Annotation Quality in Aspect-Based Sentiment Analysis: A Case Study Comparing Experts, Students, Crowdworkers, and Large Language Models

Niklas Donhauser[1] **Jakob Fehle**[1] **Nils Constantin Hellwig**[1]
Markus Weinberger[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
{niklas.donhauser, jakob.fehle, nils-constantin.hellwig,
markus.weinberger, udo.kruschwitz, christian.wolff}@ur.de

Abstract

Aspect-Based Sentiment Analysis (ABSA) enables fine-grained opinion analysis by identifying sentiments toward specific aspects or targets within a text. While ABSA has been widely studied for English, research on other languages such as German remains limited, largely due to the lack of high-quality annotated datasets. This paper examines how different annotation sources influence the development of German ABSA. To this end, an existing dataset is re-annotated by experts to establish a ground truth, which serves as a reference for evaluating annotations produced by students, crowdworkers, Large Language Models (LLMs), and experts. Annotation quality is compared using Inter-Annotator Agreement (IAA) and its impact on downstream model performance for different ABSA subtasks. The evaluation focuses on Aspect Category Sentiment Analysis (ACSA) and Target Aspect Sentiment Detection (TASD). We apply State-of-the-Art (SOTA) methods for ABSA, including BERT-, T5-, and LLaMA-based approaches to assess performance differences, spanning fine-tuning and in-context learning with instruction prompts. The findings provide practical insights into trade-offs between annotation reliability, and efficiency, offering guidance for dataset construction in under-resourced Natural Language Processing (NLP) scenarios.

Keywords: Aspect-Based Sentiment Analysis, Large Language Models, Restaurant Reviews, Annotation Quality, Dataset Annotation

1. Introduction

Aspect-Based Sentiment Analysis (ABSA) is a subfield of Natural Language Processing (NLP) concerned with identifying sentiment expressed toward specific aspects or attributes mentioned in text. With the rapid growth of online content, user-generated data has become a major source for understanding public opinion across a wide range of domains (Liu, 2022). ABSA has been applied to product and restaurant reviews, movie critiques, political discourse, educational initiatives, social events, and market campaigns, as well as government policy analysis (Hua et al., 2024). It has also been used in social media contexts such as YouTube video ranking and in the economic domain, where aspect-level models are applied to microblogs and news articles (Chauhan et al., 2023).

Established benchmark corpora have played a central role in shaping ABSA research. The SemEval shared tasks from 2014 to 2016 defined standard datasets and evaluation settings that have strongly influenced subsequent work (Pontiki et al., 2014, 2015, 2016; Chebolu et al., 2023). Within these benchmarks, the restaurant domain emerged as one of the most prominent and widely reused settings, and has since become a standard benchmark setting for ABSA across multiple languages. Beyond its methodological importance, the restaurant domain also carries clear practical relevance, as aspect-level sentiment information can support applications such as recommendation systems and customer feedback analysis (Ara et al., 2020; Singhi et al., 2024).

Despite the potential of ABSA, its success strongly depends on the availability of annotated training data. However, ABSA is an under-resourced task for many languages, including German: datasets are scarce, and existing resources are often limited in size, domain coverage, or annotation quality (Fehle et al., 2023; Hellwig et al., 2024). Constructing high-quality datasets requires careful annotation, but this process is costly, time-intensive, and prone to inconsistencies (Klie et al., 2024; Monarch, 2021; Orr and Crawford, 2024; Dobnik and Kelleher, 2023). The challenge is amplified by the fact that different annotation strategies, such as crowdsourcing (Nowak and Rüger, 2010; He et al., 2024), student annotators (Fehle et al., 2023), expert annotators (Fehle et al., 2025; Barbarestani et al., 2024), or the use of Large Language Models (LLMs) (Ostyakova et al., 2023; Hellwig et al., 2025; Mæhlum et al., 2024) can produce

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 73–88 May 11, 2026.

datasets of varying reliability and utility for machine learning models.

This raises the question of how annotation strategies influence downstream ABSA performance and whether higher annotation quality justifies increased effort. To address this, we conduct a systematic comparison of four annotator groups, (1) crowdworkers, (2) students, (3) LLMs, and (4) task experts, in the German restaurant review domain. We evaluate the resulting datasets on two central ABSA subtasks, Aspect Category Sentiment Analysis (ACSA) and Target Aspect Sentiment Detection (TASD) (Fehle et al., 2025; Hellwig et al., 2025; Bu et al., 2021; Wu et al., 2025), and complement model-based evaluation with Inter-Annotator Agreement (IAA) to assess annotation consistency.

In addition to evaluating model performance, the study also analyzes IAA as a complementary measure to assess annotation consistency across different annotator groups. To comprehensively assess the influence of annotation quality, a range of State-of-the-Art (SOTA) approaches are applied and compared. These include traditional classifier-based models such as BERT-CLF (Fehle et al., 2023), HIER-GCN (Cai et al., 2020), as well as more recent text generation and LLM techniques, including Paraphrase (Zhang et al., 2021), Multi-View Prompting (Gou et al., 2023), or LLaMA (Dubey et al., 2024) with fine-tuning and Gemma (Team et al., 2025) with few-shot prompting.

In summary, this paper presents a systematic annotation study on German restaurant reviews, comparing four annotation strategies: crowdworkers, students, LLMs, and experts, and analyzes their impact on ABSA dataset quality and downstream model performance for ACSA and TASD. The study derives practical recommendations for dataset construction and provides empirical insights into how annotation quality affects ABSA models. To support reproducibility, the code is publicly available on GitHub,[1] and the datasets can be accessed upon request for academic use.

2. Related Work

A persistent challenge in ABSA research concerns the limited availability, diversity, and transparency of annotated datasets (Hua et al., 2024). Existing benchmarks are heavily concentrated in a small number of English review domains, most prominently the SemEval Restaurant and Laptop datasets. While these resources have driven methodological progress, they represent comparatively simplified settings and often yield inflated performance estimates on narrow domain slices (Hua et al., 2024).

[1] GitHub: https://github.com/NiklasDonhauser/absa-annotation-quality

Beyond dataset size, annotation quality and documentation constitute a second critical bottleneck. Modern ABSA formulations such as triplet or quadruplet annotations translate directly into concrete dataset requirements, including a clearly specified label space and guidelines, trained annotators, annotation tools that support the intended output format, and transparent procedures for assessing annotation quality (Pontiki et al., 2016; Klie et al., 2018). Meeting these requirements is time- and cost-intensive. However, many ABSA datasets provide only limited information about sampling strategies, annotator backgrounds, agreement measures, or conflict resolution procedures, complicating reproducibility and reliability assessment. Meta-analyses of NLP datasets highlight recurring deficiencies along dimensions such as stability, reproducibility, accuracy, and unbiasedness (Klie et al., 2024).

These structural issues become even more pronounced in non-English settings, where data scarcity and domain concentration further restrict systematic comparison and reuse.

2.1. The State of ABSA Annotation in German

Against this background, German provides a representative case to examine how structural challenges of ABSA dataset construction materialize in a non-English context. In contrast to English, where shared tasks and benchmark consolidation have shaped methodological development, German ABSA resources have emerged in a more fragmented manner, varying substantially in domain coverage, annotation granularity, and accessibility. The majority of German ABSA datasets provide sentence-level annotations, including *Hotel Reviews* (Fehle et al., 2023), *GERestaurant* (Hellwig et al., 2024), *MobASA* (Gabryszak and Thomas, 2022), *Talk of Literature* (Greve et al., 2021), and *B2B Software Reviews* (Fehle et al., 2025). Review-level annotations are offered by *GermEval 2017* (Wojatzki et al., 2017), while *M-ABSA* provides sentence-level German data via automatic translation, lacking human-authored annotations and ground truth (Wu et al., 2025). Access to several datasets is restricted, as some are proprietary or require direct author contact.

2.2. Annotation Practices in the Literature

Clear annotation guidelines are essential for consistency in NLP tasks (Klie et al., 2024). They define objectives, label spaces, and decision rules, and are often refined iteratively. Their structure can influence annotator behavior and introduce biases,

making careful design crucial. In ABSA, the SemEval shared tasks (2014–2016) established standardized task definitions and guidelines (Pontiki et al., 2014, 2015, 2016), which have been widely reused and adapted. For German, GermEval 2017 followed a similar structure with stronger emphasis on practical instructions and language-specific phenomena (Wojatzki et al., 2017), and subsequent German datasets build on these principles (Hellwig et al., 2024; Fehle et al., 2025).

IAA is often reported as an indicator of annotation consistency (Klie et al., 2024), but it should not be interpreted as a complete measure of annotation quality. In ABSA, span-based and structured annotations complicate agreement measurement. While Krippendorff's α is often applied, many studies report F1-based agreement scores due to the difficulty of using chance-corrected metrics for span extraction (Pontiki et al., 2016; Chebolu et al., 2023). However, agreement alone is insufficient; prior work recommends complementary quality-control measures, such as manual inspection or control instances, and distinguishes between intrinsic evaluation (e.g., IAA) and extrinsic evaluation via downstream performance for application use cases (Klie et al., 2024; Jurafsky and Martin, 2026).

3. Methodology

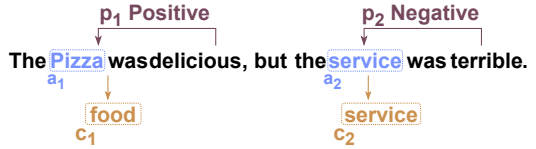
This section describes the annotation and evaluation methodology. First, an independent ground truth was constructed to serve as a reference for evaluation. Subsequently, four distinct annotation settings were implemented, involving crowdworkers, students, LLMs, and task experts. Annotation was performed in batches of 200 sentences, while different mechanisms were used to ensure annotation quality (majority vote, curation, iterative refinement). This design enables a systematic comparison of annotation styles, their interrater agreement, and their impact on downstream model performance.

We build upon GERestaurant (Hellwig et al., 2024), a German ABSA dataset available upon request. It contains 2,154 training and 924 test instances. We use the full test split as ground truth and randomly sample 1,000 training sentences due to annotation resource constraints. The restaurant domain is chosen as it constitutes a widely established benchmark setting for ABSA across languages.

3.1. Annotation Objective

The objective of our annotation studies is to identify all aspect–sentiment pairs expressed within a sentence, following the standard definition of ABSA. Each sentence constitutes one annotation unit, and multiple aspects per sentence are explicitly allowed.

This work focuses on the two established ABSA tasks in the literature: Aspect Category Sentiment Analysis (ACSA) and Target Aspect Sentiment Detection (TASD) (Zhang et al., 2023; Chebolu et al., 2023). In both tasks, annotators assign one or more predefined aspect categories — `FOOD`, `SERVICE`, `AMBIENCE`, `PRICE`, and `GENERAL` — together with a sentiment polarity label — `POSITIVE`, `NEGATIVE`, `NEUTRAL`, and `CONFLICT`. The category `GENERAL` captures overall evaluations of the restaurant that cannot be attributed to one of the other four categories. `CONFLICT` is used when opposing sentiments toward the same aspect (or aspect phrase for TASD) are expressed within a sentence.



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

Table 1: Illustration of the ABSA subtasks investigated in this study. Figure based on Fehle et al. (2025).

Implicit aspects must be identified in both tasks, i.e., cases where a sentiment clearly refers to a predefined category without explicitly mentioning it (e.g., "It tasted really good." → `FOOD`). As depicted in Table 1, the tasks differ in their annotation granularity: ACSA requires aspect category and sentiment polarity assignment only, whereas TASD additionally requires annotating the textual span that realizes the aspect category. For implicit aspects in TASD, no text span is marked. An annotated example from the test set is provided in Appendix B.

3.2. Annotation Strategies

In addition to a new **expert-based ground truth**, we applied four different annotation strategies to construct ABSA datasets: **crowdworkers**, **students**, **LLMs**, and **experts**.

Due to limited expert availability, for both ground truth and expert anntoation, no separate dataset variant was created for ACSA. Instead, the ACSA dataset was derived from the revised TASD annotations by removing aspect phrases and consolidating duplicate tuples.

Ground Truth The ground truth dataset was annotated following the guidelines of Klie et al. (2024).

It comprises 924 sentences annotated independently by two annotators with prior ABSA experience.

To enable incremental quality control, the data were divided into five batches of approximately 185 sentences, after each of which IAA was computed using micro-averaged F1. Following Chebolu et al. (2023), we prefer F1 over chance-corrected measures such as Kappa κ for span extraction tasks. Disagreements were jointly reviewed, and recurring disagreement patterns informed iterative refinements of the annotation guidelines to reduce ambiguity and improve consistency. These refinements mainly concerned the scope and specificity of valid aspect phrases, including the treatment of job titles, generic expressions, national food references, abstract quality terms, and anonymized entities. After the initial annotation phase and guideline updates, a second revision round was conducted to further improve consistency. In this phase, 48 sentences were revised, affecting only aspect phrase boundaries, while polarity and category assignments remained unchanged. Revisions were applied by one annotator and validated by the second, with remaining disagreements resolved through discussion.

Crowdworkers For the dataset based on crowdworker annotations, we recruited 30 participants via Prolific[2] to annotate 200 sentences each. Participation was restricted to German-speaking individuals located in Germany, Austria, or Switzerland with at least secondary education. Participants who had taken part in related annotation studies within this project were excluded to prevent overlap. Each participant was allowed to annotate only a single batch and could not participate in multiple studies (e.g., across the ACSA and TASD tasks). Annotators were compensated at £9 per hour in line with Prolific's recommendations. Each participant could submit only once, and unusually fast submissions were filtered using Prolific's automated quality checks. The task was restricted to desktop devices to ensure consistent interaction with the annotation interface. Before annotation, participants completed a short questionnaire and provided informed consent via Google Forms. Annotations were performed independently using Label Studio[3] following detailed written guidelines and a short instructional video.

Students For this study, computer science-related students were recruited via the university network. Participants received written guidelines and an instructional video explaining the annotation procedure and use of Label Studio. Each batch of 200 texts was annotated independently by three students. Final labels were derived using majority voting, requiring agreement on category, polarity, and aspect phrase. After removing the conflict label, sentences without remaining annotations were retained to ensure consistent dataset sizes across training sets. Two annotators systematically misinterpreted the guidelines by assigning a sentiment label to every aspect category. Instead of leaving non-mentioned categories unannotated, they labeled them as neutral (e.g., assigning neutral to Ambience and Price in a sentence that only expresses sentiment about Food and Service). These annotations were retained in the final dataset, as they reflect a common and instructive source of error in annotation studies. Before starting the annotation, participants completed a short questionnaire via Google Forms, which was identical for both students and crowdworkers to ensure consistent instructions. The questionnaire collected study information, obtained informed consent, and recorded participants' prior annotation experience (ranging from no experience to professional level) and the annotation domains they had worked in (e.g., text, audio, video, image). The results, detailing experience level and type across students and crowdworkers, are provided in Appendix D.

LLMs Following the methodology of Hellwig et al. (2025), we adapted their approach to annotate the training set using an LLM. For our experiments, we employed Gemma-3-27B with a temperature of 0.8, a context length of 4096 tokens, and five random seeds (0 – 4). Based on the results reported by Hellwig et al. (2025), we injected 30 few-shot examples into the prompt. Given that configurations with 30 and 50 examples yielded similar performance in their study, we selected the smaller setting to reduce computational cost and annotation time. To consolidate outputs across seeds, we applied a majority-voting procedure inspired by the self-consistency technique of Wang et al. (2022), whereby an annotation was included in the final annotation if it appeared in the majority of seed predictions.

Experts Expert annotation was based on the original labels provided by Hellwig et al. (2024), which were transformed to match the revised labeling schema of our annotation interface. The expert annotator was a PhD student with prior ABSA experience and involvement in the original annotation as a curator. For the TASD dataset, the expert reviewed all 1,000 texts using Label Studio's review functionality, accepting, revising, or removing annotations per the updated guidelines. In total, 1,333 annotations were accepted, 92 were revised, and 10 triplets were removed. Metadata tags were used to flag difficult or context-dependent cases.

[2] https://www.prolific.com/

[3] https://labelstud.io/

Annotation Guidelines Two separate guideline sets were developed for TASD and ACSA, adapting the framework of Hellwig et al. (2024).[4] The TASD guidelines introduce ABSA, define aspect categories and polarity labels, and specify the annotation of aspect phrases, including the distinction between explicit and implicit mentions. They further outline constraints such as the requirement that sentiment must target the aspect and that only the first occurrence of an aspect phrase is annotated. Illustrative examples of complete aspect–category–sentiment triplets and practical instructions for using the annotation interface, including meta-tags and free-text comments, are also provided. The ACSA guidelines follow the same overall structure but omit the detailed specification of aspect phrases and adjust the interface instructions accordingly. Both guideline sets build on established ABSA annotation principles (Pontiki et al., 2016) and include concrete German-language examples. In Appendix E, we provide the layout of the annotation interface for both tasks (TASD and ACSA).

4. Experiments

The experiments evaluate the same ABSA subtasks considered in the annotation study, ACSA and TASD, to ensure direct comparability between annotation quality and downstream model performance.

4.1. Baseline Methods

To capture a broad range of methodological approaches, we consider classification-based architectures, text generation models, and LLMs. Due to the limited size of our datasets, constructing a separate development set was not feasible. Instead, we adopted the hyperparameter settings proposed by Fehle et al. (2025), who applied the same models to the original GERestaurant dataset. For few-shot prompting experiments, we used Gemma 3 27B,[5] following Hellwig et al. (2025), which also ensured consistency with the model used to generate the LLM dataset. Due to resource constraints, fine-tuning Gemma 3 27B was not feasible. Instead, we employed LLaMA 3.1 8B[6] for instruction fine-tuned experiments. LLaMA-based models have demonstrated strong performance in fine-tuning settings for ABSA and related sentiment analysis tasks (Fehle et al., 2025, 2026; Šmíd et al., 2024). Building on these findings, we adopted the configuration and hyperparameters proposed by Fehle et al. (2025) for the German restaurant domain.

BERT-CLF Following Fehle et al. (2023), we implement a multi-label classification model based on gbert-base.[7] The model predicts aspect–sentiment pairs for the ACSA task, using a linear classification head on top of the [CLS] token representation from BERT.

Hier-CGN The Hierarchical Graph Convolutional Network (Hier-GCN) (Cai et al., 2020) combines contextual embeddings from gbert-base with graph convolutional layers to explicitly model dependencies between aspects and sentiments.

Paraphrase The Paraphrase method (Zhang et al., 2021) treats the TASD task as a sequence-to-sequence text generation problem. Using T5-base[8] as the base model, the input sentence is reformulated into a natural-language template that explicitly encodes the target output structure.

MvP The MvP approach (Gou et al., 2023) models TASD as a sequence-to-sequence generation task using T5-base. Multiple prompt formulations ("views") are applied to generate aspect–category–polarity tuples, and predictions are aggregated via majority voting.

Few-Shot Prompting (Gemma FS) Few-shot prompting leverages LLMs via in-context learning to perform both ACSA and TASD tasks (Simmering and Huoviala, 2023). We use the Gemma 3 27B model and provide 50 annotated examples directly in the prompt, randomly sampled from the corresponding training set. Following Hellwig et al. (2025), who report the best performance with 50 examples, we adopt the same configuration to facilitate comparability with prior work. The prompt template, adapted from Gou et al. (2023), is translated into German and tailored to the specific structure of each task. Further details, including the instruction prompt and the structure of the examples, are provided in Appendix A.

Instruction-based Fine-Tuning (LLaMA FT) Instruction fine-tuning adapts a LLM to directly map input sentences to structured ABSA outputs (Šmíd et al., 2024). We follow the implementation of Fehle et al. (2025) and fine-tune LLaMA 3.1 8B on task-specific datasets for both the ACSA and TASD tasks. The same prompt template is used as in the few-shot setup to ensure consistency in task formulation.

[4] Guidelines: https://github.com/NiklasDonhauser/absa-annotation-quality/tree/main/03_annotations/Guidelines

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

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

[7] https://huggingface.co/deepset/gbert-base

[8] https://huggingface.co/google-t5/t5-base

4.2. Evaluation Procedure

We evaluated annotation consistency, model performance, and statistical differences across datasets. All experiments were conducted on a workstation with an NVIDIA Quadro RTX 6000 (24 GB GDDR6) GPU.

Inter-Annotator Agreement For ACSA, we measured IAA using average pairwise micro-F1 and Krippendorff's alpha (Krippendorff, 2011) to account for chance agreement. For TASD, micro-F1 was used, following prior work (Chebolu et al., 2023; Pontiki et al., 2016), as it captures overlap in aspect phrase spans.

Model Evaluation Models were assessed using micro-F1 scores, averaged over five runs with different random seeds, following previous work in German ABSA (Hellwig et al., 2025; Fehle et al., 2025). To examine whether annotation sources influenced model performance, we conducted all statistical analyses separately for the ACSA and TASD tasks. For each task, we first assessed whether performance differed significantly between datasets when aggregating results across all models. In a second step, we analyzed each model individually to determine whether its performance was influenced by the dataset used, based on five independent runs per model.

Normality assumptions were evaluated using the Shapiro–Wilk test (Shapiro and Wilk, 1965). Depending on the results, either parametric tests (repeated-measures ANOVA followed by paired t-tests) (Student, 1908; Field et al., 2012) or non-parametric alternatives (Friedman test followed by Wilcoxon signed-rank tests) (Friedman, 1937; Wilcoxon, 1992) were applied. Holm–Bonferroni correction was used to account for multiple comparisons (Holm, 1979). Results were considered statistically significant at $p < 0.05$.

5. Results and Discussion

This section reports results for the ACSA and TASD subtasks, including a comparative analysis of dataset variants, inter-annotator agreement, model performance across datasets, and cost and effort considerations.

5.1. Comparative Analysis of Dataset Variants

Across annotation approaches, category distributions for ACSA remain consistent with only minor shifts between datasets. Variations are most pronounced for `GENERAL` and `FOOD`, while `PRICE` remains the most stable category across all annotations. Overall, these differences suggest that annotator type introduces small but systematic changes in category frequencies without substantially altering the overall distribution.

Since ACSA does not distinguish between explicit and implicit mentions, we additionally examine polarity distributions. Sentiment polarity is largely preserved across datasets, with only minor variation between annotation sources, indicating that polarity assignment is comparatively robust to annotator differences.

Compared to ACSA, the differences between annotation approaches are more pronounced for TASD. The expert and LLM-annotated datasets consistently contain higher counts across categories, while student and crowdworker annotations yield noticeably fewer instances, particularly for less frequent categories such as `AMBIENCE` and `PRICE`.

Similar patterns emerge for explicit and implicit mentions as well as sentiment polarity. Expert and LLM datasets maintain higher and more balanced distributions, whereas student and crowdworker datasets exhibit systematic reductions across categories, polarity labels, and mention types. Overall, these results indicate that annotation expertise and automation strongly influence dataset size and class coverage for the more complex TASD task. For a more detailed view of the datasets, including additional statistics, see Appendix C.

5.2. Interrater Agreement during Annotation

The dataset creation process revealed approach-specific trade-offs: crowdsourcing and student annotations differed in reliability and timeliness, LLM-based annotation required substantial computational resources and may reflect training-data biases, and expert annotation achieved the highest quality but was limited in availability, leading to its restriction to TASD and a reduced ACSA formulation.

Although we did not systematically evaluate the impact of guideline clarity or interface design, we assume that clear instructions and a carefully designed annotation interface likely contributed to reducing errors and improving annotation consistency overall. However, the previously observed systematic misinterpretations indicate that certain aspects of the guidelines remained ambiguous, suggesting that not all sources of error can be mitigated through interface design alone, and that particular care is needed in formulating unambiguous annotation guidelines. Questionnaire responses indicate that most students and crowdworkers had little or no prior annotation experience, with only a few reporting moderate or extensive experience. Prior

work was mainly in text and image annotation, with smaller numbers having experience in audio, video, or multimodal tasks. Table 2 presents a comparative overview of IAA for ACSA and TASD across all datasets, excluding the expert dataset due to the presence of only a single annotator.

5.2.1. IAA on the ACSA Task

The IAA results for the ACSA task show broadly consistent patterns across datasets. Crowdworker and student annotations achieve comparable agreement levels, reflecting the constrained annotation setup in which annotators select predefined category–polarity pairs. However, variability across annotation batches suggests that certain texts were more difficult or that annotators applied divergent interpretations. This effect is particularly evident in the student dataset, where unusually high variance in some batches indicates misinterpretation of the guidelines, inflating the overall agreement variance compared to the crowdworker dataset. These observations underscore that annotation errors can occur even with clear instructions, highlighting the importance of carefully designed guidelines and interfaces (Klie et al., 2024). In response, an additional warning was introduced in the crowdworker interface to mitigate similar issues. In contrast, LLM-generated annotations show very high agreement. Despite using a non-zero temperature to encourage output diversity, repeated prompting produced highly consistent annotations, explaining the strong IAA scores. As noted by Klie et al. (2024), however, high agreement alone does not necessarily imply high annotation quality.

5.2.2. IAA on the TASD Task

For the TASD task, IAA differs more strongly across datasets, reflecting the increased annotation complexity. Unlike ACSA, annotators were required to freely select text spans, which substantially increased variability. Student annotations generally adhered to the guidelines, whereas crowdworker annotations often included overly long spans, partial sentence fragments, or inconsistent handling of implicit aspects. Additional errors included splitting multi-word aspects into several single-token annotations, resulting in highly variable annotation quality. These issues occurred far less frequently in the student dataset and are reflected in the higher variance observed for crowdworker annotations. As in ACSA, LLM-generated annotations exhibit very high agreement, despite the use of a non-zero temperature, due to the fixed prompting setup. Overall, IAA for TASD is notably lower than for ACSA, consistent with the added difficulty of aspect phrase extraction. This finding aligns with prior work (Monarch, 2021), which shows that tasks challenging for human annotators also tend to be difficult for machine learning models.

5.2.3. IAA on the Ground Truth

IAA for the ground truth TASD dataset is consistently high and clearly exceeds that of the student and crowdworker annotations. Agreement improves steadily across annotation batches, indicating that iterative discussions, calibration, and guideline refinements led to increasingly consistent annotations with low variance. These findings highlight the importance of expert collaboration and regular feedback in producing reliable gold-standard annotations for complex tasks such as TASD, in line with prior work emphasizing the role of careful guideline design and calibration in improving annotation quality (Klie et al., 2024; Fehle et al., 2025).

5.3. Model Performance on the different Datasets

This section analyzes model performance across the two ABSA subtasks: ACSA and TASD. We compare classical baselines and LLM-based approaches trained on datasets annotated by different annotator types.

5.3.1. Performance on the ACSA Task

Table 3a summarizes model performance on the ACSA task. Overall, models trained on expert-annotated data achieve the strongest results across nearly all approaches, although differences between datasets are relatively small. LLM-based models outperform classical baselines, with fine-tuned and few-shot LLMs on expert annotations achieving the strongest overall results, in line with prior findings by Fehle et al. (2025, 2026). An exception is the few-shot setting, where the student dataset yields the highest score, indicating that non-expert annotations can occasionally be competitive.

Across models, performance remains relatively stable regardless of the annotation source, suggesting that all datasets are broadly suitable for ACSA. Nevertheless, expert annotations consistently provide a small but reliable advantage when optimizing for performance.

Analysis by category and polarity shows that positive sentiment is easiest to predict, followed by negative, while neutral sentiment remains the most challenging. Performance drops are particularly pronounced for neutral polarity in the `AMBIENCE` category, whereas `FOOD` benefits from a higher number of neutral examples.

Pairwise tests reveal isolated significant differences for Hier-GCN (Experts vs. Students) and

	ACSA				TASD			
	GT	**Crowd**	**Students**	**LLMs**	**GT**	**Crowd**	**Students**	**LLMs**
Batch 1	83.93	$66.75_{\pm 11.31}$	$85.11_{\pm 1.43}$	$98.12_{\pm 0.55}$	63.33	$44.47_{\pm 16.20}$	$41.29_{\pm 17.67}$	$90.20_{\pm 2.11}$
Batch 2	88.38	$84.57_{\pm 1.43}$	$81.55_{\pm 1.01}$	$96.46_{\pm 1.10}$	70.25	$61.55_{\pm 6.41}$	$63.81_{\pm 2.34}$	$87.66_{\pm 2.82}$
Batch 3	89.66	$78.94_{\pm 2.19}$	$50.65_{\pm 25.84}$	$96.23_{\pm 1.49}$	75.78	$28.78_{\pm 20.15}$	$45.85_{\pm 7.19}$	$90.74_{\pm 2.21}$
Batch 4	88.25	$84.24_{\pm 1.60}$	$52.03_{\pm 27.54}$	$97.32_{\pm 1.27}$	74.41	$19.91_{\pm 21.97}$	$45.71_{\pm 12.52}$	$89.30_{\pm 1.95}$
Batch 5	85.78	$83.54_{\pm 2.64}$	$81.56_{\pm 1.51}$	$97.86_{\pm 0.69}$	76.95	$26.33_{\pm 26.49}$	$55.26_{\pm 9.64}$	$92.95_{\pm 1.19}$
Overall	**87.22**	**78.95**$_{\pm 2.27}$	**63.38**$_{\pm 16.75}$	**97.20**$_{\pm 0.88}$	**72.18**	**32.38**$_{\pm 10.18}$	**50.50**$_{\pm 5.84}$	**90.22**$_{\pm 1.82}$

Table 2: Batch-wise IAA (micro-F1) for ACSA and TASD across annotation groups. GT (ground truth) was annotated on the test split by two expert annotators. The remaining groups were annotated on the training split. For the experts group on the training split, only one annotator was available, so no IAA could be computed. Values denote per-batch mean $\pm$ standard deviation; standard deviation is not reported for GT.

Method	**Crowd**	**Students**	**LLMs**	**Experts**
BERT-CLF	76.99	77.81	77.44	**78.26**
Hier-GCN	79.66	78.97	79.13	**79.78**
Gemma FS	86.03	**86.43**	85.60	86.29
LLaMA FT	85.64	85.71	84.85	**86.39**

(a) Aspect Category Sentiment Analysis (ACSA)

Method	**Crowd**	**Students**	**LLMs**	**Experts**
Paraphrase	52.77	57.33	57.37	**61.65**
MvP	51.29	56.83	60.65	**64.01**
Gemma FS	58.56	62.28	**65.58**	63.38
LLaMA FT	65.46	69.33	66.24	**71.47**

(b) Target Aspect Sentiment Detection (TASD)

Table 3: Micro-F1 scores averaged over five seeds. Results are reported for ACSA and TASD across annotation sources (Crowd, Students, LLMs, Experts). Bold indicates the best performance per row.

Gemma-FS (Experts vs. LLMs; LLMs vs. Students), without systematic effects over all datasets.

5.3.2. Performance on the TASD Task

Table 3b summarizes TASD performance across datasets. As in ACSA, expert-annotated data yields the strongest results overall, with the highest scores for most models. The best performance is achieved by the fine-tuned LLM, which clearly outperforms all other approaches on the expert dataset. An exception is the few-shot LLM setting, where the LLM-annotated dataset performs slightly better, likely due to the shared underlying model.

Across models, the crowdworker dataset consistently results in the lowest performance, while student and LLM datasets show comparable results, occasionally outperforming each other depending on the model. Compared to ACSA, performance differences between datasets are more pronounced for TASD, reflecting the increased task complexity and lower annotation agreement.

Category–polarity analysis follows similar patterns to ACSA: positive sentiment is easiest to predict, followed by negative, while neutral sentiment remains the most challenging. Performance drops are especially pronounced for infrequent classes such as `PRICE` and for neutral polarity, particularly in the `AMBIENCE` category.

Statistical testing reveals significant overall differences between Crowdworker and Student datasets, with further pairwise effects for Paraphrase and MvP.

5.4. Cost and Effort Analysis

Creating the datasets involved varying levels of cost and effort. Crowdworker studies were completed within two days per study at a total cost of roughly $£828 (\hat{=} £0.41$ per three way annotated example), including platform fees. Student annotations required several weeks while engagement relied on course credits. For LLM-based annotation, assuming $\$0.05 - \$0.25$ per million tokens (MTok) for a self-hosted model (Knoop and Holtmann, 2026) and a budget of ≈10,000 tokens per example, the upper-bound inference cost is at approximately $\$0.0025$ per example. By contrast, commercial frontier APIs such as GPT-5.2 Pro ($\$21/\$168$ per 1M input/output tokens) yield an estimated cost of $\approx \$0.36$ per example. Thus, self-hosted LLM annotation is substantially cheaper per example than crowd-based annotation, while frontier APIs approach similar cost levels. Expert refinement required several hours; without pre-existing labels, large-scale expert annotation would be costly and difficult to scale due to limited availability.

5.5. Summary

The discussion highlights key insights into dataset creation, annotation quality, and model performance for ABSA. Clear annotation guidelines and well-designed interfaces are crucial for reducing errors and improving consistency. While IAA is a use-

ful indicator of task complexity and reliability, high agreement does not necessarily translate into superior model performance, as shown by LLM annotations, which achieved high IAA but only comparable performance to student-annotated data. Each annotation strategy involves trade-offs: crowdsourcing can be costly and inconsistent, student annotations are slower, LLM-generated datasets require computational resources and may reflect training biases, and expert annotations are time-intensive but yield the highest quality. Across tasks, LLM-based models perform competitively, benefiting from large-scale pretraining, while expert annotations remain the most reliable basis for achieving peak performance.

Overall, these findings suggest that LLMs provide a fast and scalable annotation (Dietz et al., 2025; Li et al., 2024) alternative for ABSA, whereas expert annotation remains the gold standard when maximum accuracy and reliability are required.

6. Conclusion

This study systematically analyzed the impact of annotation quality and annotator type on ABSA datasets and model performance. We created a ground truth dataset annotated by two experts and four training datasets produced by crowdworkers, students, LLMs, and task experts, and evaluated SOTA models on ACSA and TASD, alongside IAA analyses. Expert-annotated datasets consistently achieved the highest performance, with LLM-based models generally outperforming classical approaches. For ACSA, LLM annotations approached expert-level quality, while TASD performance was similar across crowdworker, student, and LLM datasets. IAA was lowest for crowdworkers and students, highest for LLMs in ACSA, and more variable for TASD. These findings highlight the importance of structured annotation guidelines and careful interface design. Expert annotations improve dataset quality but are time-intensive, whereas LLM-generated annotations offer a scalable alternative with competitive performance.

Future work includes evaluating models and annotations in other domains, combining LLM and expert annotations, increasing annotator numbers, exploring alternative aggregation strategies, and relaxing strict phrase boundaries to improve coverage and model robustness.

Limitations

This study has several limitations. While crowd annotations enabled rapid data collection, they incur financial costs that scale with dataset size. Student-based annotations were particularly time-intensive, as recruitment and task completion often spanned the entire one-week period. It should be noted that students and crowdworkers are not inherently distinct groups. However, student annotators enable more controlled sampling with respect to demographic characteristics, while crowdworkers generally represent a more diverse but less controllable population. LLM-generated annotations may reflect biases present in the model's training data and require substantial computational resources, including high-memory GPUs, which can limit reproducibility. Furthermore, as the same LLM (Gemma 3 27B) was used for both annotation and few-shot inference, results on LLM-annotated data may be biased due to underlying model. This should be considered when interpreting the performance of the few-shot prompting approach on the LLM dataset. In addition, potential data contamination cannot be fully ruled out, as restaurant reviews are a widely used benchmark domain and may already be represented in the pretraining data of LLMs. Expert annotations yielded the highest-quality data but depend on scarce expertise and are difficult to scale. Furthermore, despite clear instructions prohibiting external assistance, we cannot fully exclude the possibility that student or crowdworker annotators used LLMs as supportive tools or partially outsourced their work, which may have influenced annotation characteristics. Finally, all datasets and analyses are confined to the restaurant review domain, limiting the generalizability of our findings to other domains and languages.

Ethical Considerations

We used OpenAI's GPT-4.5 as a coding assistant to support implementation tasks and as a writing aid to improve clarity and formulation of the manuscript. The dataset and its annotations are available upon request from the authors to ensure responsible academic use, while the Python code for data collection and preprocessing is publicly available on GitHub.[9]

No demographic information was collected from crowdworkers or students, thereby minimizing privacy risks. All annotation procedures followed established data protection and ethical guidelines and were reviewed to prevent potential harms. All participants signed an informed consent form permitting the use of their annotations for research purposes. As with any annotated dataset, individual judgment and bias cannot be fully avoided. For the crowdworker, student, and LLM annotations, this was mitigated via majority voting. Because bias is particularly critical in the ground truth, two expert annotators independently annotated the data and resolved disagreements by consensus. Some residual subjectivity may remain.

[9] GitHub: https://github.com/NiklasDonhauser/absa-annotation-quality

7. Bibliographical References

Jinat Ara, Md. Toufique Hasan, Abdullah Al Omar, and Hanif Bhuiyan. 2020. Understanding Customer Sentiment: Lexical Analysis of Restaurant Reviews. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 295–299. ISSN: 2642-6102.

Baran Barbarestani, Isa Maks, and Piek T.J.M. Vossen. 2024. Content Moderation in Online Platforms: A Study of Annotation Methods for Inappropriate Language. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 96–104, Torino, Italia. ELRA and ICCL.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A Chinese Review Dataset Towards Aspect Category Sentiment Analysis and Rating Prediction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2069–2079.

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.

Ganpat Singh Chauhan, Ravi Nahta, Yogesh Kumar Meena, and Dinesh Gopalani. 2023. Aspect based sentiment analysis using deep learning approaches: A survey. *Computer Science Review*, 49:100576.

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)*, pages 611–628.

Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, pages 218–229, New York, NY, USA. Association for Computing Machinery.

Simon Dobnik and John Kelleher. 2023. On the role of resources in the age of large language models. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 191–197, Gothenburg, Sweden. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jakob Fehle, Niklas Donhauser, Udo Kruschwitz, Nils Constantin Hellwig, and Christian Wolff. 2025. German Aspect-based Sentiment Analysis in the Wild: B2B Dataset Creation and Cross-Domain Evaluation. In *21st Conference on Natural Language Processing (KONVENS 2025)*, volume 9, pages 213–227.

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. *Knowledge-Based Systems*, 336:115277.

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.

Andy Field, Jeremy Miles, and Zoe Field. 2012. *Discovering Statistics Using R*. Sage Publications.

Milton Friedman. 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701.

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.

Zhibin Gou, Qingyan 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, Toronto, Canada. Association for Computational Linguistics.

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.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–25, New York, NY, USA. Association for Computing Machinery.

Nils Constantin Hellwig, Jakob Fehle, Markus Bink, and Christian Wolff. 2024. GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 123–133.

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.

Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian journal of statistics*, pages 65–70. Publisher: JSTOR.

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

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Stanford University. Online manuscript released January 6, 2026.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing Dataset Annotation Quality Management in the Wild. *Computational Linguistics*, 50(3):817–866.

Jonathan Knoop and Hendrik Holtmann. 2026. Private llm inference on consumer blackwell gpus: A practical guide for cost-effective local deployment in smes. *arXiv preprint arXiv:2601.09527*.

Klaus Krippendorff. 2011. *Computing Krippendorff's Alpha-Reliability*. University of Pennsylvania, Department of Communication.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *arXiv preprint arXiv:2412.05579*.

Bing Liu. 2022. *Sentiment Analysis and Opinion Mining*. Synthesis lectures on human language technologies. Morgan & Claypool, San Rafael, California.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Petter Mæhlum, David Samuel, Rebecka Maria Norman, Elma Jelin, Øyvind Andresen Bjertnæs, Lilja Øvrelid, and Erik Velldal. 2024. It's Difficult to Be Neutral – Human and LLM-based Sentiment Annotation of Patient Comments. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 8–19, Torino, Italia. ELRA and ICCL.

Stefanie Nowak and Stefan Rüger. 2010. How Reliable are Annotations via Crowdsourcing: A Study about inter-annotator Agreement for Multi-label Image Annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Will Orr and Kate Crawford. 2024. The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media & Society*, 26(9):4955–4972. Publisher: SAGE Publications.

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. Crowdsourcing vs. Experts: Annotating Open-Domain Conversations with Speech Functions. In *Proceedings*

of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 242–254, Prague, Czechia. Association for Computational Linguistics.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, and Orphée De Clercq. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. 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.

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

Paul F. Simmering and Paavo Huoviala. 2023. Large language models for aspect-based sentiment analysis. *arXiv preprint arXiv:2310.18025*.

Vishal Singhi, Charulata Chauhan, and Piyush Kumar Soni. 2024. Exploring Progress in Aspect-based Sentiment Analysis: An In-depth Survey. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pages 1–10.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25. Publisher: JSTOR.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and et al. 2025. Gemma 3 technical report.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Frank Wilcoxon. 1992. Individual Comparisons by Ranking Methods. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics*, pages 196–202. Springer New York, New York, NY. Series Title: Springer Series in Statistics.

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, Yun Xue, and Barbara Plank. 2025. M-ABSA: A Multilingual Dataset for Aspect-Based Sentiment Analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2530–2557, Suzhou, China. Association for Computational Linguistics.

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.

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. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

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, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A. Prompts for Few-Shot LLMs

```
Gemäß der folgenden Definition der Sentiment-Elemente:

- Der 'Aspektbegriff' ist das genaue Wort oder die genaue Wortgruppe im Text, die eine
spezifische Eigenschaft, ein Merkmal
oder einen Aspekt eines Produkts oder einer Dienstleistung darstellt, über die ein
Nutzer eine Meinung äußern kann. Der Aspektbegriff kann 'NULL' sein, wenn der Aspekt
implizit ist.

- Die 'Aspektkategorie' bezieht sich auf die Kategorie, zu der der Aspekt gehört, und
die verfügbaren Kategorien sind: [[aspect_category]].

- Die 'SentimentPolarität' beschreibt den Grad der Positivität, Negativität oder
Neutralität, die in der Meinung zu einem bestimmten Aspekt oder Merkmal eines Produkts
 oder einer Dienstleistung ausgedrückt wird. Die verfügbaren Polaritäten
sind: 'Positiv', 'Negativ' und 'Neutral'.

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

[[ examples ]]
```

Listing 1: Sample prompt for the TASD task showing few-shot examples before the task sentence.

```
Text: Furztrocken.
Sentiment Elements: [('Essen', 'Negativ', 'NULL')]
Text: Die schönsten Plätze sind draußen an den Mauern der Kirche!
Sentiment Elements: [('Ambiente', 'Positiv', 'Plätze')]
Text: Das Bier schmeckt und die Köbes haben die liebenswerte witzige Art.
Sentiment Elements: [('Essen', 'Positiv', 'Bier'),
('Service', 'Positiv', 'Köbes')]
Text: Ich weiß nicht was das soll.
Sentiment Elements: [('Gesamteindruck', 'Negativ', 'NULL')]
Text: Vor dem Eingang war eine beeindruckende Schlange von wartenden Gästen.
Sentiment Elements: []
...
Text: Wir kommen gerne wieder!
Sentiment Elements: [('Gesamteindruck', 'Positiv', 'NULL')]
Text: [Sentence to predict]
Sentiment Elements:
```

Listing 2: Listing of 30 few-shot examples for the TASD prompt and the corresponding sentence to predict. For space reasons, only a subset is shown.

B. Annotation Examples

Category	**ID**	**Extracted Triplets** *(aspect, sentiment, target)*	**Sentence**
FOOD	736	[["essen", "positive", "Essen"], ["essen", "positive", "Wein"]]	*"Das Essen geschmackvoll, der Wein ein lecker Tröpfchen."*
SERVICE	913	[["service", "positive", "Personal"]]	*"Das Personal war freundlich und zuvorkommend."*
GENERAL	832	[["gesamteindruck", "negative", NULL]]	*"Wir würden diesen Ort nicht empfehlen."*
AMBIENCE	11	[["ambiente", "positive", "Brauhaus"]]	*"Ein tolles uriges Brauhaus mit viel Platz."*
PRICE	303	[["preis", "positive", "Preis-/Leistungsverhältnis"]]	*"Fazit: Preis-/Leistungsverhältnis mehr als stimmig!"*

Table 4: Representative ground truth annotations covering all five aspect categories. Each entry lists the sample ID, the extracted opinion triplets in *(aspect, sentiment, target)* format, and the corresponding source sentence.

(a) Crowdworkers

(b) Students

(c) Large Language Models

(d) Experts

Figure 1: Example annotations of the same text by four groups (crowdworkers, students, LLMs, and experts). For crowdworkers, students, and LLMs, labels are aggregated via majority voting across multiple annotators.

C. Dataset Statistics

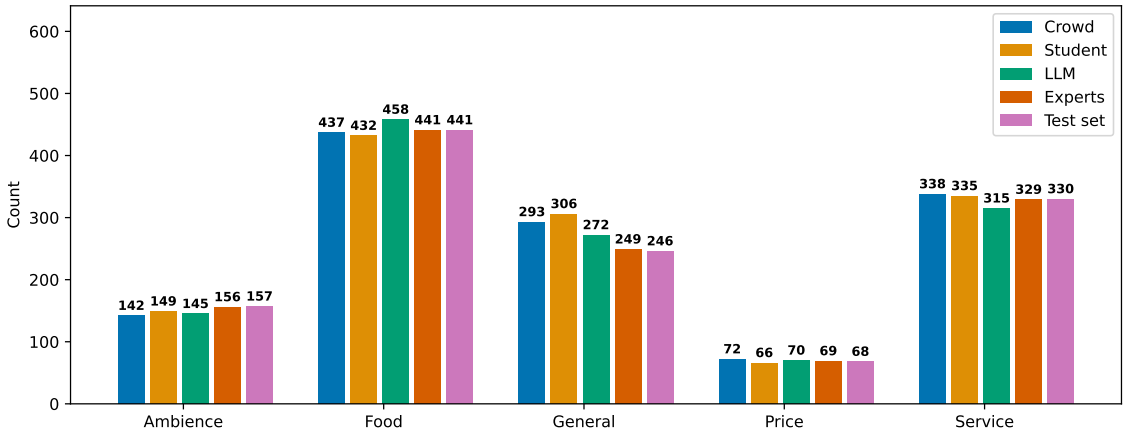
	Positive		Negative		Neutral		Total	
Annotator	**Explicit**	**Implicit**	**Explicit**	**Implicit**	**Explicit**	**Implicit**	**Explicit**	**Implicit**
Crowd	475	111	275	146	27	6	777	263
Student	518	104	291	136	42	1	851	241
LLM	596	177	336	221	64	11	996	409
Experts	613	169	364	220	51	9	1,028	398
Test set	526	146	340	199	48	16	914	361

(a) Target Aspect Sentiment Detection (TASD)

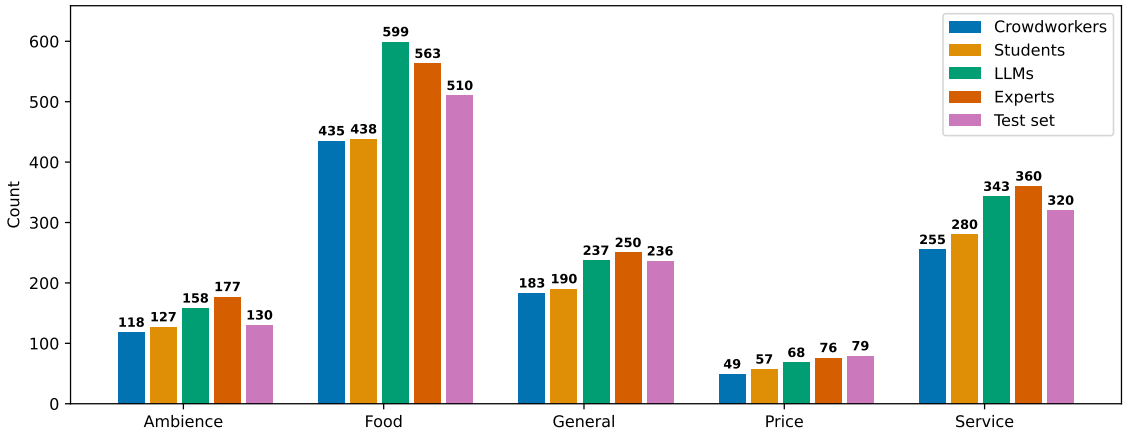
Annotator	**Positive**	**Negative**	**Neutral**	**Total**
Crowd	690	524	68	1,282
Student	686	541	61	1,288
LLM	670	515	75	1,260
Experts	674	514	56	1,244
Test set	671	515	56	1,242

(b) Aspect Category Sentiment Analysis (ACSA)

Table 5: Distribution of sentiment labels (positive, negative, neutral) across datasets for the TASD and ACSA tasks, with a breakdown into explicit and implicit cases, reported for different annotation sources (crowd workers, students, LLMs, experts) and the test set. For the ACSA task, no explicit–implicit distinction is made; therefore, all values are reported under the explicit category.



(a) Aspect Category Sentiment Analysis (ACSA)



(b) Target Aspect Sentiment Detection (TASD)

Figure 2: Distribution of aspect categories across the five datasets for the TASD and ACSA tasks, reported for different annotation sources (crowdworkers, students, LLMs, experts) and the test set. Note that the test set contains fewer texts (924) compared to 1000 for the other datasets.

D. Questionnaire Results

	Students		Crowd	
Experience Level	ACSA	TASD	ACSA	TASD
No experience	9	8	6	4
less than 10 hours	4	3	3	4
10–50 hours	2	2	2	4
more than 50 hours	0	2	3	2
Work in field	0	0	0	1
Don't know	0	0	1	0

(a) Experience level distribution

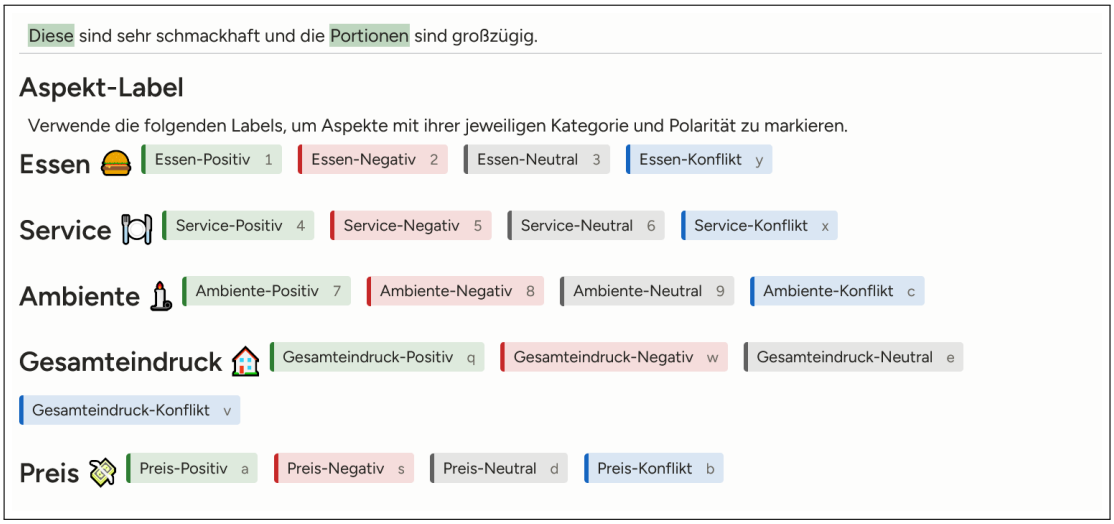
	Students		Crowd	
Type of Experience	ACSA	TASD	ACSA	TASD
Text	4	2	0	9
Image	4	3	3	8
Audio	1	0	3	4
Video	0	0	4	5
Multimodal	1	0	1	2
Miscellaneous	0	0	0	0

(b) Annotation modality distribution

Table 6: Comparison of annotators' prior experience in terms of expertise levels and annotation modalities across student and crowdworker groups in the ACSA (n=15) and TASD (n=15) studies.

E. Label Interface

(a) Label interface for the ACSA task in Label Studio.



(b) Label interface for the TASD task in Label Studio.

Figure 3: Label interfaces used for the ACSA and TASD annotation tasks. While not shown in the screenshots, both interfaces also included two meta tags (one for missing context and one to indicate difficult annotations) and a free-text field for annotator comments.