms, as well as the overall annotation objective. These were then tailored to the software feedback domain by refining category sets, clarifying domain-specific

[1] https://github.com/JakobFehle/German-ABSA-in-the-Wild/blob/main/Annotation_Guidelines_English.pdf

terminology, and extending examples to reflect typical B2B software user concerns. Annotations follow a triplet format **(aspect term, aspect category, polarity)**, with `NULL` used for implicit aspect terms. Sentences may contain multiple triplets, while all elements are required for validity.

We derived our aspect categories from a combination of prior ABSA research on software systems and user feedback (Rotovei and Negru, 2020; Guzman et al., 2017), established software quality frameworks such as ISO/IEC 25010 (International Organization for Standardization, 2023), and practical taxonomies employed by commercial review platforms such as G2 Crowd[2] and Capterra.[3] The resulting categories cover a broad spectrum of quality dimensions and user concerns, including technical attributes, user experience, and economic factors. Several categories align closely with standardized evaluation dimensions (e.g., usability or performance-related feedback), while others reflect terminology commonly found in real-world software reviews (e.g., comments on available features or pricing). To account for opinions not tied to a specific aspect, we also included a general-purpose category, as seen in previous shared tasks and annotation studies (Pontiki et al., 2014, 2016; Wojatzki et al., 2017).

3.3 Annotation Process

The annotation was conducted by two ABSA task experts (a PhD and a Master's student in Computer Science) in a multi-stage process with regular validation and review. Given the high annotation effort and limited resources, we opted against full double annotation with majority voting. Each annotator worked on separate portions of the dataset, followed by supervision / review through mutual validation to ensure consistency and quality. Discrepancies were discussed to resolve ambiguities, after which the existing annotations were reviewed for consistency across the dataset, in line with quality assurance practices recommended by Klie et al. (2024).

Annotation was carried out in two iterations. In the first iteration, besides annotating all samples in the dataset, the primary focus was on the refinement and expansion of aspect category definitions using examples from the dataset to improve clarity and separation, in alignment with annotation quality dimensions such as stability and accuracy (Klie et al., 2024). Each aspect was annotated with aspect term (if any) and a polarity label $p \in \{\text{positive, negative, neutral}\}$. Based on insights from extensive review discussions, which revealed recurring difficulties in annotating precise phrase boundaries, a second iteration was conducted to further improve the overall consistency of the dataset and, more specifically, to refine phrase boundary consistency for the 1,775 annotations with explicit aspect terms.

3.4 Quality Control and Annotator Agreement

During first iteration, 3,918 sentences were annotated; 1,075 instances (27.4 %) received objections or clarifications during review. Of these, 243 cases (6.2 %) required direct discussion between annotators to reach a final decision. The remaining comments were resolved through individual revision. Finally, the first iteration resulted in 37.1 % of the dataset annotated as containing no sentiment-bearing content, 5.9 % annotated as segmentation errors, 5.5 % annotated as ambiguous content, and 0.6 % annotated as grammatically invalid. During second iteration, 31.9 % of annotations were revised or flagged, though most could be resolved without further discussion, indicating growing consistency in interpretation.

To assess inter-annotator agreement (IAA), we double-annotated a shared subset of 100 samples and measured micro F1 scores, involving both ABSA task experts (annotators of this study) and domain experts responsible for maintaining and providing the B2B software platform from which the dataset originated, to enable a direct comparison. We chose F1 over Kappa as an IAA metric because F1 is more suitable for span extraction tasks such as opinion target phrase identification and for datasets where classification and extraction subtasks are combined (Chebolu et al., 2023b; Pontiki et al., 2016). In our case, exact phrase matching was required for targets and opinions, and IAA was computed by treating the annotations of one annotator as the gold standard and the other as the prediction.

Agreement decreased as more elements of the ABSA triplet were considered (see Table 2). For domain experts, agreement dropped from 60.80 % for aspects alone to 55.20 % when polarity was included, and to 35.20 % for full triplets with opinion target phrases. Task experts achieved consistently higher agreement across all levels (68.00 %,

[2] https://www.g2.com/
[3] https://www.capterra.com.de/

Evaluation Level	Domain Experts	Task Experts
Aspects	60.80	68.00
Aspects + Polarity	55.20	67.20
Full Triplet	35.20	57.76

Table 2: Inter-annotator agreement (micro-F1) for domain experts and ABSA task experts at different levels of the ABSA triplet.

67.20 %, and 57.76 %, respectively). This progressive drop, particularly for full triplets, underscores both the importance of clear, task-specific guidelines and the need for a dedicated, consistently annotated dataset for further analysis.

3.5 Impact of Re-annotation

To analyze the effect of our two-stage annotation process, we conducted an evaluation using SOTA ABSA models on standard subtasks (ACD, ACSA, and TASD).[4]

The second iteration aimed to improve the overall consistency of aspect category assignments and, more specifically, to refine the boundaries of opinion target phrases. Model evaluations showed slight performance gains in ACD (F1: 74.31 → 75.60), but strong improvement in ACSA (F1: 66.72 → 69.71) and TASD (F1: 47.51 → 50.27), indicating increased dataset consistency and more accurate phrase boundary handling.

3.6 Dataset Statistics

Aspect Category	Pos	Neu	Neg	Total
Functional Scope	402	13	577	992
Ease of Use	418	2	422	842
Interface Design	254	3	227	484
General Experience	254	34	184	472
Technical Performance	128	1	336	465
Customer Support	81	0	88	169
Pricing	2	1	52	55
Total	1,539	54	1,886	3,479

Table 3: Aspect-based sentiment distribution for the B2B software reviews dataset across aspect categories.

Table 3 summarizes the sentiment distribution across the seven aspect categories. The most frequently annotated aspects include functional scope (28.5 %), ease of use (24.2 %), and interface-related issues (13.9 %), reflecting a strong user focus on core functionality, usability, and visual layout. In contrast, customer support (4.9 %) and cost-related factors (1.6 %) were mentioned less frequently.

The dataset is skewed toward negative sentiment (54.2 %), followed by positive (44.2 %) and neutral statements (1.6 %), indicating that users more often express dissatisfaction, particularly with technical performance and pricing, both of which show high proportions of negative sentiment.

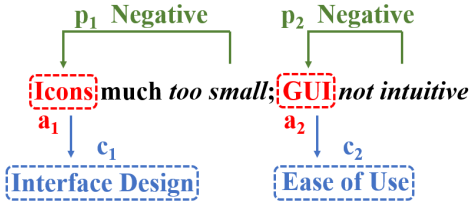
Around 32 % of all sentiment targets are implicit, i.e., the aspect term is not explicitly stated but inferred from context. This underscores the complexity of the annotation task and the importance of contextual understanding in ABSA for B2B software feedback.

The final dataset comprises 2,464 sentences and 3,479 annotated aspects. For model training and evaluation, the data was split 70 / 10 / 20 into training (1,707), development (249), and test (508) sets.

4 Experiments

4.1 Tasks

ABSA comprises several sub-tasks that focus on different components of sentiment expressed within a text. In this work, we address the three most common ABSA tasks supported by German datasets: Aspect Category Detection (ACD), Aspect Category Sentiment Analysis (ACSA), and Targeted Aspect Sentiment Detection (TASD).



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

Table 4: Illustration of the ABSA subtasks used in this study, each with the expected output of the respective task. Input is only the text sentence.

An example of the tasks including outputs can be seen in Table 4. ACD involves identifying all aspect categories mentioned or implied in a given text. ACSA extends this by additionally determining the sentiment polarity associated with each identified aspect. TASD further refines the analysis by not

[4]We used BERT-CLF for ACD and ACSA and Paraphrase for TASD. For more detailed descriptions, see 4.3.

Name	Annotation	Domain	Size			
			Train	Dev	Test	Total
Hotel Reviews (Fehle et al., 2023)	AC, SP	Hospitality (Hotels)	3,403	-	851	4,254
MobASA (Gabryszak and Thomas, 2022)	AT, AC, SP	Public Transportation	3,119	1,054	1,028	5,201
GERestaurant (Hellwig et al., 2024)	AT, AC, SP	Hospitality (Restaurants)	2,135	-	919	3,054
GermEval (Wojatzki et al., 2017)	AT, AC, SP	Public Transportation	16,200	1,917	3,642	21,759
B2B Software Reviews (Ours)	AT, AC, SP	Software Products	1,707	249	508	2,464

Table 5: Overview of the datasets used in this study. Abbr.: AT = Aspect Term, AC = Aspect Category, SP = Sentiment Polarity.

only classifying the aspect category and its sentiment polarity, but also extracting the specific phrase or expression in the text that serves as the sentiment target. Thus, TASD requires models to simultaneously perform classification and term extraction.

4.2 Datasets

To evaluate performance across domains, we compare results on our B2B software dataset with German-language customer reviews from hospitality and public transport domains (see Table 5; more details in Appendix A.1). *GermEval* (Wojatzki et al., 2017) covers public transport, focusing on customer feedback to Deutsche Bahn, comprising 21,772 sentences and 29,439 triplets over 19 categories. *MobASA* (Gabryszak and Thomas, 2022) contains 5,201 tweets on accessibility in public transport, with 5,927 annotations across 17 categories. *Hotel Reviews* (Fehle et al., 2023) consists of 4,254 hotel reviews from TripAdvisor with 5,617 aspect annotations across 5 categories, while *GERestaurant* (Hellwig et al., 2024) includes 3,078 restaurant reviews with 4,314 aspects across 6 categories.

4.3 Baseline Methods

To ensure comparability, we implemented a set of baseline methods based on recent SOTA approaches in English, most of which were originally developed for the widely used SemEval 2016 restaurant dataset (Hua et al., 2024). As these methods have rarely been evaluated on German data, we adopted the hyperparameters and model configurations from the original publications. Baselines were selected based on reported performance and the availability of reproducible code, ensuring reliable adaptation to our experimental setup.

BERT-CLF. A multi-label classification model based on BERT, following the approach of Fehle et al. (2023). The model predicts one or more labels per sentence, where each label corresponds to either an aspect category (e.g., `FOOD` in ACD) or an aspect–sentiment pair (e.g., `FOOD:POSITIVE` in ACSA).

Hier-GCN. Combines BERT sentence embeddings with hierarchical graph convolutional networks (GCNs) to model dependencies between aspects and sentiments (Cai et al., 2020).

Paraphrase. A seq-to-seq method using fixed natural language output templates to convert input into structured ABSA elements (Zhang et al., 2021).

MvP. Multi-View Prompting (MvP) generates sentiment tuples by augmenting the model with different permutations of the target elements. The outputs are aggregated via a majority voting strategy to better capture interdependencies (Gou et al., 2023).

Few-Shot Prompting. This approach extends prior work (Simmering and Huoviala, 2023; Šmíd et al., 2024) and uses in-context learning (ICL) with LLMs with up to 50 annotated ABSA examples embedded in prompts. The prompt template is adapted from Gou et al. (2023), translated into German and adjusted to match the specific structure of each ABSA subtask. Examples are depicted in Appendix A.2.

Instruction-based Fine-Tuning. A LLM is fine-tuned on a task-specific dataset to directly learn the mapping from input sentences to ABSA element outputs. Unlike prompting-based methods, this approach updates the model parameters during training, enabling more specialized behavior (cf. Šmíd et al., 2024). We use the same prompt template as in the few-shot setting, ensuring consistency in task formulation.

Since most ABSA approaches were developed for English, we adapt them to German. For classification tasks, we use *gbert-base*[5] (Chan et al., 2020), a monolingual German BERT. TASD models (Paraphrase, MvP) rely on the multilingual *t5-base*,[6] pretrained on English, German, and French. The LLM-based methods (Few-Shot Prompting, Instruction Fine-Tuning) use *LLaMA 3.1 8B*,[7] previously shown effective in English ABSA (Šmíd et al., 2024).

4.4 Evaluation Procedure

Each dataset was evaluated based on the available annotation depth. Similarly to previous studies, we only use main aspects for aspect category representation (Wojatzki et al., 2017). All datasets except *Hotel Reviews*, which lacks sentiment targets, were assessed on ACD, ACSA, and TASD. Applicable baseline methods were used per task, following reference implementations with minimal changes to model settings.

We tested different epoch counts based on dataset size and selected the best-performing run on the development set for test evaluation. For datasets without a dev split, 20 % of the training data was used for validation. Test evaluations were averaged over five fixed seeds (5, 10, 15, 20, and 25). For *GermEval*, both test sets were averaged. Details on configurations are provided in Appendix A.3.

As in previous studies (Pontiki et al., 2016; Wojatzki et al., 2017), we report micro-averaged F1 scores as the primary evaluation metric, and provide additional metrics such as macro-averaged F1-score, precision, recall and accuracy on GitHub.[8]

To assess the significance of method differences within each dataset ($p_{\text{adj}} \leq 0.05$), we apply parametric and non-parametric tests (e.g., ANOVA and paired t-tests (Field et al., 2012) or Friedman and-Wilcoxon tests (Wilcoxon, 1992)) with Bonferroni-Holm correction (Holm, 1979), based on normality-tested samples using the Shapiro-Wilk test (Shapiro and Wilk, 1965).

[5] deepset/gbert-base
[6] google-t5/t5-base
[7] meta-llama/Llama-3.1-8B
[8] https://github.com/JakobFehle/German-ABSA-in-the-Wild

5 Results and Discussion

5.1 Creation of a B2B Software Reviews Dataset

Our B2B Software Reviews dataset addresses a gap in ABSA resources for unstructured, domain-specific user feedback. The dataset was carefully curated through a multi-stage annotation process, resulting in high-quality sentence-level ABSA annotations across seven domain-relevant categories.

The annotation process revealed the inherent difficulty of the task: substantial amount of implicit aspects, sentiment skew toward negative feedback, and domain-specific phrasing patterns increased annotation complexity. Inter-annotator agreement results further underscore the need for task-specific training and iterative guideline development, confirming observations from large-scale analyses of annotation practices (Klie et al., 2024). These findings highlight the particular challenges posed by the B2B software domain for ABSA tasks.

5.2 Benchmarking Results

In contrast to earlier work in enterprise settings, which primarily relied on lexicon-based or traditional machine learning methods such as SVMs and rule-based pipelines (Atoum and Otoom, 2016; Alkalbani et al., 2016; Rotovei and Negru, 2020; Alqaryouti et al., 2020), our evaluation systematically explores the performance of modern transformer-based architectures, including fine-tuned LLMs.

TASD emerged as the most challenging task (see Table 6). Fine-tuned LLMs achieved the best results (F1: 55.77), outperforming Paraphrase and MvP approaches. Extraction accuracy varied considerably across categories: interface-related phrases were detected most reliably, likely due to their linguistic consistency, while terms related to the applications functionality were frequently misclassified, mirroring higher annotator uncertainty

Target Aspect Sentiment Detection (TASD)

Method	MobASA	Rest	GermEval	B2B
Paraphrase	78.69	65.72	54.03	50.27
MvP	79.65	67.00	**55.75**	50.50
LLaMA Few-Shot	64.62	61.13	43.78	42.34
LLaMA Fine-Tune	**81.56**	**73.22**	31.06	**55.77**

Table 6: Micro-F1 scores as averages over five seeds for TASD across datasets. Highest values are bold, significant differences are underscored. Abbr.: Rest = GERestaurant, B2B = B2b Software Reviews.

Method	Hotel	MobASA	Rest	GermEval	B2B
Aspect Category Detection (ACD)					
BERT-CLF	**89.06**	**94.07**	**91.09**	**78.10**	**75.60**
LLaMA Few-Shot	79.09	79.70	83.68	46.51	66.98
LLaMA Fine-Tune	87.69	92.18	88.06	41.27	74.22
Aspect Category Sentiment Analysis (ACSA)					
BERT-CLF	78.75	83.57	84.34	65.83	69.71
Hier-GCN	78.02	84.82	83.31	**67.87**	**69.80**
LLaMA Few-Shot	74.80	70.29	80.79	39.36	64.78
LLaMA Fine-Tune	**80.51**	**87.22**	**85.22**	33.41	69.13

Table 7: Micro-F1 scores as averages over five seeds for ACD and ACSA across datasets. Highest values are bold, significant differences are underscored. Abbr.: Hotel = Hotel Reviews, Rest = GERestaurant, B2B = B2b Software Reviews.

during labeling. Neutral sentiment was hardest to detect, while positive phrases were classified more reliably than negative ones, despite being less frequent, suggesting that clarity may outweigh frequency in model performance.

By contrast, classification-based tasks (ACD and ACSA) were less sensitive to these issues (see Table 7). For ACD, BERT-CLF performed best (F1: 75.60), confirming the strength of simple multi-label classification. While technical and general feedback categories posed challenges, feature-related aspects were more reliably detected, indicating that category classification is more robust than phrase extraction. In ACSA, fine-tuned LLMs and Hier-GCN slightly outperformed BERT-CLF. Positive sentiment was again predicted more accurately than negative or neutral sentiment.

5.3 General Performance Comparison across Datasets

While some methods have previously been evaluated on individual datasets, we re-implemented their approaches to ensure consistency and comparability across tasks and domains. BERT-CLF follows prior work by Fehle et al. (2023), Aßenmacher et al. (2021), and Hellwig et al. (2024), while Paraphrase corresponds to Hellwig et al. (2024). For *GermEval*, we apply a standard F1 evaluation instead of the shared task script. Thus, as discussed by Wunderle et al. (2023), previously reported results on this dataset are not directly comparable, and are therefore not considered here.

5.3.1 Performance on ACD and ACSA

The results reflect strong influences of the number of aspect categories, label imbalance, and domain-specific variability.

Datasets with few categories, such as *Hotel Reviews* and *GERestaurant*, yield consistently high and balanced scores (micro ≈ macro F1). In contrast, *MobASA* and *GermEval* show substantial macro F1 drops, even with BERT-CLF, falling to 74.36 and 39.54 respectively. Despite its limited class count, *B2B Software Reviews* yields lower overall scores, likely due to higher linguistic variability.

On ACD, BERT-CLF performs best overall, confirming its strength in simpler multi-label classification. In ACSA, fine-tuned LLMs and Hier-GCN outperform BERT-CLF by better modeling aspect-sentiment relations. However, performance drops notably on *GermEval*, where fine-tuned LLMs struggle with domain-specific noise.

Error Analysis The shift from ACD to ACSA revealed considerable performance drops in datasets with high class counts or skewed distributions, a trend also observed in previous studies (Fehle et al., 2023; Hellwig et al., 2024). On *GermEval*, a specific issue emerged with fine-tuned LLMs: instead of structured ABSA tuples, the model often returned shortcuts like single digits, primarily for the overrepresented `("general", "negative")` class (40 % of the training data). This suggests a form of shortcut learning driven by class dominance and noise. While similar behavior has been noted in few-shot prompting (Tang et al., 2023; Du et al., 2024), it remains under-researched in fine-tuning scenarios.

5.3.2 Performance on TASD

TASD proves to be the most challenging task, requiring models to jointly extract aspect terms and classify their categories and sentiments. Overall scores are notably lower than in ACD and ACSA, reflecting the increased complexity of structured prediction.

Results on *MobASA* and *GERestaurant* are relatively stable, whereas *GermEval* and *B2B Software Reviews* show substantially weaker performance across all methods. Fine-tuned LLMs achieve the best results overall, usually performing better than other seq-to-seq approaches and few-shot prompting. Among the T5-based methods, MvP provides only marginal improvements over Paraphrase despite incurring significantly longer training times and higher memory usage, raising concerns about its efficiency. Few-shot prompting performs worst, especially on *GermEval* (F1: 43.78) and *B2B Soft-*

ware Reviews (F1: 42.34). A full comparison of resource consumption is provided in Appendix A.4.

Error Analysis While fine-tuned LLMs generally perform best in TASD, recurring error patterns persist and often intensify. For the *GermEval* dataset and similar to ACSA, the model frequently fails to produce correctly structured ABSA outputs for the dominant `("general", "negative")` class, especially if combined with implicit aspect phrases. Instead of complete aspect-sentiment-term triples, the model often outputs incomplete results, such as "0". As discussed in Section 5.3.1, this behavior appears linked to shortcut learning, driven by class imbalance and noisy inputs. Moreover, *GermEval* is review-level and features longer, noisier inputs, which are frequently mapped to a single, overrepresented label. This increases the risk of models collapsing their output to template-like patterns. The tendency to simplify structured predictions under such conditions reveals a broader limitation of fine-tuned LLMs in imbalanced and noisy ABSA scenarios. In contrast, *MobASA* shows minimal TASD performance drops. This likely results from low linguistic variability: categories like "Lift" or "Escalator" map consistently to single terms (e.g., "Aufzug"), simplifying extraction.

5.4 Summary

We report insights from the multi-stage annotation and evaluation of a B2B ABSA dataset, analyzed alongside four existing German resources across common subtasks. While different methods prove most effective per task, e.g., simple classifiers for ACD and fine-tuned LLMs for TASD, their relative strengths are consistent across domains. Still, performance varies significantly due to class imbalance, linguistic variability, and domain-specific phrasing. These findings highlight that strong models require high-quality, task-specific annotation, and that broader, domain-sensitive evaluations are essential, not only to assess cross-domain robustness, but to account for the fact that ABSA methods do not generalize uniformly and face domain-specific challenges that can substantially impact their effectiveness.

6 Conclusion

This paper presents a systematic evaluation of ABSA methods across five German-language datasets, including a newly developed corpus of B2B software reviews. We benchmarked SOTA approaches for the ACD, ACSA, and TASD tasks, demonstrating that domain characteristics, class imbalance, and annotation quality strongly influence model performance. While simpler classification tasks such as ACD still benefit from lightweight BERT-based architectures, more complex tasks (ACSA and especially TASD) show clear advantages for fine-tuned LLMs, provided that output structure and domain variability are well managed.

Beyond model benchmarking, our multi-stage annotation process revealed key insights into the challenges of domain-specific ABSA: the difficulty of handling implicit aspects, the impact of phrasing variability, and the importance of clear, iterative guidelines. These findings underscore the central role of annotation practices and data design in enabling robust model evaluation.

Future work may explore improved decoding strategies for structured LLM outputs (Beurer-Kellner et al., 2024), sampling strategies for label balancing (Henning et al., 2023), or new approaches specifically tailored to the difficulties of B2B ABSA. More broadly, our results emphasize the value of high-quality annotation and domain-sensitive evaluation as important steps toward robust sentiment analysis, especially in business-critical applications.

Limitations

This work is subject to several limitations. First, due to confidentiality constraints, the B2B software dataset cannot be released publicly. While detailed documentation and comparative evaluation are provided, full reproducibility on the same data is not possible. Second, the annotations focus on a specific domain and may not generalize across all enterprise contexts. Additionally, although strong baseline and LLM models were evaluated, more advanced domain-adaptation methods were beyond the scope of this study.

Ethical Consideration

All data used in this study was collected as part of a partnership with a B2B software provider and consisted solely of anonymized, voluntarily submitted user feedback. No personal or sensitive information was included in the annotated corpus. The annotation process followed ethical guidelines for data protection and was reviewed to avoid potential harms such as exposing identifiable entities or commercially sensitive content. The aim of this re-

search is to improve methodological understanding of ABSA in enterprise contexts, not to evaluate or critique specific customers, products, or organizations.

References

Asma Musabah Alkalbani, Ahmed Mohamed Ghamry, Farookh Khadeer Hussain, and Omar Khadeer Hussain. 2016. Sentiment analysis and classification for software as a service reviews. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pages 53–58. IEEE.

Omar Alonso and Ricardo Baeza-Yates, editors. 2024. *Information Retrieval: Advanced Topics and Techniques*, 1 edition, volume 60. Association for Computing Machinery, New York, NY, USA.

Omar Alqaryouti, Nur Siyam, Azza Abdel Monem, and Khaled Shaalan. 2020. Aspect-based sentiment analysis using smart government review data. *Appl. Comput. Inform.*, ahead-of-print(ahead-of-print).

Nouf Alturaief, Hamoud Aljamaan, and Malak Baslyman. 2021. AWARE: Aspect-based sentiment analysis dataset of apps reviews for requirements elicitation. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, pages 211–218. IEEE.

Muhammad Arslan and Christophe Cruz. 2024. Business-RAG: Information extraction for business insights. In *Proceedings of the 21st International Conference on Smart Business Technologies*, pages 88–94. SCITEPRESS - Science and Technology Publications.

Issa Atoum and Ahmed Otoom. 2016. Mining software quality from software reviews: Research trends and open issues. *Int. j. comput. trends technol.*, 31(2):74–83.

M. Aßenmacher, A. Corvonato, and C. Heumann. 2021. Re-Evaluating GermEval17 Using German Pre-Trained Language Models. *arXiv preprint*. ArXiv:2102.12330 [cs].

Luca Beurer-Kellner, Marc Fischer, and Martin T Vechev. 2024. Guiding LLMs the right way: Fast, non-invasive constrained generation. *ICML*, abs/2403.06988.

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.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *Int Conf Comput Linguistics*, pages 6788–6796.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2023a. 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. 2023b. 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)*, pages 611–628.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. 2012. MLSA - a multi-layered reference corpus for german sentiment analysis.

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.

Laleh Davoodi, József Mezei, and Markku Heikkilä. 2025. Aspect-based sentiment classification of user reviews to understand customer satisfaction of e-commerce platforms. *Electron. Commer. Res.*

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2024. Shortcut learning of large language models in natural language understanding. *Commun. ACM*, 67(1):110–120.

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 Lingustics.

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

Aleksandra Gabryszak and Philippe Thomas. 2022. MobASA: Corpus for aspect-based sentiment analysis and social inclusion in the mobility domain. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 35–39, Marseille, France. European Language Resources Association.

Zhibin Gou, Qi Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. *Annual Meeting of the Association for Computational Linguistics*, pages 4380–4397.

Lore De Greve, Pranaydeep Singh, Cynthia Van Hee, Els Lefever, and Gunther Martens. 2021. Aspect-based Sentiment Analysis for German: Analyzing "Talk of Literature" Surrounding Literary Prizes on Social Media. *Computational Linguistics in the Netherlands Journal*, 11:85–104.

Emitza Guzman, Rana Alkadhi, and Norbert Seyff. 2017. An exploratory study of twitter messages about software applications. *Requir. Eng.*, 22(3):387–412.

Rajalaxmi Hegde and Seema S. 2017. Aspect based feature extraction and sentiment classification of review data sets using incremental machine learning algorithm. In *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pages 122–125. IEEE.

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. 2025. 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.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving Aspect Sentiment Quad Prediction via Template-Order Data Augmentation. *arXiv preprint*. ArXiv:2210.10291 [cs].

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).

International Organization for Standardization. 2023. ISO/IEC 25010:2023 – Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Product quality model. Available at: `https://www.iso.org/standard/78176.html`.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Comput. Linguist. Assoc. Comput. Linguist.*, pages 1–50.

Roman Klinger. 2014. The USAGE review corpus for fine-grained, multi-lingual opinion analysis.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Ji-Ung Lee, Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. UKP TU-DA at GermEval 2017: Deep learning for aspect based sentiment detection. *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 22–29.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv [cs.CL]*.

Alexander Ligthart, Cagatay Catal, and Bedir Tekinerdogan. 2021. Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, 54(7):4997–5053.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Pruthwik Mishra, Vandan Mujadia, and Soujanya Lanka. 2017. Germeval 2017: sequence based models for customer feedback analysis. *Proceedings of the GermEval*, pages 36–42.

Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 22(1):146–163.

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.

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

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.

Doru Rotovei and Viorel Negru. 2020. Multi-agent recommendation and aspect level sentiment analysis in B2B CRM systems. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 238–245. IEEE.

Eugen Ruppert, Abhishek Kumar, and Chris Biemann. 2017. LT-ABSA: An extensible open-source system for document-level and aspect-based sentiment analysis. *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 55–60.

Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint Aspect and Polarity Classification for Aspect-based Sentiment Analysis with End-to-End Neural Networks.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

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

Mark Swillus and Andy Zaidman. 2023. Sentiment overflow in the testing stack: Analyzing software testing posts on stack overflow. *J. Syst. Softw.*, 205(111804):111804.

Mario Sänger, Ulf Leser, Steffen Kemmerer, Peter Adolphs, and Roman Klinger. 2016. SCARE - the sentiment corpus of app reviews with fine-grained annotations in German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1114–1121, Portorož, Slovenia. European Language Resources Association (ELRA).

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zengzhi Wang, Qiming Xie, and Rui Xia. 2023. A Simple yet Effective Framework for Few-Shot Aspect-Based Sentiment Analysis. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 1765–1770, New York, NY, USA. Association for Computing Machinery.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial intelligence review*, 55(7):5731–5780.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In Samuel Kotz and Norman L Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer New York, New York, NY.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*, pages 1–12.

Chengyan Wu, Bolei Ma, Yihong Liu, Zheyu Zhang, Ningyuan Deng, Yanshu Li, Baolan Chen, Yi Zhang, Barbara Plank, and Yun Xue. 2025. M-ABSA: A Multilingual Dataset for Aspect-Based Sentiment Analysis. *arXiv preprint*. ArXiv:2502.11824 [cs].

Julia Wunderle, Jan Pfister, and Andreas Hotho. 2023. Pointer Networks: A Unified Approach to Extracting German Opinions.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. 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.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2024. 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.

A Appendix

A.1 Additional Dataset Descriptions

A.1.1 Hotel Reviews

The *Hotel Reviews* dataset (Fehle et al., 2023) contains customer reviews of hotels collected from Tripadvisor, annotated for aspect-based sentiment at the sentence level. Each annotation includes a main aspect, a subcategory, and a sentiment polarity. Annotation was performed by student annotators based on majority decisions, with expert oversight for curation and supervision (Fehle et al., 2023). As no predefined data splits are available, we perform an 80 / 20 random split into training and test sets. The dataset comprises 4,254 sentences and 5,617 annotated aspects.

A.1.2 MobASA

The Mobility Aspect-based Sentiment Analysis (*MobASA*) dataset (Gabryszak and Thomas, 2022) consists of German-language tweets related to public transportation, annotated for relevance and aspect-based sentiment with a focus on aspects of barrier-free travel. Aspect categories are divided into main aspects and subcategories and reflect issues relevant to passengers with limited mobility due to disability, age, or traveling with young children. Annotations are based on majority decision of two annotators and include relevance, document-level sentiment, and aspect-based sentiment, structured along aspect categories, sentiment polarity and explicit sentiment targets (Gabryszak and Thomas, 2022). Multiple dataset versions exist based on annotator expertise (experts, crowdsourcing, or mixed). For this work, we use the expert-annotated version, comprising 5,927 aspect annotations across 5,201 sentences from the predefined train, development, and test splits.

A.1.3 GERestaurant

The *GERestaurant* dataset (Hellwig et al., 2024) contains German-language restaurant reviews from TripAdvisor and follows an annotation scheme similar to the SemEval 2016 restaurant dataset, but with an adapted set of aspect categories. Annotations include aspect categories, sentiment polarities, and sentiment targets at the sentence level, with both explicit and implicit aspect terms being considered. The annotation was conducted by an experienced computer science student and subsequently validated by an expert annotator (Hellwig et al., 2024). We use the predefined train / test splits, which include 3,078 sentences and 4,314 aspect annotations.

A.1.4 GermEval 2017

The *GermEval 2017* dataset (Wojatzki et al., 2017) was developed for the GermEval Shared Task on ABSA and comprises German customer feedback related to "Deutsche Bahn", the national railway operator. Annotations include relevance, document-level sentiment, and aspect-based sentiment, consisting of aspect category, sentiment polarity, and sentiment target. Aspect categories are structured into main categories and subcategories, covering a wide range of service- and travel-related topics. Both explicit and implicit aspects were annotated. Annotations were generated through majority decisions by trained student annotators under expert supervision (Wojatzki et al., 2017). The dataset provides predefined splits for training, development, and testing, including two distinct test sets: one acquired synchronously with the training data, and another collected several months later to allow for diachronic evaluation. We utilize the complete dataset, including both test sets, totaling 21,759 sentences and 29,439 aspect annotations across all splits.

A.2 Prompts Examples for Few-Shot and Fine-Tuned LLMs

```
Entsprechend der folgenden Definitionen der
Sentiment-Elemente:

Die "Aspektkategorie" bezeichnet die Kategorie,
zu der der Aspekt gehört. Die verfügbaren
Kategorien sind: ['food', 'ambience',
'service', 'price', 'general'].

Bitte befolge die Anweisungen sorgfältig.
Stelle sicher, dass Aspektkategorien aus den
angegebenen Kategorien stammen.

Erkenne alle Aspekte mit ihrer jeweiligen
Aspektkategorien im folgenden Text im Format
['Aspektkategorie', ...]:

Text: Der Burger total durchgebraten und
trocken.
Sentiment-Elemente: ['food']
```

Listing 1: Prompt example for the instruction prompt of the ACD task for the GERestaurant dataset. Few shot examples are provided in front of the task sentence.

```
Entsprechend der folgenden Definitionen der
Sentiment-Elemente:

Die "Aspektkategorie" bezeichnet die Kategorie,
zu der der Aspekt gehört. Die verfügbaren
Kategorien sind: ['food', 'ambience',
'service', 'price', 'general'].

Die 'Sentiment-Polarität' beschreibt den Grad
der Positivität, Negativität oder Neutralität,
```

```
die in der Meinung zu einem bestimmten Aspekt
oder Merkmal eines Produktes oder einer
Dienstleistung ausgedrückt wird. Die
verfügbaren Polaritäten sind: 'positive',
'negative' und 'neutral'.

Bitte befolge die Anweisungen sorgfältig.
Stelle sicher, dass Aspektkategorien aus den
angegebenen Kategorien stammen. Stelle sicher,
dass die Polaritäten aus den verfügbaren
Polaritäten stammen.

Erkenne alle Sentiment-Elemente mit ihren
jeweiligen Aspektkategorien und Sentiment-
Polaritäten im folgenden Text im Format
[('Aspektkategorie', 'Sentiment-Polarität'),
...]

Text: Der Burger total durchgebraten und
trocken.
Sentiment-Elemente: [('food', 'negative')]
```

Listing 2: Prompt example for the instruction prompt of the ACSA task for the GERestaurant dataset. Few shot examples are provided in front of the task sentence.

```
Entsprechend der folgenden Definitionen der
Sentiment-Elemente:

Der "Aspektbegriff" bezieht sich auf ein
bestimmtes Merkmal, eine Eigenschaft oder einen
Aspekt eines Produktes oder einer
Dienstleistung, zu dem eine Person eine Meinung
äußern kann. Explizite Aspektbegriffe kommen
explizit als Teilzeichenkette im gegebenen Text
vor. Der Aspektbegriff kann "null" sein, wenn
es sich um einen impliziten Aspekt handelt.

Die "Aspektkategorie" bezeichnet die Kategorie,
zu der der Aspekt gehört. Die verfügbaren
Kategorien sind: ['food', 'ambience',
'service', 'price', 'general'].

Die 'Sentiment-Polarität' beschreibt den Grad
der Positivität, Negativität oder Neutralität,
die in der Meinung zu einem bestimmten Aspekt
oder Merkmal eines Produktes oder einer
Dienstleistung ausgedrückt wird. Die
verfügbaren Polaritäten sind: 'positive',
'negative' und 'neutral'.

Bitte befolge die Anweisungen sorgfältig.
Stelle sicher, dass Aspektbegriffe exakt mit
dem Text übereinstimmen oder "null" sind, wenn
sie implizit sind. Stelle sicher,
dass Aspektkategorien aus den angegebenen
Kategorien stammen. Stelle sicher, dass die
Polaritäten aus den verfügbaren Polaritäten
stammen.

Erkenne alle Sentiment-Elemente mit ihren
jeweiligen Aspektbegriffen, Aspektkategorien
und Sentiment-Polaritäten im folgenden Text im
Format
[('Aspektbegriff', 'Aspektkategorie',
'Sentiment-Polarität'), ...]

Text: Der Burger total durchgebraten und
trocken.
Sentiment-Elemente: [('Burger', 'food',
'negative')]
```

Listing 3: Prompt example for the instruction prompt of the TASD task for the GERestaurant dataset. Few shot examples are provided in front of the task sentence.

```
According to the following sentiment elements
definition:

The 'aspect term' refers to  specific feature,
attribute, or aspect of 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:
['food', 'ambiance', 'service', 'price',
'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 includes: '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: Der Burger total durchgebraten und
trocken.
Sentiment-Elemente: [('Burger', 'food',
'negative')]
```

Listing 4: Prompt example for the instruction prompt of the TASD task for the GERestaurant dataset in the English language. Few shot examples are provided in front of the task sentence.

A.3 Model Configurations and Hyperparameter Tuning

We provide an overview of the hyperparameter settings and model configurations used for the baseline experiments. Settings are based on the implementations specified by Hellwig et al. (2024) (BERT-CLF), Cai et al. (2020) (Hier-GCN), Zhang et al. (2021) (Paraphrase), Gou et al. (2023) (MvP), and Šmíd et al. (2024) (LLaMA Fine-Tune). Configuration for Few-Shot Prompting is based on previous experiments by Simmering and Huoviala (2023) and Šmíd et al. (2024).

BERT-CLF and Hier-GCN. Both approaches use `gbert-base` as a baseline model. For BERT-CLF, we set the learning rate to 2×10^{-5} and a batch size of 16. Hier-GCN is trained with a learning rate of 5×10^{-5} and a batch size of 8. For both models, the number of training epochs was validated on development data, with values ranging from e $\in$ [2, 3, 4, 5, 6] for the *Transport* dataset and e $\in$ [20, 25, 30, 35, 40] for all other datasets.

Paraphrase and MvP. For both approaches we employ `t5-base` as the base model with a batch size of 16. For Paraphrase, the learning rate is set to 3×10^{-4}; for MvP, we use 1×10^{-4}. Epochs

were validated over e ∈ [2, 3, 4, 5, 6] for *Transport*, and e ∈ [15, 20, 25, 30] for the remaining datasets.

LLaMA Fine-Tuning. We fine-tune the `LLaMA-3.1-8B` base model using LoRA with rank 64 and α scaling factor 16. The model is trained with a context length of 2,096 tokens, learning rate 2×10^{-4}, and batch size 16. Training is performed in 4-bit quantization, while inference uses full precision. Generation is performed using greedy decoding with temperature set to 0. The number of training epochs is validated with e ∈ [2, 3, 4] for *Transport* and e ∈ [4, 5, 6, 7] for all other datasets.

LLaMA Few-Shot. We use the `LLaMA-3.1-8B` Instruct model with a maximum context length of 8,192 tokens. Inference is conducted via greedy decoding (temperature = 0). For each dataset, we evaluate performance using n ∈ [10, 25, 50] few-shot examples for in-context learning. Text length for few shot examples is limited to 500 characters per example to prevent exceeding the context window.

A.4 Resource Requirements and Efficiency Analysis

Table 8 summarizes training times and GPU memory usage across models, tasks, and datasets.

BERT-CLF is highly efficient, requiring under 10 minutes and < 3.5 GB GPU across all tasks, making it a strong baseline for computing resource-constrained settings. Hier-GCN introduces moderate overhead (10 to 36 min, 3.3 GB) while offering slight gains in ACSA performance. LLaMA Fine-Tune achieves the best results but comes at higher computational cost, highlighting a clear trade-off between performance and efficiency. Paraphrase is a lightweight seq-to-seq baseline (13 to 14 min, 8 to 10 GB), while MvP substantially increases training time (up to 2 h) without notable performance gains, raising concerns about cost-effectiveness.

LLaMA Few-Shot was evaluated with *vllm* (Kwon et al., 2023), whereby the GPU utilisation can be customised in favour of the inference speed. Therefore, the values are not listed in the table.

Task	Model	Hotel Reviews		MobASA		GERestaurant		GermEval		B2B Software	
		Time	**GPU**	**Time**	**GPU**	**Time**	**GPU**	**Time**	**GPU**	**Time**	**GPU**
ACD	BERT-CLF	5m	2.7 GB	5m	2.5 GB	3m	2.4 GB	8m	3.2 GB	3m	2.4 GB
	LLaMA Fine-Tune	33m	8.7 GB	1h 17m	8.8 GB	37m	8.4 GB	2h 57m	10.2 GB	21m	8.5 GB
ACSA	BERT-CLF	6m	2.7 GB	4m	2.5 GB	5m	2.4 GB	6m	3.3 GB	4m	2.4 GB
	Hier-GCN	36m	3.2 GB	19m	3.3 GB	23m	3.3 GB	20m	3.3 GB	10m	3.3 GB
	LLaMA Fine-Tune	1h 07m	9.5 GB	1h 01m	9.6 GB	42m	9.3 GB	4h 09m	11.3 GB	27m	9.3 GB
TASD	Paraphrase	–	–	14m	7.5 GB	13m	10.5 GB	1h 02m	15.5 GB	13m	8.6 GB
	MvP	–	–	1h 59m	11.8 GB	45m	11.8 GB	1h 59m	11.8 GB	45m	11.8 GB
	LLaMA Fine-Tune	–	–	2h 07m	10.6 GB	1h 14m	10.6 GB	5h 50m	12.8 GB	50m	10.6 GB

Table 8: Training times and GPU memory usage per model, task, and dataset during test evaluation as an average over five seeds.