



Universität Regensburg

Annotation Quality and Its Influence on Aspect-Based Sentiment Analysis: A Case Study on German Restaurant Reviews

Master's Thesis in Medieninformatik
at Institute for Information and Media, Language and Culture (I:IMSK)

Handed in by:	Niklas Donhauser
Address:	Stadtweg 6, 92334 Berching
E-Mail (University):	niklas.donhauser@stud.uni-regensburg.de
E-Mail (private):	niklasdonhauser97@freenet.de
Student Number:	2111397
Primary Corrector:	Prof. Dr. Christian Wolff
Secondary Corrector:	Prof. Dr. Udo Kruschwitz
Supervisor:	Mr. Jakob Fehle
Current Semester:	6. Semester M.S. Medieninformatik
Date handed in:	30.09.2025

Contents

1. Introduction	11
2. Related Work	14
2.1. Sentiment Analysis	14
2.2. Aspect-based Sentiment Analysis	15
2.2.1. Single-task ABSA	16
2.2.2. Joint ABSA	16
2.3. Datasets for ABSA in German	18
2.4. Annotation Practices in Literature	22
2.4.1. Guidelines	23
2.4.2. Annotation Tools	24
2.4.3. Inter-Annotator Agreement Metrics	26
2.5. Approaches to Aspect-Based Sentiment Analysis	28
2.5.1. Rule-based & Lexicon-based Approaches	29
2.5.2. Traditional Machine Learning Approaches	30
2.5.3. Deep Learning & Large Language Models	30
2.5.4. Recent Trends in ABSA Research	32
2.6. ABSA Evaluation Metrics	34
2.7. Summary	36
3. Methodology	38
3.1. Annotation Setup	38
3.1.1. Interface	38
3.1.2. Guidelines	40
3.2. Annotation Strategies	41
3.2.1. Ground Truth	41
3.2.2. Crowd	43
3.2.3. Students	44
3.2.4. LLM	46
3.2.5. Experts	47
3.3. Methods	48
3.3.1. Classification Models	48
3.3.2. Text generation Models	49
3.3.3. Large Language Models	50
3.4. Summary	51
4. Results	53
4.1. Evaluation Procedure	53
4.1.1. Inter-Annotator Agreement	53
4.1.2. Model Evaluation	54
4.1.3. Statistical Testing	54
4.2. Annotation Evaluation	55
4.2.1. IAA for the ACSA Task	56
4.2.2. IAA for the TASD Task	57

4.2.3.	Reliability of Ground Truth Annotations	57
4.2.4.	Cross-Dataset Comparison to the Ground Truth	58
4.3.	Evaluation of Model Performance and Statistical Testing	59
4.3.1.	Performance for the ACSA Task	59
4.3.2.	Performance for the TASD Task	60
4.3.3.	Statistical Evaluation of the ACSA Task	61
4.3.4.	Statistical Evaluation of the TASD Task	64
4.4.	Cost and Effort Analysis	66
4.5.	Comparative Analysis of Dataset Variants	67
4.5.1.	ACSA	68
4.5.2.	TASD	68
4.5.3.	Ground Truth	70
4.6.	Summary	71
5.	Discussion	73
5.1.	Creating of the different Datasets	73
5.1.1.	IAA on the ACSA Task	74
5.1.2.	IAA on the TASD Task	74
5.1.3.	IAA on the Ground Truth	75
5.2.	Performance across Datasets	76
5.2.1.	Performance on the ACSA Task	76
5.2.2.	Performance on the TASD Task	77
5.3.	Limitations & Ethical Considerations	78
5.4.	Summary	79
6.	Conclusion & Future Work	80
	Bibliography	82
A.	Appendix	96
A.1.	Prompts Examples for Few-Shot LLMs	96
A.2.	Study Description on Prolific	98
A.3.	Annotation Example	98
A.4.	Statistical Testing	99
A.5.	Accuracy Scores for ACSA and TASD	100
A.6.	Training Time and Memory Usage	101
A.7.	Label Interface ACSA	101
A.8.	Dataset Statistics	102
A.9.	Category and Polarity F1-scores	105
	Erklärung zur Urheberschaft	108
	Erklärung zur Lizenzierung und Publikation dieser Arbeit	109

List of Figures

1.	Combination of different ABSA subtasks and their combination. Adapted from W. Zhang et al. (2023).	16
2.	Label interface for the T ASD task in Label Studio.	39
3.	Example for the category and polarity description in the T ASD guidelines.	40
4.	Micro-F1 scores for ACSA across annotation datasets. Each box shows the distribution of performance averaged over seeds for each model, combined across all models within the dataset.	62
5.	Micro-F1 scores for the ACSA task across annotation datasets. Each box represents the distribution of performance over five random seeds for a given model–dataset combination, shown separately per model. Note: the y-axis scale is adjusted individually for each model.	63
6.	Micro-F1 scores for T ASD across annotation datasets. Each box shows the distribution of performance averaged over seeds for each model, combined across all models within the dataset.	65
7.	Micro-F1 scores for the T ASD task across annotation datasets. Each box represents the distribution of performance over five random seeds for a given model–dataset combination, shown separately per model. Note: the y-axis scale is adjusted individually for each model.	66
8.	Category distribution across datasets for the ACSA datasets. The figure shows the number of annotated aspects per category across the different datasets. The total number of aspects per dataset is displayed at the end of each bar.	69
9.	Category distribution across datasets for the T ASD datasets. The figure shows the number of annotated aspects per category across the different datasets. The total number of aspects per dataset is displayed at the end of each bar.	70
10.	Category distribution across datasets for the ground truth datasets. The figure shows the number of annotated aspects per category across the different datasets. The total number of aspects per dataset is displayed at the end of each bar.	71
11.	Study description shown to participants on Prolific.	98
12.	Label interface for the ACSA task in Label Studio.	101
13.	Micro-F1 scores for the T ASD task across categories, phrase prediction. Results are averaged over methods and seeds for each dataset.	105
14.	Micro-F1 scores for the ACSA task across categories, separated by polarity (positive, negative, neutral). Results are averaged over methods and seeds for each dataset.	106

15.	Micro-F1 scores for the TASD task across categories, separated by polarity (positive, negative, neutral). Results are averaged over methods and seeds for each dataset.	107
-----	---	-----

List of Tables

1.	Illustration of the different ABSA subtasks. Figure based on Fehle et al. (2025).	15
2.	Overview of the datasets, including the number of documents, aspect annotations, and sentiment distribution. Older datasets such as MLSA, PoTS, and Scare are excluded due to differences in the ABSA task definition.	23
3.	Annotation tool usage frequency across reviewed corpora.	25
4.	Overview of annotators' experience levels and types of annotation across the ACSA (n=15), TASD Pilot (n=5), and TASD (n=10) studies.	45
5.	Comparison of micro-F1 scores (Gemma 3 27B) across different few-shot counts (Hellwig et al., 2025). Best results are highlighted in bold; second-best results are underlined.	47
6.	Model configurations used in this study. Values for epochs, batch size, and learning rate were chosen based on the paper by Fehle et al. (2025).	51
7.	Batch-wise evaluation metrics for the three datasets (ACSA). Values are per batch averages with the \pm standard deviation. At the bottom, average macro and micro scores across batches are provided. Note: The expert dataset includes only one annotator, so reliability metrics cannot be computed. Abbr.: Pair-F1 = average pairwise micro-F1, α = Krippendorff's alpha.	56
8.	Batch-wise evaluation metrics for the three datasets (TASD). Average pairwise micro-F1 are per batches with the \pm standard deviation. At the bottom, average macro and micro scores across batches are provided. Note: The expert dataset includes only one annotator, so reliability metrics cannot be computed.	57
9.	Batch-wise evaluation metrics for the ground truth. Values are per batch averages with the \pm standard deviation. Abbr.: Pair-F1 = average pairwise micro-F1, α = Krippendorff's alpha.	58
10.	Comparison of IAA and model-independent metrics on predefined 100-text subsets drawn from each dataset's training portion. For ACSA, both micro-F1 and Krippendorff's alpha are reported. For TASD, micro-F1 is reported. Ground truth values correspond to agreement between two expert annotators.	59
11.	Micro- and macro-F1 scores for ACSA, averaged over five seeds across datasets. Highest values are shown in bold.	60
12.	Micro- and macro-F1 scores for TASD, averaged over five seeds across datasets. Bold indicates the highest values.	61
13.	Examples of ground truth annotations showing all aspect categories, IDs, extracted triplets, and their corresponding sentences.	98

14.	Adjusted p-values (Holm–Bonferroni) for pairwise comparisons of micro-F1 across datasets per model for the ACSA task. Significant values ($p < 0.05$) are in bold.	99
15.	Adjusted p-values (Holm–Bonferroni) for pairwise comparisons of micro-F1 across datasets per model for the TASD task. Significant values ($p < 0.05$) are in bold.	100
16.	Accuracy scores for ACSA, averaged over five seeds across datasets. Bold indicates the highest values	100
17.	Accuracy scores for TASD, averaged over five seeds across datasets. Bold indicates the highest values.	101
18.	Training times and GPU memory usage per model and task on the GERestaurant (LLM) dataset, averaged over five seeds. Because of the similar size of the training sets and the shared test set, we only report the values for one dataset.	101
19.	Ground truth distribution of aspect categories across sentiment polarities and reference types for the original GERestaurant test sets. . .	102
20.	Ground truth distribution of aspect categories across sentiment polarities and reference types for the new annotated test sets.	102
21.	Counts of polarity triplets by category, with explicit/implicit split for the crowd dataset.	103
22.	Counts of polarity triplets by category, with explicit/implicit split for the student dataset	103
23.	Counts of polarity triplets by category, with explicit/implicit split for the LLM dataset.	103
24.	Counts of polarity triplets by category, with explicit/implicit split for the Experts dataset.	104
25.	Counts of polarity triplets by category, with explicit/implicit split for the original GERestaurant train set.	104

List of Listings

1. Sample prompt for the TASD task showing few-shot examples before the task sentence. 96
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. 97

Zusammenfassung

Aspect-based Sentiment Analysis (ABSA) ermöglicht eine fein granulare Bewertung von Meinungen, indem nicht nur die allgemeine Stimmung, sondern auch die Haltung gegenüber spezifischen Aspekten oder Zielen innerhalb eines Textes identifiziert werden. Während die Aufgabe im Englischen bereits intensiv untersucht wurde, ist die Forschung zu deutscher ABSA noch begrenzt, da hochwertige, annotierte Datensätze selten sind. Zuverlässige Annotationen sind entscheidend für das Training und die Evaluierung von Machine-Learning-Modellen, wobei Qualität und Konsistenz dieser Daten stark von der eingesetzten Annotationsstrategie abhängen.

Diese Arbeit untersucht, wie unterschiedliche Quellen von Annotationen die Entwicklung der deutschen ABSA beeinflussen. Ein bestehender Datensatz wird von Experten erneut annotiert, um eine Ground-Truth zu erstellen, die als Referenzpunkt zur Evaluierung von Annotationen durch Studierende, Crowdfworker, Large Language Models in Few-Shot-Settings und Experten dient. Die Studie vergleicht die Qualität dieser Annotationen unter Verwendung des Inter-Annotator-Agreements als zentrales Maß, analysiert die Konsistenz über verschiedene Datensätze und Annotationsquellen hinweg und untersucht, wie diese Unterschiede die Performance nachfolgender Modelle beeinflussen.

Die Evaluierung konzentriert sich auf zwei zentrale ABSA-Aufgaben: Aspect Category Sentiment Analysis und Target Aspect Sentiment Detection. State-of-the-art-Methoden, darunter Paraphrase, Multi-View Prompting, BERT-CLF, fine-tuned LLMs und Few-Shot-Prompting-Ansätze, werden angewendet, um zu beurteilen, wie Unterschiede in der Annotationsqualität in messbare Unterschiede bei den Modellergebnissen übersetzt werden.

Die Ergebnisse tragen zu einem tieferen Verständnis der Beziehung zwischen Annotationsstrategie, Inter-Annotator-Agreement und Modellperformance in der ABSA bei und liefern praxisnahe Erkenntnisse für die Erstellung annotierter Datensätze in ressourcenarmen Sprachen wie dem Deutschen.

Abstract

Aspect-based sentiment analysis (ABSA) enables fine-grained evaluation of opinions by identifying not only overall sentiment but also sentiments toward specific aspects or targets within a text. While the task has been extensively studied in English, research on German remains limited due to the scarcity of high-quality annotated datasets. Reliable annotations are essential for training and evaluating machine learning models, yet the quality and consistency of such data strongly depend on the annotation strategy employed.

This thesis investigates how different sources of annotation affect the development of German ABSA. An existing dataset is re-annotated to establish a ground truth created by experts, which serves as the reference point for evaluating annotations generated by students, crowdworkers, large language models in few-shot settings and experts. The study compares the quality of these annotations using inter-annotator agreement as a central measure, analyzing consistency across different datasets and annotator types, and examines how these differences impact downstream model performance.

The evaluation focuses on two central ABSA tasks: Aspect Category Sentiment Analysis and Target Aspect Sentiment Detection. State-of-the-art methods, including Paraphrase, Multi-View Prompting, BERT-CLF, fine-tuned large language models and few-shot prompting approaches, are applied to assess how differences in annotation quality translate into measurable differences in model outcomes.

The findings contribute to a deeper understanding of the relationship between annotation strategy, inter-annotator agreement, and model performance in ABSA, providing practical insights for the creation of annotated datasets in poor-resource languages such as German. By highlighting trade-offs between annotation reliability, efficiency, and cost, the thesis offers guidance for future work in resource construction and evaluation for sentiment analysis in German.

1. Introduction

The growing availability of user-generated content such as product reviews, customer feedback, and social media posts has made sentiment analysis (SA) one of the most widely studied tasks in natural language processing (Wankhade et al., 2024; Brauwiers & Frasincar, 2022; Chauhan et al., 2023). SA aims to automatically identify and classify subjective opinions in texts, thereby enabling the extraction of valuable insights from large volumes of unstructured data. While traditional SA focuses on the overall polarity of a text, aspect-based sentiment analysis (ABSA) extends this perspective by examining sentiments tied to specific aspects or attributes (B. Liu, 2022). This finer granularity offers deeper insights, making ABSA particularly useful in domains where user opinions cover multiple dimensions. Restaurant reviews are especially suitable for this purpose: they are widely used as benchmark domains in ABSA research across many languages (Pontiki et al., 2014, 2015, 2016; Chebolu et al., 2023) and they also have clear practical relevance for applications like 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 a low-resource 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 & Crawford, 2024). The challenge is amplified by the fact that different annotation strategies, such as crowdsourcing, student annotators, expert annotators, or the use of large language models (LLMs) can produce datasets of varying reliability and utility for machine learning models. This raises the central question of how

annotation quality influences the performance of ABSA systems and whether the benefits of high-quality annotations justify their associated costs.

The main goal of this thesis is therefore to evaluate the influence of different annotation strategies on ABSA for German restaurant reviews. Specifically, the study investigates annotations produced by four groups:

- Crowdworkers
- Students
- Large Language Models
- Task Experts

By comparing these approaches, the thesis seeks to uncover quality differences in the resulting datasets and to examine their consequences for machine learning performance. The evaluation concentrates on two key subtasks of ABSA: Aspect Category Sentiment Analysis (ACSA) and Target Aspect Sentiment Detection (TASD) (Fehle et al., 2025; Hellwig et al., 2025; Bu et al., 2021; Wu, Ma, Liu, et al., 2025). In addition to model performance, the study systematically analyzes inter-annotator agreement as a central measure of annotation reliability 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, HIER-GCN, as well as more recent text generation and large language model techniques, including Paraphrase, Multi-View Prompting, few-shot LLMs, and fine-tuned LLMs.

The scope of this study is restricted to the restaurant domain, where user feedback is particularly aspect-rich and relevant to both industry stakeholders and academic research (Singhi et al., 2024; Pontiki et al., 2014, 2015, 2016). While the analysis covers multiple annotation strategies and ABSA subtasks, it does not extend to multilingual or cross-domain applications. Instead, the emphasis is on conducting an in-depth comparison within one domain to generate clear and interpretable findings.

The structure of the thesis is organized as follows: Chapter 2 reviews related work, beginning with a brief introduction to sentiment analysis and its extension to aspect-

based sentiment analysis, including both single and joint tasks. This is followed by an overview of existing German datasets, a discussion of annotation practices in the literature, covering guidelines, annotation tools, and inter-annotator agreement and a brief survey of approaches in ABSA together with commonly used evaluation metrics. Chapter 3 presents the methodology, detailing the annotation setup, the annotation strategies pursued across different annotator groups, and the modeling approaches employed. Chapter 4 reports the results, structured around the evaluation procedure: annotation evaluation (with a focus on agreement and reliability), model performance and statistical testing, as well as cost and effort analysis. A comparative analysis of the resulting datasets is also included. Chapter 5 discusses the findings with respect to the creation of reliable datasets, agreement between annotators, and model performance across datasets. It further addresses limitations and ethical considerations of the study. Finally, Chapter 6 concludes the thesis by summarizing the main contributions, outlining limitations, and suggesting possible directions for future work.

In summary, this thesis conducts a systematic annotation study on German restaurant reviews, comparing four different annotation strategies: crowdworkers, students, large language models, and experts. This work evaluates their impact on the quality of ABSA datasets and examines how annotation differences influence model performance across two key subtasks: ACSA and TASD. By doing so, it provides both practical recommendations for dataset construction and theoretical insights into the role of annotation quality in natural language processing (NLP). To support reproducibility and further research, all code for this thesis is provided on GitHub¹, while the datasets are available upon request to ensure responsible usage for academic purposes and to preserve the original intent of the dataset.

¹GitHub:<https://github.com/ValdrDarmir/Annotation-Quality-and-Its-Influence-on-ABSA-A-Case-Study-on-German-Restaurant-Reviews>

2. Related Work

The chapter starts with a definition of sentiment analysis and its extension, aspect-based sentiment analysis. Utilizing these definitions as a foundation, a subsequent analysis encompasses a review of pertinent German datasets, alongside an examination of annotation practices. This includes guidelines, tools, and the inter-annotator agreement. The review then progresses to approaches in ABSA and the discussion of frequently employed evaluation metrics.

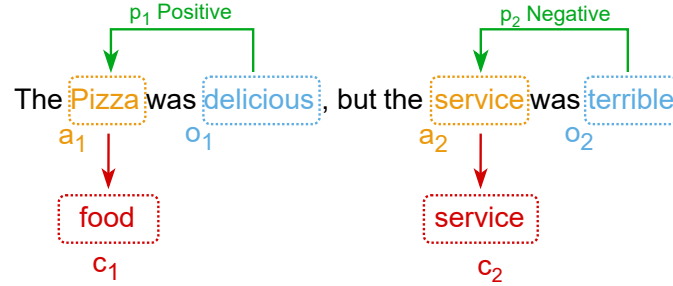
2.1. Sentiment Analysis

Sentiment analysis (SA), also referred to as opinion mining, is the study of people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions toward various entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (B. Liu, 2022). According to B. Liu (2022), SA represents a broad and complex problem space, as sentiments can be expressed in many forms and with varying degrees of subtlety.

With the explosive growth of social media and user-generated content, including reviews, forum discussions, blogs, micro-blogs, Twitter posts, comments, and other social network interactions, individuals and organizations increasingly rely on this data for decision making (Wankhade et al., 2024; Brauwers & Frasincar, 2022; Chauhan et al., 2023). SA enables the extraction and aggregation of such opinions to understand public perception, improve products and services, and support strategic decisions (B. Liu, 2022).

2.2. Aspect-based Sentiment Analysis

ABSA represents a finer-grained approach within the broader field of SA. Instead of classifying an entire document or sentence as positive or negative, ABSA explicitly links sentiments to specific aspects of entities. According to B. Liu (2022), an opinion can be understood as a combination of a sentiment (e.g., positive, negative or neutral) and its corresponding target. Identifying sentiment without its target is of limited use, since applications often require knowing what the sentiment refers to.



Subtask	Output
Aspect-Opinion Pair Extraction (AOPE)	(a, o)
End-to-End ABSA (E2E-ABSA)	(a, o)
Aspect Category Sentiment Analysis (ACSA)	(c, p)
Aspect Sentiment Triplet Extraction (ASTE)	(a, o, p)
Target Aspect Sentiment Detection (TASD)	(a, c, p)
Aspect-based Sentiment Analysis Quad Prediction (ASQP)	
Category-Opinion-Sentiment Quadruple Extraction (ACOS)	(a, c, o, p)

Table 1.: Illustration of the different ABSA subtasks. Figure based on Fehle et al. (2025).

In practice, opinion targets are typically entities (e.g., a restaurant) and their associated aspects (e.g., food, price, service). The goal of ABSA is therefore to detect both the target aspects and the sentiment expressed towards them. This process enables the transformation of unstructured text, such as customer reviews, into structured data. Such structured outputs can then be used to create detailed summaries of opinions across aspects, supporting both qualitative and quantitative analyses (B. Liu, 2022).

The following sections providing definitions and discussions of the individual ABSA tasks. The following sentence, presented in Table 1 ,should help to better understand the extraction of the individual elements and the joint tasks:

2.2.1. Single-task ABSA

ABSA is typically divided into several subtasks that capture the entities, opinions, and sentiments expressed in a text (W. Zhang et al., 2023). **Aspect Term Extraction (ATE)** identifies explicit aspect phrases that are the targets of opinions. For example, in the sentence, the terms *Pizza* and *service* would be extracted as aspect terms. **Aspect Category Detection (ACD)** assigns aspect terms to predefined domain-specific categories, such as *food* for *Pizza* and *service* for *service*. **Opinion Term Extraction (OTE)** identifies the opinion expressions that refer to aspects, such as *delicious* and *terrible* in the example sentence. Finally, **Aspect Sentiment Classification (ASC)** determines the sentiment polarity for a specific aspect in context, assigning a *positive* polarity to *Pizza* and a *negative* polarity to *service*.

2.2.2. Joint ABSA

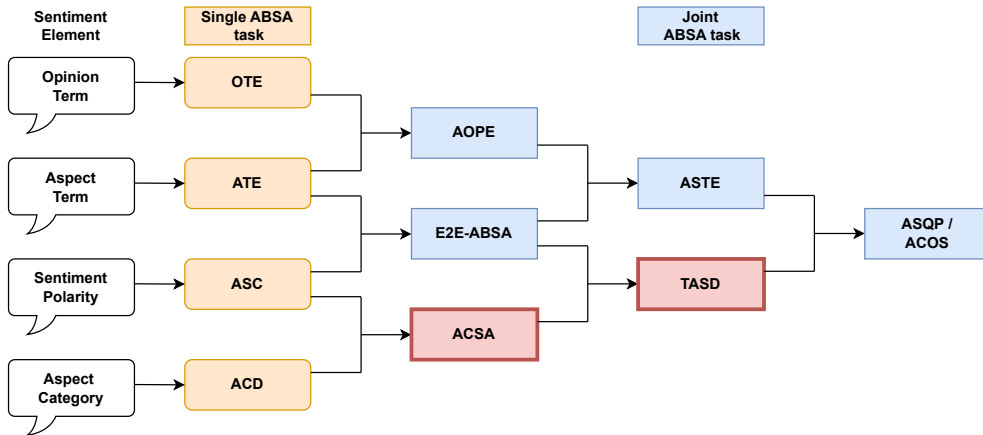


Figure 1: Combination of different ABSA subtasks and their combination. Adapted from W. Zhang et al. (2023).

While the subtasks of ABSA can be studied individually, they are often combined into joint tasks that aim to extract multiple sentiment elements simultaneously (see Figure 1). These so-called joint tasks differ in the number and type of elements

they capture. **Aspect-Opinion Pair Extraction (AOPE)** identifies pairs of aspect terms and their corresponding opinion expressions, such as (*pizza*, *delicious*) and (*service*, *terrible*). **End-to-End ABSA (E2E-ABSA)** links each aspect directly with its polarity, e.g., (*pizza*, *positive*) and (*service*, *negative*).

More complex tasks extend this idea to triplets or quadruples. **Aspect Sentiment Triplet Extraction (ASTE)** combines aspect, opinion, and polarity, e.g., (*pizza*, *delicious*, *positive*). **Aspect-based Sentiment Analysis Quad Prediction (ASQP)** predicts quads consisting of aspect, category, opinion, and polarity, such as (*pizza*, *food*, *delicious*, *positive*). Here only aspect-phrases can be implicit. Finally, **Aspect-Category-Opinion-Sentiment Quadruple Extraction (ACOS)** also targets quads of aspect, category, opinion, and polarity, but additionally allows for implicit aspect or opinion expressions, e.g., (*pizza*, *food*, *delicious*, *positive*) (W. Zhang et al., 2023; Wankhade et al., 2024; Cai et al., 2021).

Among the various ABSA formulations, two tasks are of particular relevance to this work: Aspect Category Sentiment Analysis and Target Aspect Sentiment Detection.

In **Aspect Category Sentiment Analysis (ACSA)**, the aspect is represented by its predefined *category*, and the goal is to determine the associated *polarity*, such as (*food*, *positive*) and (*service*, *negative*). Traditional approaches often rely on a pipelined setup, where aspect categories are detected first and sentiment is assigned in a second step. While conceptually simple, such pipelines suffer from error propagation, since misclassified categories directly affect sentiment prediction. To address this, more recent work proposes unified approaches that jointly model categories and polarities, leveraging their interdependence to improve robustness and contextual consistency (W. Zhang et al., 2023).

Target Aspect Sentiment Detection (TASD) extends ACSA by explicitly linking aspect terms, aspect categories, and their sentiment polarity, e.g., (*Pizza*, *food*, *positive*) and (*service*, *service*, *negative*). The structure of this task was already anticipated in early benchmark datasets, such as SemEval 2016, which defined all the

necessary subtasks but still treated them as separate steps (Pontiki et al., 2016). The first joint formulation, integrating all elements in one framework, was presented by Brun & Nikoulina (2018), and to the best of our knowledge, the term *Target Aspect Sentiment Detection* was introduced later by Wan et al. (2020). T ASD is therefore a relatively recent but increasingly influential formulation, as it captures both the explicit mention of an aspect and its abstract category together with the expressed sentiment.

2.3. Datasets for ABSA in German

In this section, datasets for ABSA in German are presented. The majority of the resources provide sentence-level annotations, including *Hotel Reviews*, *Talk of Literature*, *GERestaurant*, *MobASA*, and *B2B Software Reviews*. Other datasets, such as *GermEval 2017*, offer review-level annotations. Older resources like *MLSA*, *Scare*, and *PoTS* contain entity-level annotations but do not align with current definitions of ABSA. Finally, *M-ABSA* provides sentence-level data as well, but it is automatically translated into German and therefore lacks human-provided annotations for the German subset. In addition, some datasets are not publicly available (*B2B Software Reviews*, *Talk of Literature*) or require contacting the authors (*GERestaurant*) to gain access.

GermEval 2017 The *GermEval 2017* dataset introduced by Wojatzki et al. (2017) consists of German-language customer reviews and news articles centered on “Deutsche Bahn”, the major German public railway operator. The dataset was created to support ABSA and includes both synchronic and diachronic test sets. The latter was developed to evaluate the temporal robustness of sentiment analysis systems and was annotated using the same procedure as the main corpus. Annotations were performed using the *WebAnno* tool Yimam et al. (2014). The annotation process involved six trained student annotators and a curator. Each document was annotated by two different annotators in rotating pairings, with the curator resolving any disagreements. The dataset includes 19 predefined aspect categories and

contains a total of 26,209 annotated aspects.

Talk of Literature The *Talk of Literature* dataset introduced by Greve et al. (2021) comprises German-language Twitter posts focusing on the Ingeborg-Bachmann-Preis (TDDL). The annotations include aspect terms, aspect categories, sentiment polarities, and named entities. The annotation was performed with *INCEpTION*. The dataset encompasses a total of 4,521 tweets and includes 8,264 annotated aspects, categorized under seven main aspect categories. The dataset is not publicly available. Furthermore, no specific annotation guidelines were published.

MobASA The *MobASA* dataset introduced by Gabryszak & Thomas (2022) is a novel German-language corpus consisting of tweets annotated for their relevance to public transportation and for sentiment toward aspects related to barrier-free travel. The dataset includes 19 aspect categories and applies a three-point polarity scale: positive, negative, and neutral. Annotations were conducted using two strategies (expert annotation and crowdsourcing) with the expert-annotated subset containing a total of 5,927 aspects. The annotation process was carried out using the *INCEpTION* tool Klie et al. (2018), and the guidelines for labeling were partly based on the annotation instructions from the GermEval 2017 and SemEval 2015 datasets.

Hotel Reviews The dataset *Hotel Reviews* (Fehle et al., 2023) contains customer reviews of hotels collected from Tripadvisor and annotated at sentence level for aspect-based sentiment. Each annotation includes a main aspect, a subcategory and a sentiment polarity. Annotation was performed by student annotators based on majority voting, with expert curation and supervision, inside the *INCEpTION* platform (Fehle et al., 2023). The dataset consists of 4,254 sentences and 5,617 annotated aspects.

GERestaurant The dataset *GERestaurant* (Hellwig et al., 2024) consists of German-language restaurant reviews from TripAdvisor and follows an annotation scheme

similar to the SemEval 2016 restaurant dataset, but the aspect categories have been adapted. The aspect classes used are `FOOD`, `SERVICE`, `AMBIENCE`, `PRICE` and `GENERAL`. The 3,212 user ratings were divided into 12,795 sentences using Tokenizer and then 5,000 sentences were randomly filtered out for annotation. After sorting out wrongly split sentences, sentences without sentiment and sentences with conflict labels, 3,212 sentences remained in the final data set, which was split into training and test splits using a 70/30 split. The annotations include aspect categories, sentiment polarities and sentiment targets at sentence level, taking into account both explicit and implicit aspect terms. The annotation was performed by an experienced computer science student and subsequently validated by an experienced annotator, inside *Label Studio* (Hellwig et al., 2024).

M-ABSA The *M-ABSA* dataset (Wu, Ma, Liu, et al., 2025) comprises texts from seven domains: `Hotel`, `Food`, `Coursera`, `Phone`, `Laptop`, `Restaurant`, and `Sight` across 21 languages. Most of the data originates from existing corpora originally created for the ABSA triplet extraction task. The `Sight` domain is the only newly annotated dataset, specifically enriched with triplet-level sentiment annotations. Annotation guidelines were adapted from SemEval 2016 Pontiki et al. (2016). A key characteristic of M-ABSA is its multilinguality: the English datasets were translated into 20 additional languages using the Google Translate API.

B2B Software Reviews The *B2B Software Reviews* dataset presented by Fehle et al. (2025) is derived from a proprietary collection of user feedback collected by a B2B software provider. The dataset includes 1,500 reviews, segmented into 3,918 sentences, and annotated with 3,479 aspects across 8 aspect categories. Annotation was conducted using the *Label Studio* annotation tool. The annotation process was guided by methodologies established in Hellwig et al. (2024) and Pontiki et al. (2016). Two domain experts (a PhD and a Master’s student in computer science) carried out the annotation. The dataset is not publicly available.

In addition to the datasets that directly align with the modern ABSA task definitions, there also exist a number of earlier German resources that have been applied in related SA settings. While these datasets do not fully match today’s subtask structure, they represent important precursors. For completeness, we briefly describe three such datasets below.

PoTS The dataset introduced by Sidarenka (2016) comprises a comprehensive collection of 7,992 German-language tweets, manually annotated with fine-grained opinion relations. The tweets cover four domains: the 2013 German federal elections, the papal conclave, general politics, and everyday conversation. Annotations include three main components: (1) sentiment targets, described as objects or events evaluated by sentiment expressions; (2) opinion sources, defined as the authors or holders of evaluative expressions; and (3) the sentiment expressions themselves, marked as minimal syntactic or discourse units in which both the target and sentiment co-occur. Two human annotators carried out the labeling process. To improve interrater reliability, differences between the annotations were automatically identified in a post-processing step and recorded as a separate category of labels. Annotation was performed using the *MMAX2* tool.

Scare *SCARE* (Sänger et al., 2016) is a corpus of German-language app reviews designed to support fine-grained sentiment analysis in the domain of mobile applications. It includes 1,760 annotated user reviews from the Google Play Store, containing 2,487 application aspects and 3,959 subjective phrases, along with their semantic relations. Annotation was carried out using the *BRAT* tool by four trained annotators over multiple iterations, supported by refined guidelines.

MLSA *MLSA* (Clematide et al., 2012) is a publicly available multi-layered corpus for German-language sentiment analysis, containing 270 annotated sentences. It includes three levels of annotation: sentence level, word and phrase level, and expression level. Layer 2 focuses on annotating polarity at the word and phrase level, particularly for nominal and prepositional phrases. All layers were annotated by

multiple raters, with inter-annotator agreement evaluated to ensure quality.

Summary

In summary, the reviewed datasets cover a wide range of domains, including hospitality, transportation, politics, app reviews, and B2B software, and they exhibit notable diversity in annotation scope and granularity. At the same time, this section highlights the limited availability of up-to-date and openly accessible German ABSA datasets. By mapping the existing resources, this section provides the foundation for selecting a suitable dataset for re-annotation in the following chapters. While some corpora focus on sentence-level polarity or aspect categorization, others incorporate richer structures such as subjective phrases, opinion sources, and sentiment targets.

Annotation was performed using a variety of tools such as *BRAT*, *WebAnno*, *INCEpTION*, and *Label Studio*, and involved strategies ranging from crowdsourcing to expert annotation with curated resolution procedures. Public availability, annotation guidelines, and agreement measures varied significantly, with some corpora well-documented and openly accessible, while others remain proprietary or lack formal guideline publications.

Among these domains, the restaurant sector stands out as one of the most extensively studied and benchmarked in ABSA research, not least due to its central role in the SemEval shared tasks (Pontiki et al., 2014, 2015, 2016). In addition, recent work has introduced German ABSA corpora for this domain (Hellwig et al., 2024; Wu, Ma, Liu, et al., 2025), further strengthening its position as a well-resourced benchmark setting. The availability of established resources and annotation practices therefore makes it a natural focus for more detailed investigation in the following chapters.

2.4. Annotation Practices in Literature

In this section, annotation practices in the literature are reviewed. First, annotation guidelines are presented, describing how consistency and reliability are main-

Dataset	#Doc	#Asp	Pos	Neg	Neu
GermEval <small>Wojatzki et al. (2017)</small>	27,800	26,209	17,758	6,911	1,540
Talk of Literature <small>Greve et al. (2021)</small>	4,521	8,264	2,637	2,757	2,870
MobASA <small>Gabryszak & Thomas (2022)</small>	5,201	13,533	2,361	9,520	1,652
Hotel Reviews <small>Fehle et al. (2023)</small>	1,512	5,617	4,032	628	957
GERestaurant <small>Hellwig et al. (2024)</small>	3,212	4,314	2,339	1,795	180
M-ABSA <small>Wu, Ma, Liu, et al. (2025)</small>	12,794	18,484	13,562	4,088	834
B2B Software Reviews <small>Fehle et al. (2025)</small>	1,500	3,479	1,539	1,886	54

Table 2.: Overview of the datasets, including the number of documents, aspect annotations, and sentiment distribution. Older datasets such as MLSA, PoTS, and Scare are excluded due to differences in the ABSA task definition.

tained across annotators. This is followed by an overview of annotation tools commonly used to support the annotation process. Finally, metrics for measuring inter-annotator agreement are discussed, highlighting methods for assessing consistency between annotators.

2.4.1. Guidelines

Following Klie et al. (2024), the establishment of clear annotation guidelines is essential to ensure consistency and reliability in annotating textual data. Well-designed guidelines provide a common framework that helps maintain a uniform annotation style and promotes agreement in how annotators interpret and label the text. Beyond consistency, guidelines significantly shape the annotation process itself: the way they are structured and formulated can influence annotators’ decisions and even introduce unintended biases. In practice, annotation guidelines are rarely static. They are often iteratively revised across multiple versions, for example during pilot studies or as new challenges emerge throughout the annotation process. Furthermore, guidelines may span several pages, offering detailed instructions, examples, and clarifications to cover a wide range of possible cases. In some settings, existing guidelines are reused and adapted to a new task, which accelerates the development process and provides a solid base (Klie et al., 2024).

The work by Pontiki et al. (2014) represents an important milestone in the devel-

opment of ABSA guidelines, as it was among the first to refine and consolidate existing definitions and methodologies for the ABSA task within the context of the SemEval workshop. The guidelines introduced a clear structure consisting of: a brief introduction to the annotation task, precise definitions of relevant elements such as aspect categories or opinion phrases, and examples with detailed instructions for handling different cases. This structure was largely maintained in the subsequent SemEval shared tasks from 2014 to 2016 (Pontiki et al., 2014, 2015, 2016).

A similar approach can be found in the GermEval 2017 guidelines, which likewise provide an introduction, definitions of annotation elements, and illustrative examples (Wojatzki et al., 2017). However, in contrast to SemEval, GermEval placed additional emphasis on practical aspects of the annotation process. The guidelines explicitly included instructions on how to use the annotation platform, offered more detailed rules for handling special cases, and considered language-specific phenomena relevant for German. These foundational guidelines have not only shaped the design of shared tasks but also served as a reference point for later annotation efforts. For instance, subsequent works such as Hellwig et al. (2024) and Fehle et al. (2025) adapted and extended the principles established in SemEval and GermEval to their own domains and the German language.

2.4.2. Annotation Tools

According to Colucci Cante et al. (2024), companies and public organizations are increasingly seeking to unlock the full potential of the vast amounts of information they collect. However, a significant portion of this information exists in unstructured formats, making it difficult to analyze and leverage effectively. To address this, natural language understanding (NLU) and NLP technologies have become essential tools to transform unstructured data into structured forms that are suitable for analysis and machine learning applications. As highlighted by Colucci Cante et al. (2024), the management of unstructured data remains one of the central challenges in the field of big data. To enable meaningful analysis, raw data must be systematically organized, with annotation remaining one of the most effective ap-

proaches to achieve this (Colucci Cante et al., 2024). Based on this premise, we analyzed all datasets referenced in the review by Chebolu et al. (2023), along with papers on German-language datasets, to determine which annotation tools were used. To the best of our knowledge, the work of Chebolu et al. (2023) represents the only comprehensive survey of ABSA datasets across English, making it a valuable foundation for our analysis. The results are summarized in Table 3.

Many of the reviewed papers did not specify which annotation tools were used, and in most cases, no visual materials such as interface screenshots were provided. Among those that did mention their tools, recent publications more commonly relied on INCEpTION or Label Studio, suggesting a growing preference for these platforms in current annotation practices.

Annotation Tool	Amount	References
BRAT	6	Pontiki et al. (2014, 2015, 2016); Sanger et al. (2016); Saeidi et al. (2016); Y. Li et al. (2023)
INCEpTION	3	Greve et al. (2021); Gabryszak & Thomas (2022); Fehle et al. (2023)
Label Studio	2	Hellwig et al. (2024); Fehle et al. (2025)
WebAnno	1	Wojatzki et al. (2017)
Knowtator (Protege)	1	Kessler et al. (2010)
MMAX2	1	Sidarenka (2016)
YEDDA	1	Cai et al. (2021)
In-House Tool	1	Regatte et al. (2020)
Not Mentioned	20	Hu & Liu (2004); Ding et al. (2008); H. Wang et al. (2011); Clematide et al. (2012); Steinberger et al. (2014); Dong et al. (2014); Q. Liu et al. (2015), Yin et al. (2017); Basile et al. (2018); de Frana Costa & da Silva (2018); Rahman & Kumar Dey (2018); Jiang et al. (2019); Fan et al. (2019), Peng et al. (2020); De Mattei et al. (2020); Hamborg et al. (2021); Bu et al. (2021); W. Zhang et al. (2021), T. Xu et al. (2023); Wu, Ma, Liu, et al. (2025)

Table 3.: Annotation tool usage frequency across reviewed corpora.

In the following section, the three most frequently used annotation tools are described in greater detail. **BRAT** (Stenetorp et al., 2012), short for brat rapid annotation tool, is a web-based platform designed for intuitive text annotation. It builds on the STAV text annotation visualizer and places particular emphasis on ease of use and clear visualization. It supports collaborative annotation through a client-server

architecture and can incorporate machine learning or statistical methods to assist annotators. *BRAT* was especially prominent in the early phases of ABSA corpus development. Another widely adopted tool is **INCEpTION** (Klie et al., 2018), an open-source platform for semantic and interactive annotation. It features a modular architecture and provides machine learning-based assistance to support both the annotation process and project management. In addition, **Label Studio**² is a flexible, open-source data labeling platform designed to accommodate a wide range of annotation tasks, data formats, and user roles. It also supports the integration of machine learning models to generate label predictions, helping streamline and accelerate annotation workflows.

Colucci Cante et al. (2024) provides a comparative analysis of a wide range of semantic and non-semantic textual annotation tools, focusing on core functionalities essential for complex annotation tasks. Label Studio, WebAnno, and Labelbox are identified as the most comprehensive tools, as they all support features such as multilabeling, annotation suggestions, relation annotation, customizable labels, and collaborative workflows. These shared capabilities make them well-suited for a broad range of annotation scenarios, like ABSA.

2.4.3. Inter-Annotator Agreement Metrics

Inter-annotator agreement (IAA) is a central concept in annotation studies, as it provides a measure of consistency between annotators and serves as an indicator of annotation quality (Klie et al., 2024). Assessing agreement is important for quality estimation, ensuring that annotations are reliable and reproducible. Common approaches include manual inspection of a subset or the full dataset, as well as the use of control instances, where examples with known ground truth are injected to verify annotator performance.

A common way to quantify the reliability of annotations and annotators is to compute their IAA (Klie et al., 2024). Several agreement measures exist, each with its advantages, limitations, and prerequisites:

²Label Studio:<https://labelstud.io/>

Percent Agreement Measures the percentage of units on which two annotators agree. While simple to compute, it is sensitive to imbalanced datasets, does not account for chance agreement, and can be difficult to compare across different annotation schemas.

Cohen’s Kappa (κ) Introduces chance-correction and is suitable for two annotators assigning categorical labels. All instances must be annotated by both annotators, and missing entries are not allowed (Cohen, 1960). The formula is:

$$\kappa_C = \frac{p_o - p_e}{1 - p_e} \quad (2.1)$$

Fleiss’ Kappa (κ) Extends Scott’s π (Scott, 1955) to multiple annotators. Each instance must be labeled by the same annotators, and annotators are assumed to be sampled randomly (Fleiss, 1971). Its formula is:

$$\kappa_F = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (2.2)$$

Krippendorff’s Alpha (α) Based on the ratio of observed disagreement to expected disagreement. It is more versatile than Fleiss’ κ , handling multiple annotators, missing data, and various data types, including categorical, ordinal, hierarchical, and continuous data (Krippendorff, 2011). Its formula is:

$$\alpha = 1 - \frac{D_o}{D_e} \quad (2.3)$$

Correlation Measures Applicable when instances are scored on numerical, continuous, or discrete scales, such as Likert scales. However, using correlation coefficients as an agreement measure is controversial, as they capture covariation rather than true agreement. In other words, the indicators suggest the tendency for two variables to co-move, but do not provide a definitive assessment of the similarity between the assigned scores (Fisher, 1992).

Classification Metrics Often used for sequence labeling tasks like named entity recognition, providing pairwise agreement measures. However, they are typically not chance-corrected, may not be symmetric, and averaging can result in information loss. Measuring agreement in ABSA tasks is particularly challenging due to the need for span extraction. As noted by Chebolu et al. (2023), Kappa may not be well-suited for span-based annotations, since it requires counting negative cases, which is not straightforward for sequences of words. Instead, some studies report F1-scores to assess annotation consistency (Pontiki et al., 2016; Chebolu et al., 2023). Overall, agreement metrics are an essential part of ensuring high-quality annotation, but they are not sufficient on their own. High agreement does not automatically guarantee high-quality labels. Complementary methods, such as manual inspection or other quality-control measures, are necessary. Additionally, relying on a single agreement coefficient is generally insufficient for a robust evaluation of annotation reliability (Klie et al., 2024).

2.5. Approaches to Aspect-Based Sentiment Analysis

This chapter provides an overview of the main approaches to ABSA as reported in the literature. It begins with lexicon-based methods, which rely on manually or semi-automatically constructed sentiment dictionaries and rule-based techniques to determine aspect-level sentiment. Next, traditional machine learning approaches are discussed, including feature-based classifiers and conventional supervised learning techniques. The chapter then covers deep learning and large language model (LLM) approaches, which have become dominant in recent years and are capable of capturing complex patterns in text. Throughout the sections on lexicon-based, traditional machine learning, and deep learning approaches, three recent survey papers (Sadia et al., 2018; Chauhan et al., 2023; Wankhade et al., 2024) are used as references to summarize the state of the art. Finally, recent trends in ABSA research are highlighted, with a particular focus on the increasing use of large language models and the exploration of more fine-grained task formulations.

2.5.1. Rule-based & Lexicon-based Approaches

At the beginning of ABSA research, rule-based and lexicon-based methods played a central role, relying on handcrafted resources and linguistic heuristics rather than learned statistical models.

Lexicon-based Methods Lexicon-based approaches determine sentiment by relying on pre-defined sentiment lexicons that assign polarity values to words or phrases (Sadia et al., 2018). They can be broadly divided into two categories: dictionary-based methods, which use lexical resources such as WordNet, and corpus-based methods, which derive sentiment information from corpora using statistical or semantic techniques. Early work in ABSA has also applied lexicon-based strategies. For instance, Wogenstein et al. (2013) constructed a phrase-based opinion lexicon for German to detect strong positive and negative expressions concerning products and services in the insurance domain, linking opinion-bearing phrases to their associated aspects. Similarly, Wiegand et al. (2014) developed a rule-based system that relies on a predicate lexicon with extraction rules for verbs, nouns, and adjectives, combined with linguistic preprocessing steps such as named-entity recognition and syntactic parsing.

Syntax-based Methods Another important line of research relies on syntactical relations rather than raw frequency counts. According to Chauhan et al. (2023); Hua et al. (2024), syntax-based techniques exploit grammatical dependencies to extract aspect terms while efficiently pruning away low-frequency candidates. For instance, L. Zhang et al. (2010) propose a double propagation approach, which iteratively extracts aspects and opinion words through dependency relations. To improve its precision and recall on both small and large corpora, the authors enhance the method with part-whole patterns and negation handling, followed by feature ranking to refine the set of aspect candidates.

2.5.2. Traditional Machine Learning Approaches

Before the rise of neural architectures, ABSA often relied on manually designed features and classical classifiers. These approaches typically combined linguistic preprocessing and feature engineering with established machine learning algorithms.

Linear Classifiers Linear classifiers represent the decision function as a linear predictor, where features such as word frequencies are combined with learned coefficients to separate sentiment classes (Chauhan et al., 2023). For example, Zainuddin et al. (2016) proposed a hybrid method for Twitter ABSA that integrates Principal Component Analysis (PCA) for feature selection with SentiWordNet lexicon-based features, feeding the resulting representation into a Support Vector Machine classifier.

Decision Tree Classifiers Decision trees recursively partition the feature space into hierarchical rules to classify sentiment associated with aspects (Chauhan et al., 2023). Bhoi & Joshi (2018) explored various preprocessing strategies and features in combination with decision trees and other classifiers, highlighting their usage for aspect-level polarity classification.

Probabilistic Classifiers Probabilistic classifiers, such as Naïve Bayes, assume a generative process for the data and estimate class probabilities based on the distribution of features (Chauhan et al., 2023). For instance, Mubarak et al. (2017) proposed a pipeline involving preprocessing, feature selection with Chi-Square statistics, and final sentiment classification with Naïve Bayes for aspect-level analysis.

2.5.3. Deep Learning & Large Language Models

Neural Networks Neural network-based models have become a central approach for aspect-based sentiment analysis, capable of learning complex patterns from text. Recurrent neural networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants, are feed-forward networks designed to analyze sequences such as sentences or time series (Wankhade et al., 2024). For

instance, Cheng et al. (2017) propose a Hierarchical Attention (HEAT) network for aspect-level sentiment classification, which uses aspect attention to guide sentiment attention and thereby better capture aspect-specific sentiment words.

Recursive neural networks (RcNNs) leverage syntactic tree structures to learn compositional semantics (Wankhade et al., 2024). Dong et al. (2014) introduce an Adaptive Recursive Neural Network (AdaRNN) for target-dependent sentiment classification on Twitter, utilizing dependency parsing to identify words syntactically connected to the target.

Convolutional neural networks (CNNs) are also employed for ABSA, originally popular in image processing but effective in capturing local n-gram features in text (Wankhade et al., 2024). Pham & Le (2018) propose a joint model of multiple CNNs, each handling a different representation of the input, including word embeddings from Word2Vec and GloVe, as well as one-hot character vectors.

M. Zhang & Qian (2020) introduce an architecture that convolves over hierarchical syntactic and lexical graphs. They employ a global lexical graph to encode corpus-level word co-occurrence information, construct concept hierarchies to differentiate types of dependencies or lexical pairs, and design a bi-level interactive graph convolution network (BiGCN) to exploit both graphs fully.

Transformer-based Methods Transformer-based architectures have become a strong paradigm in ABSA due to their ability to model long-range dependencies and contextual information (Wankhade et al., 2024). Attention mechanisms, originally developed to focus on relevant parts of the input sequence, allow the model to assign different weights to tokens depending on their importance for the task. The Transformer model (Vaswani et al., 2017) consists of stacked encoder and decoder layers, where self-attention in the encoder generates contextualized representations that are subsequently processed by feed-forward layers with linear and softmax operations.

Building on this idea, X. Li et al. (2019) investigate the use of pretrained contextualized embeddings, such as BERT, for end-to-end ABSA. In their approach, BERT generates contextualized token representations that are fed through addi-

tional transformer layers and a task-specific prediction layer to output the aspect and sentiment tag sequence.

Large Language Models Large language models (LLMs) are transformer-based models with tens to hundreds of billions of parameters, pretrained on massive text corpora. They exhibit strong language understanding and generation abilities, including novel capabilities not found in smaller models (Minaee et al., 2024). LLMs have recently become a dominant approach in aspect-based sentiment analysis, offering strong performance in zero-shot, few-shot, and fine-tuned settings. In zero-shot evaluations, Mughal et al. (2024) assess models such as ATAE-LSTM, Flan-T5-Large-ABSA, DeBERTa, PaLM, and GPT-3.5-Turbo across a diverse set of datasets, including DOTSA, MAMS, and SemEval16. For few-shot scenarios, Z. Wang et al. (2023) propose FS-ABSA, which frames end-to-end ABSA as a text-infilling task combined with domain-adaptive pre-training and fine-tuning.

Wu, Ma, Zhang, et al. (2025) evaluated LLMs under zero-shot conditions using multiple strategies, including vanilla zero-shot, chain-of-thought, self-improvement, self-debate, and self-consistency, across nine models and five languages (EN, ES, FR, NL, RU). Their results suggest that while LLMs show promise for multilingual ABSA, they generally underperform compared to fine-tuned, task-specific models. Simmering & Huoviala (2023) further explore zero-shot, few-shot, and fine-tuned setups using GPT-4 and GPT-3.5, experimenting with different prompting strategies such as guideline summaries, roleplay, reference prompts, and minimal instructions. Finally, Zhou et al. (2024) investigate instruction-based multi-task fine-tuning for LLMs, alongside three demonstration selection strategies to enhance few-shot performance. Across 13 datasets, eight ABSA subtasks, and six LLMs.

2.5.4. Recent Trends in ABSA Research

To gain an overview of current trends, we surveyed papers published at ACL, EACL, EMNLP, and NAACL in the last two years containing the topic ABSA. In total, 63 papers were identified, of which 8 were excluded because they did not focus on text-based ABSA (e.g., dialogue or multimodal tasks). The majority of works consider

datasets in English or Chinese, with only three addressing other languages. Commonly investigated subtasks include ASQP and ACOS, while some studies focus on ASTE. In contrast, the tasks of TASD and ACSA remains relatively underexplored.

The following recent work has examined the task of TASD, with several papers proposing new frameworks and evaluations, particularly on the *Rest15* and *Rest16* datasets in English. Šmíd et al. (2024) evaluate the performance of open-source LLaMA-based models fine-tuned for ABSA. Across four tasks and eight English datasets, the fine-tuned Orca 2 model outperforms SOTA methods, including MvP and Paraphrase, particularly for TASD on the *Rest15* and *Rest16* datasets. However, performance in zero- and few-shot settings remains limited compared to fully fine-tuned approaches. Lv et al. (2023) propose Efficient Hybrid Generation, a framework designed for improved TASD performance. On English datasets (*Rest15* and *Rest16*), the EHG-Para variant achieves superior results compared to GAS-Para and GAS-R, demonstrating the potential of generation-based approaches for ABSA. Bai et al. (2024) introduce ChatABSA, a framework to assess the performance of large language models in ABSA tasks. Experiments on English datasets show that, in few-shot settings (FS-10), ChatABSA underperforms compared to supervised baselines and SOTA models such as Hier-GCN, AAGCN, MvP, and GAS.

Beyond TASD, recent research has also concentrated on the tasks of ASQP and ACOS. Several studies have introduced new methods and evaluation strategies on benchmark datasets such as *Rest15*, *Rest16*, and *Laptop*, often in comparison with established models like MvP or with large language models in zero- and few-shot settings. Hellwig et al. (2025) investigate the capabilities of LLMs for zero- and few-shot learning in ASQP across five datasets, including *Rest15*, *Rest16*, *FlightABSA*, *OATS Coursera*, and *OATS Hotels*. Using Gemma 3 (4B and 27B) with up to 50 examples, they report that in the 20-shot setting on *Rest16*, LLMs achieve an F1-score of 51.54, compared to 60.39 by the fine-tuned MvP model. Jun & Lee (2025) propose Dynamic Order Template (DOT), a method for ACOS and ASQP that dynamically generates order templates for each instance. Their two-stage framework first predicts the number of sentiment tuples before refining the template for final pre-

diction. On five datasets (*Rest15*, *Rest16*, *Laptop*, *Restaurant*, *MEMD*), DOT achieves 51.91 F1-score for ASQP on *Rest15*. Yang et al. (2025) introduce a fully automated pipeline to expand evaluation sets by adding alternative valid terms for aspects and opinions. This approach, termed ZOOM IN-N-OUT, accommodates multiple candidate answers and improves agreement in human evaluation. On datasets including *Rest15*, *Rest16*, and *Laptop*, they report an improvement of 5.29% for ASQP on *Rest15* with a fully fine-tuned MvP model (57.03 F1). W. Zhang et al. (2024) provide a broad investigation into the capabilities of LLMs for sentiment analysis, spanning conventional classification, ABSA, and more fine-grained tasks. On *Rest15*, ChatGPT achieves an F1-score of 35.54 in a 10-shot ASQP setting, illustrating the challenges LLMs still face in comparison to specialized architectures.

The DimABSA shared task, introduced at SIGHAN 2024, focused on Chinese restaurant reviews and defined three subtasks: intensity prediction in the valence-arousal dimensions, triplet extraction of aspect, opinion phrase, and intensity, and quadruple extraction adding the aspect category. In total, 214 submissions from 61 teams were received (Lee et al., 2024). H. Xu et al. (2024) presented the winning system, which integrates BERT with large language models to combine their strengths. Their approach achieved a 41.7 F1-score on quadruple extraction, with BERT showing stronger performance for continuous intensity prediction and aspect-opinion extraction, while LLMs proved more effective for integer-level intensity prediction. Zhu et al. (2024) proposed a coarse-to-fine in-context learning approach based on the Baichuan2-27B model. By using a two-stage prediction process, they obtained a 37.6 F1-score in quadruple extraction. Tong & Wei (2024) introduced CL-Span, a contrastive learning-enhanced span-based framework. The method employs a modular pipeline for aspect-opinion pairing, sentiment scoring, and category prediction, achieving 38.9 F1-score in quadruple extraction.

2.6. ABSA Evaluation Metrics

In order to evaluate model performance in ABSA, several standard classification metrics are commonly applied. The most frequently used are accuracy, precision,

recall, and the F1-score (Wankhade et al., 2024). We also looked at alternative measures, such as Macro-Averaged Mean Absolute Error (MAE) and Ranking Loss. MAE is recommended for tasks involving continuous sentiment scores, as it directly captures prediction errors independent of dataset balance (Wankhade et al., 2024). Ranking Loss penalizes small and large errors in a more balanced way (Brauwers & Frasincar, 2022).

Accuracy Accuracy measures the proportion of correctly classified instances out of all instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.4)$$

where TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) represent the counts of classification outcomes.

Precision Precision quantifies how many of the instances predicted as positive are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.5)$$

Recall Recall measures how many of the actual positive instances were correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.6)$$

F1-Score The F1-score is the harmonic mean of precision and recall. It provides a single measure that balances both:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.7)$$

In multi-class settings, F1 can be averaged across classes in different ways. The *macro-averaged* F1 computes the F1-score for each class and then averages them, giving each class equal weight regardless of its frequency. The *micro-averaged* F1 considers all instances together and computes global counts of TP , FP , and FN , without distinguishing between classes (Grandini et al., 2020).

Macro-Averaged Mean Absolute Error (MAE) MAE measures the average absolute difference between predicted and true labels. It is more robust to class imbalance and is often used for continuous sentiment scores (Wankhade et al., 2024). The macro-averaged variant is defined as:

$$MAE(y, y') = \frac{1}{p} \sum_{j=1}^p \frac{1}{|y_j|} \sum_{y_i \in (y_j)} |y_i - y'_i| \quad (2.8)$$

Ranking Loss Ranking loss is closely related to the MAE but penalizes small and large errors in a more equal manner (Wankhade et al., 2024; Brauwiers & Frasincar, 2022). It is defined as:

$$RankingLoss = \sum_{i=1}^q \frac{|y_i - y'_i|}{p \cdot q} \quad (2.9)$$

2.7. Summary

This chapter reviewed the current state of research in ABSA. We began with a brief introduction to SA and ABSA, covering task definitions, subtasks, and joint formulations. An overview of available German ABSA datasets was provided, along with annotation guidelines, tools, and common IAA metrics such as Krippendorff’s alpha and Cohen’s kappa. The review highlighted that only a small number of German datasets exist, some of which are outdated or not publicly accessible. Subsequently, the main approaches to ABSA were discussed, ranging from lexicon-based methods to traditional machine learning and neural models, up to LLMs. Recent trends emphasizing the use of LLMs and more fine-grained task formulations were also highlighted. Finally, evaluation metrics such as F1-score and accuracy were introduced to contextualize results reported in the literature.

Building on this foundation, the purpose of this work is to systematically investigate how different annotation strategies influence the development of ABSA in the German language. While SA and ABSA are well established in English, research on German remains limited, and the availability of suitable training data is scarce. Creating reliable annotated datasets is therefore a crucial step toward enabling accurate machine learning models. However, annotation quality may vary consider-

ably depending on who performs the annotation and under which conditions. This thesis addresses this gap by comparing annotations generated by experts, students, crowdworkers, and LLMs, using separate expert annotations as ground truth. The study focuses on identifying the strengths and weaknesses of each approach and analyzing how annotation quality affects downstream model performance on two ABSA tasks: ACSA and TASD.

3. Methodology

In this section, the methodological framework of the thesis is presented. First, the setup of the label interface used for data annotation is described, followed by the guidelines that ensured consistency across annotators. Subsequently, five different annotation strategies are discussed, namely expert annotation, student annotation, crowdsourcing, and the use of large language models in few-shot settings. The procedure for generating the expert-based ground truth dataset, which serves as the reference point for evaluation, is then outlined. Finally, the chapter concludes with an overview of the methods applied and a summary highlighting the key points.

3.1. Annotation Setup

In this section, we present the annotation setup used to create the datasets. We describe the annotation interface that guided the annotators through the tasks, as well as the detailed annotation guidelines that ensured consistency and clarity. An example sentence for each category can be viewed in the Appendix A.3.

3.1.1. Interface

The annotation process was conducted using the web-based tool Label Studio³ (utilized in the Academic Version 2.26.0.dev11), which was customized to support ABSA. Figure 2 shows the interface used by the annotators to perform the TASD task (the interface for the ACSA task can be found in the Appendix A.7). The sentence to be annotated appears at the top of the interface, in this case: *"Diese sind sehr schmackhaft und die Portionen sind großzügig."* ("These are very tasty and the portions are generous."). Below the sentence, the interface presents a clearly structured labeling panel organized by aspect categories: `ESSEN (FOOD)`, `SERVICE (SERVICE)`, `AMBI-`

³Label Studio: <https://labelstud.io/>

3. Methodology

ENTE (AMBIENCE), GESAMTEINDRUCK (GENERAL) and PREIS (PRICE). Each category includes four polarity options: `Positiv` (positive), `Negativ` (negative), `Neutral` (neutral) and `Konflikt` (conflict). The polarities are represented as colored buttons with associated keyboard shortcuts to facilitate efficient annotation.

Additionally, an `Implizit` (implicit) label is available to mark implicit aspects. This label must be used in conjunction with an existing aspect label and cannot be assigned on its own. A warning is provided to prevent invalid combinations that may lead to tool instability.⁴

At the bottom of the interface, annotators can mark metadata related to the annotation task, such as whether the sentence was difficult to annotate or whether a lack of context may have influenced their decision. A comment field is available for free-text input, allowing annotators to justify or explain decisions where necessary. The current version of the annotation guidelines is also linked at the bottom of the interface for quick reference.

Diese sind sehr schmackhaft und die Portionen sind großzügig.

Aspekt-Label
Verwende die folgenden Labels, um Aspekte mit ihrer jeweiligen Kategorie und Polarität zu markieren.

Essen Essen-Positiv 1 Essen-Negativ 2 Essen-Neutral 3 Essen-Konflikt y

Service Service-Positiv 4 Service-Negativ 5 Service-Neutral 6 Service-Konflikt x

Ambiente Ambiente-Positiv 7 Ambiente-Negativ 8 Ambiente-Neutral 9 Ambiente-Konflikt c

Gesamteindruck Gesamteindruck-Positiv q Gesamteindruck-Negativ w Gesamteindruck-Neutral a Gesamteindruck-Konflikt v

Preis Preis-Positiv s Preis-Negativ z Preis-Neutral d Preis-Konflikt b

Implizite Aspektennung
Das "implizit"-Label darf nur ergänzend zu einem bereits bestehenden Aspekt-Label angewendet werden (es darf also nicht alleine stehen). Zwei Labels gleichzeitig neu zu vergeben kann zu einem Absturz von Label Studio führen. In diesem Fall bitte die Seite neu laden.

☐ Implizit: o

Metadaten

☐ Die Annotation war schwierig.³¹

☐ Fehlender Kontext aus der Bewertung könnte die Annotation beeinflussen.³¹

Die Annotationsanleitung kann hier abgerufen werden: <https://drive.google.com/file/d/1uZR-cH1TVARBBj7n6uXc-R-cEIrUzmt/view?usp=sharing>

Figure 2: Label interface for the TASD task in Label Studio.

⁴In our experiments, we observed that assigning two labels simultaneously to the same text span may cause the interface to disappear, which seems to be a bug in Label Studio.

3.1.2. Guidelines

Two separate sets of annotation guidelines⁵ were developed for the TASD and ACSA tasks as an adaptation of the original annotation guidelines proposed by Hellwig et al. (2025).

The TASD guidelines begin with an brief introduction to ABSA, followed by precise definitions of the aspect categories and their scope. Polarity is then defined in terms of four labels (*positive*, *negative*, *neutral*, *conflict*), alongside a description of the aspect phrase, distinguishing between explicit and implicit mentions and specifying how implicit aspects should be marked in Label Studio. Chapter two provides a detailed specification of aspect phrases in 13 points, addressing issues such as the requirement that sentiment must be expressed towards the phrase and that only the first occurrence of an aspect phrase should be annotated. Chapter three illustrates the annotation of category and polarity through examples, presenting complete triplets for selected sentences. The final chapter offers a step-by-step guide for using Label Studio, including annotated screenshots of the interface, explanations of meta-tags, and instructions for when to apply them (e.g., free-text comments or highlighting difficult annotation decisions).

The ACSA guidelines follow the same structure but omit the detailed chapter on aspect phrases and instead adapt the interface description to the requirements of the ACSA task. Both sets of guidelines were developed with the established ABSA annotation principles of Pontiki et al. (2015) in mind. To illustrate the annotation process, the guidelines also include concrete examples in German (see Figure 3).

3. Beispiel: Es war teuer, das **Menü** jedoch großartig.
1. Aspekt: **Preis-Negativ**, **Implizit** [irgendeine Phrase aus dem Text, wie z.B. "war"]
 2. Aspekt: **Essen-Positiv** **Menü**

Die Aspekt-Kategorie "Preis" wird nur implizit adressiert - es gibt keinen Aspektbegriff. Für die Annotation muss eine beliebige Phrase im Satz ausgewählt werden; welche genau, ist dabei nicht relevant.

Figure 3: Example for the category and polarity description in the TASD guidelines.

⁵Annotation guidelines: https://github.com/ValdrDarmir/Annotation-Quality-and-Its-Influence-on-ABSA-A-Case-Study-on-German-Restaurant-Reviews/tree/main/04_annotations/Guidelines

3.2. Annotation Strategies

In this section, we describe the five annotation strategies employed to construct datasets for ABSA. The first four are presented below:

- Crowdworkers
- Students
- LLMs
- Experts

The fifth strategy, the independent expert-based ground truth, is discussed in a separate section, as it constitutes a special case that serves as the evaluation benchmark. This ground truth serves as a benchmark for comparing the consistency, reliability, and accuracy of the different annotation strategies.

3.2.1. Ground Truth

The annotation process for the dataset was conducted in accordance with the guidelines proposed by Klie et al. (2024). To ensure systematic and incremental quality control, the dataset comprising 924 sentences was divided into five consecutive batches, each consisting of approximately 185 texts. After the completion of each batch, we computed the IAA using the micro-average F1 score, as recommended by Chebolu et al. (2023).

Annotations were independently carried out by two annotators, both of whom were Master’s students in Media Informatics with prior experience in ABSA. The annotation results were compared using a shared spreadsheet, in which matches and mismatches between the two annotators were manually tracked and reviewed. Discrepancies were analyzed collaboratively, and ambiguous cases (where a consensus could not easily be reached) were flagged.

After each batch, the annotation guidelines were iteratively refined to address recurring sources of disagreement and to improve annotation consistency. These refinements were based on patterns observed during the evaluation of mismatches and aimed to reduce ambiguity and promote uniform interpretation of the guide-

lines across batches.

The following adjustments were made after each batch:

- **Batch 1:** Annotation scope was expanded to include job titles along with names (e.g., “waiter Frank” instead of just “Frank”). Additionally, the term “menu” was explicitly categorized under the `Food` category to ensure consistent labeling.
- **Batch 2:** The phrase “Quality” was excluded from being annotated as an aspect phrase due to its lack of sufficient contextual information (Note: This is changed back in the dataset version 1).
- **Batch 3:** Generic expressions such as “evening” or “visit” were excluded from annotation as aspect phrases, as they were found to be too vague and lacking in specificity.
- **Batch 4:** National food references were refined to include the country adjective (e.g., annotating “German food” instead of just “food”) for greater contextual clarity.
- **Batch 5:** Previously anonymized entities such as restaurant and location names (e.g., placeholders like `RESTAURANT_NAME` or `LOC`) were allowed as valid aspect phrases when they appeared with sufficient contextual relevance.

Following the first iteration of the annotation process and the incorporation of the adjustments described above, a second annotation round was carried out to ensure consistency across the entire dataset. In this phase, 48 sentences were revisited. All modifications concerned the aspect phrase, while polarity and category remained unchanged in every instance. Most changes were related to the length of the aspect phrase or to the inclusion of quality terms within the phrase. To implement these updates, one annotator systematically reviewed the full dataset to apply the revised guidelines and flag any deviations from the agreed annotation scheme. The second annotator then examined these changes, either accepting or rejecting the proposed modifications. In cases of disagreement, both annotators engaged in discussion to reach a mutually acceptable resolution. Subsequently, the second annotator con-

ducted an additional review of the dataset to identify any overlooked adjustments, which were again subject to validation by the first annotator, following the same process of acceptance, rejection, or discussion to ensure alignment.

3.2.2. Crowd

We recruited 30 participants via Prolific⁶, with each participant tasked to annotate 200 sentences. Recruitment was restricted to participants located in Germany, Austria, or Switzerland who were fluent in German and had at least secondary education. Participants who had previously taken part in our other annotation studies were excluded.

Participants were compensated at a rate of £9 per hour, following Prolific’s guidance (minimum £6 per hour) and labeled as “Good” by the platform. Each participant could submit only once, and extremely fast submissions were rejected based on Prolific’s automated checks. Before starting the annotation task, participants completed a brief questionnaire, which is described in Section 3.2.3. The annotation task was restricted to desktop computers to ensure consistency in display and interaction. The study description shown to participants on Prolific can be found in the Appendix A.2. The stated completion times reported by Prolific for the different annotation studies are summarized as follows. For the pilot study, times ranged from 1:15 to 2:20 hours, with an average of 1:43 hours and a maximum allowed time of 3:52 hours (same limits applied to the TASD task). For the ACSA study, completion times ranged from 0:37 to 2:18 hours, with an average of 1:17 hours and a maximum allowed time of 3:07 hours. For the TASD study, times ranged from 1:25 to 3:42 hours, with an average of 2:15 hours and a maximum allowed time of 3:52 hours.⁷ At the end of each study, participants received a completion code via the Google Forms questionnaire to confirm their participation.

Informed consent was obtained from all participants via a detailed checkbox procedure in Google Forms, with further information provided in a PDF document.

⁶Prolific: <https://www.prolific.com>

⁷Note: These times were reported by Prolific and do not necessarily reflect the actual time spent on the annotation task, as participants could take breaks or start a few minutes after accepting the study.

Annotations were collected using Label Studio, with participants working independently according to detailed instructions. These instructions consisted of a comprehensive guideline in PDF format, accompanied by a short video demonstrating the annotation process within Label Studio. The study description shown to participants on Prolific is provided in Appendix Figure 11. Further details on the annotators' prior experience are reported in Table 4.

Pilot Study

To ensure the feasibility of the annotation setup, we conducted a pilot study for the TASD task with five available slots. In total, 15 participants registered for the study. One participant timed out after annotating only a few sentences, while nine additional participants started the study but did not complete it: four dropped out without providing any annotations, and five began annotating but withdrew within the first ten sentences. Furthermore, one participant used an automatic translation tool within Label Studio, which resulted in translated text being stored in the annotation interface. This participant discontinued the task voluntarily after a few sentences.

3.2.3. Students

Students were recruited to perform the annotation either via the university's *Versuchspersonenstunden* forum or through personal contacts of the author. All annotators had a background in Media Informatics, either as current students or graduates. As compensation, each participant received 2 *Versuchspersonenstunden*.

Upon confirming participation, students were provided with a detailed guideline explaining the annotation procedure, which consisted of six steps:

1. Register at Label Studio.
2. Complete the pre-annotation questionnaire (see Section 3.2.3).
3. Read the annotation guidelines carefully.
4. Watch the annotation video, which explains the Label Studio interface and the correct procedure for labeling.

5. Annotate a batch of 200 texts.
6. Contact the author via email for any questions or problems encountered during the annotation.

The annotation process was conducted over a three-week period, from August 8 to September 3. In the ACSA task, two annotators misinterpreted the guidelines and assigned a polarity to every category for each text, rather than only to the categories actually mentioned. These annotations were retained and incorporated in the creation of the final dataset.

Each batch of 200 texts was annotated independently by three annotators. To generate the final datasets, a majority voting procedure was applied: a label was assigned only when at least two out of three annotators agreed exactly on the category, polarity, and phrase. After removing the conflict label, texts that no longer contained any annotations were still kept in the final dataset to preserve consistent dataset lengths across all training sets.

Experience Level	Students		Crowd		
	ACSA	TASD	ACSA	TASD Pilot	TASD
No experience	7	6	6	2	2
<10 hours	4	3	3	2	2
10–50 hours	2	2	2	0	4
>50 hours	0	2	3	1	1
Work in field	0	0	0	0	1
Don't know	0	0	1	0	0
Type of Experience	ACSA	TASD	ACSA	TASD Pilot	TASD
Text	4	2	0	1	8
Image	4	3	3	3	5
Audio	1	0	3	2	2
Video	0	0	4	3	2
Multimodal	1	0	1	1	1
Other	0	0	0	0	0

Table 4.: Overview of annotators' experience levels and types of annotation across the ACSA (n=15), TASD Pilot (n=5), and TASD (n=10) studies.

Questionnaire

Before beginning the annotation, all participants were required to complete a short questionnaire, which was distributed via Google Forms⁸. The questionnaire first provided information about the overall goal of the annotation study and outlined the annotation procedure. Participants were then asked to provide informed consent for data collection by checking a box confirming that their annotations could be stored and used for publication purposes.

Subsequently, participants were asked about their prior experience with data annotation.

3.2.4. LLM

Based on the methodology proposed by Hellwig et al. (2025), we adapted their approach for annotating the train set using a LLM. The LLM parameters used in our experiments were:

Model: Gemma-3-27B⁹

Temperature: 0.8

Few-Shots: 30

Seeds: 0, 1, 2, 3, 4

Context length: 4096

To determine the optimal number of few-shot examples, we used the micro-F1 scores reported by Hellwig et al. (2025) for each dataset. We then calculated the average micro-F1 score across all datasets, as well as specifically for the restaurant domain. As shown in Table 5, comparison of the aggregated micro-F1 scores indicates that 50 and 30 few-shot examples yield the highest performance. We chose 30 few-shot examples for the LLM dataset annotation, as this number provides a balance between expected model performance and practical constraints, minimizing the manual annotation effort required for the LLM approach.

⁸Google Forms: <https://docs.google.com/forms/>

⁹Gemma 3 27B: <https://ollama.com/library/gemma3:27b>

Few-Shots	F1 Overall Datasets	F1 Avg Rest15 + Rest16
0	39.23	37.94
10	55.96	60.61
20	58.44	63.44
30	60.60	<u>65.11</u>
40	<u>59.91</u>	64.62
50	60.60	65.33

Table 5.: Comparison of micro-F1 scores (Gemma 3 27B) across different few-shot counts (Hellwig et al., 2025). Best results are highlighted in bold; second-best results are underlined.

3.2.5. Experts

For the expert annotation phase, we built upon the original annotations provided by Hellwig et al. (2024). These initial labels were used as the foundation and subsequently transformed into the revised labeling schema defined by our new annotation interface. During this transformation, all aspect phrases were re-aligned with their corresponding tokens in the text. In cases where a phrase was marked as implicit in the original annotation, the first word in the sentence was selected as the anchor point for labeling.

The expert annotator conducting this phase is a PhD student with prior experience in annotating textual data and ABSA, and was also involved in the original annotation process as a reviewer. For the TASD dataset, the expert revisited the entire dataset, consisting of 1,000 texts, to check for annotation inconsistencies and to revise labels in accordance with the updated guidelines. The review was carried out using the Label Studio platform’s review feature, which allowed the annotator to accept or reject each annotation.

In instances where an annotation was rejected, the expert was required to provide a correction (e.g., modifying the phrase or label) and submit the updated annotation. For complex or ambiguous cases, the annotator could also apply metadata tags to mark examples as difficult or to indicate missing context. During the TASD expert review process, the annotator accepted 1,333 annotations, updated 92 annotations to better align with the revised guidelines, and removed 10 triplets.

Due to the limited availability of experts and the time-intensive nature of the task, the ACSA dataset did not undergo a expert review. Instead, aspect phrases were removed from the revisited T ASD dataset, and duplicate tuples were consolidated into a single entry to create the final ACSA dataset.

3.3. Methods

In this section, we present the models used to predict sentiment tuples in the ACSA task and sentiment triplets in the T ASD task. To cover a broad range of methodological approaches, we include classification-based architectures, text generation models, and LLMs. Due to the limited size of our datasets, it was not possible to construct a separate development set. Therefore, we adopted the hyperparameter settings suggested by Fehle et al. (2025), who applied the same models to the original GERestaurant dataset. The following subsections provide details on each group of models, and Table 6 summarizes the most important hyperparameters and configuration details.

3.3.1. Classification Models

In this subsection, we outline the parameter settings and training configurations for the classification-based approaches, namely BERT-CLF and Hier-GCN.

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

We fine-tuned the model for 40 epochs with a batch size of 16 and a learning rate of $2 \cdot 10^{-5}$, using the AdamW-optimizer with a binary cross entropy loss function. Sentences were tokenized using the `deepset/gbert-base` tokenizer via Hugging Face’s `AutoTokenizer` and truncated or padded to a maximum sequence length of 256 tokens.

Hier-GCN 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. Tokens are represented as nodes, and syntactic as well as semantic relations define the graph edges. The resulting graph representations are pooled and passed to a classifier to predict aspect-sentiment pairs in the ACSA task.

We fine-tuned the model for 40 epochs with a batch size of 8 and a learning rate of $5 \cdot 10^{-5}$, using the BERT-specific Adam optimizer (BertAdam) with cross-entropy loss. Tokenization was performed using the `deepset/gbert-base` tokenizer via Hugging Face’s `BertTokenizer` and sentences were truncated or padded to a maximum sequence length of 128 tokens.

3.3.2. Text generation Models

Here, we describe the experimental settings for the text generation-based approaches, specifically Multi-View Prompting (MvP) and Paraphrase. Both models treat sentiment triplet extraction as a sequence generation task rather than a classification problem.

Multi-View Prompting The MvP approach (Gou et al., 2023) formulates the TASD task as a text-to-text generation problem. Using `T5-base` as the underlying sequence-to-sequence model, the method generates aspect-category-polarity tuples directly from the input sentence by applying multiple prompt formulations (“views”). The outputs from all views are then aggregated via a majority voting strategy to produce the final set of predictions.

The different outputs are aggregated via majority voting, which helps reduce inconsistencies and improves robustness across diverse linguistic constructions.

We fine-tuned the model for 15 epochs with a batch size of 16 and a learning rate of $1 \cdot 10^{-4}$. Training was conducted using the AdamW optimizer with a cross-entropy loss objective. Input sentences were tokenized with the standard T5 tokenizer and truncated or padded to a maximum sequence length of 128 tokens.

Paraphrase The Paraphrase method (W. Zhang et al., 2021) treats the TASD task as a sequence-to-sequence generation problem. Using T5-base as the base model, the input sentence is reformulated into a natural-language template that explicitly encodes the target output structure. The model is trained to generate aspect–category–polarity triplets directly from this reformulated input. This template-based approach ensures that outputs follow a predefined format and reduces structural errors common in free-form generation.

We fine-tuned the model for 15 epochs with a batch size of 16 and a learning rate of $3 \cdot 10^{-4}$. Input sentences were processed with the standard T5-base tokenizer and truncated or padded to a maximum sequence length of 256 tokens.

3.3.3. Large Language Models

This subsection details the configurations for LLMs, focusing on few-shot prompting and instruction fine-tuned models. We describe the prompting strategies, parameter choices, and evaluation setup applied to both settings. For the few-shot prompting experiments, we used Gemma 3 27B, following the approach of Hellwig et al. (2025) and ensuring a consistent evaluation setup with the model used to generate the LLM dataset. For the instruction fine-tuned experiments, resource constraints prevented us from fine-tuning Gemma 3 27B, so we used LLaMA 8B instead, following the configuration and hyperparameters reported in Fehle et al. (2025), which had previously been applied to the German restaurant domain.

Few-Shot Prompting Few-shot prompting leverages LLMs via in-context learning to perform both ACSA and TASD tasks (Simmering & Huoviala, 2023). We use the Gemma 3 27B model and provide 50 annotated examples directly in the prompt. The prompt template, adapted from Gou et al. (2023), is translated into German and tailored to match the structure of each task (see Appendix A.1).

No parameter updates are performed during inference. The model generates aspect–sentiment pairs for ACSA or aspect–category–polarity triplets for TASD based solely on the examples provided in the prompt. Input sentences and few-shot examples were tokenized using the OlaMA tokenizer, and the total input was truncated

to fit within the model’s maximum context length of 4,096 tokens.

Instruction Fine-Tuning Instruction fine-tuning adapts a LLM to directly map input sentences to structured ABSA outputs (Šmíd et al., 2024). We use LLaMA 3.1 8B and fine-tune it 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.

The model was trained for 4 epochs with a batch size of 16 and a learning rate of $2 \cdot 10^{-4}$ using the Adam optimizer. The LLaMA 3.1 8B model was loaded from `meta-llama/Llama-3.1-8B` using `FastLLaMA`. Input sentences were tokenized with the model tokenizer and truncated or padded to a maximum sequence length of 2,048 tokens. During generation, a maximum of 200 new tokens was produced per prediction. Fine-tuning and inference were performed in 4-bit quantization for efficiency. Fine-tuning updates the model parameters, enabling task-specialized behavior, in contrast to prompting-based approaches that rely solely on in-context learning.

Method	Base-Model	Epochs	Batch Size	Learn Rate
BERT-CLF	gbert-base	40	16	$2 \cdot 10^{-5}$
Hier-GCN	gbert-base	40	8	$5 \cdot 10^{-5}$
Paraphrase	t5-base	15	16	$3 \cdot 10^{-4}$
MvP	t5-base	15	16	$1 \cdot 10^{-4}$
LLM-Few-Shot	Gemma 3 27B	–	–	–
LLM-Fine-Tune	LLaMA 3.1 8B	4	16	$2 \cdot 10^{-4}$

Table 6.: Model configurations used in this study. Values for epochs, batch size, and learning rate were chosen based on the paper by Fehle et al. (2025).

3.4. Summary

This chapter outlined the methodological framework employed in this work. We first described the annotation setup, including the interface and the guidelines designed to ensure consistency across annotators. Next, the different annotation strate-

gies were presented, covering expert annotators, students, crowdworkers, and the use of LLMs. Finally, the chapter introduced the models used for evaluation, including classification models, text generation models, and large language models. Overall, the methodological design establishes the conditions under which the influence of the different annotation styles on model performance and annotation quality can be systematically assessed, providing the groundwork for the subsequent analysis.

4. Results

This section summarizes the outcomes of the studies. The evaluation procedure is outlined, IAA for the different annotation tasks is reported, and the performance of the tested models together with the corresponding statistical tests is presented. In addition, the costs and required effort of the annotation studies are analyzed, providing a comprehensive perspective on the trade-offs associated with different annotation strategies. The section concludes with a summary of the main results.

4.1. Evaluation Procedure

To ensure the reliability and validity of our findings, we conducted a structured evaluation procedure comprising three components: assessment of annotation consistency, model performance evaluation, and statistical testing of performance differences.

4.1.1. Inter-Annotator Agreement

For the ACSA task, we evaluated IAA using multiple complementary measures. First, we computed the average pairwise micro-F1 across all annotator pairs, providing an overall estimate of agreement. Finally, we calculated Krippendorff’s alpha (Krippendorff, 2011), which accounts for chance agreement and is applicable to ordinal and nominal annotations, offering a robust alternative to F1-based measures. These complementary metrics allow for a nuanced assessment of annotation reliability across both extraction and classification tasks.

Following prior work (Chebolu et al., 2023; Pontiki et al., 2016), we chose the micro-F1 score as the IAA metric for the TASD task. Micro-F1 is particularly suitable for span extraction tasks such as aspect phrase identification, since it directly captures the overlap between annotated spans rather than relying on categorical agree-

ment. While traditional reliability coefficients such as Cohen’s κ or Krippendorff’s alpha are primarily designed for categorical labels and require averaging over annotator pairs (Klie et al., 2024), micro-F1 has been demonstrated to be an effective measure for phrase-level or named-entity annotations in prior studies (Deleger et al., 2012; Hripcsak & Rothschild, 2005).

In our setup, exact phrase matching was required for aspect phrases, and IAA was computed by treating one annotator’s annotations as the gold standard and the others’ annotations as predictions. For tasks with more than two annotators, we report the average agreement across all pairwise comparisons, along with the corresponding standard deviation. The resulting micro-F1 was reported as a proxy for consensus-based agreement.

4.1.2. Model Evaluation

Model performance was evaluated using micro-F1 and macro-F1 scores as the primary metrics. Since each model was trained and evaluated with five different random seeds, the reported results correspond to the average across five runs in order to reduce the influence of seed variability. Accuracy scores were also computed but are provided in the appendix A.5 for completeness. Due to resource constraints, we did not perform hyperparameter tuning with a development set. Instead, we adopted the model configurations as reported in Fehle et al. (2025).

4.1.3. Statistical Testing

To evaluate whether differences in annotation style (crowd, students, LLMs, experts) influenced model performance, we conducted a series of statistical analyses using the micro-F1 score. The analysis was performed in two steps. First, we aggregated each model’s performance across all datasets to assess whether the choice of dataset had a significant effect on overall model performance. Second, for each model separately, we assessed whether performance differences between datasets were statistically significant by aggregating results from five independent runs with random seeds.

To decide on the appropriate statistical procedure, we first assessed the distribution of the scores using the Shapiro–Wilk normality test (Shapiro & Wilk, 1965). If the assumption of normality was met, we applied a repeated-measures one-way ANOVA to test for overall differences between datasets, followed by pairwise paired t-tests (Student, 1908; Field et al., 2012) for post-hoc comparisons. If normality was violated, we instead used the non-parametric Friedman test (Friedman, 1937), followed by Wilcoxon signed-rank tests (Wilcoxon, 1992) for pairwise comparisons. To control for inflated Type I error rates due to multiple testing, all pairwise comparisons were corrected using the Holm–Bonferroni procedure (Holm, 1979).

System Specifications

The training and evaluation of the models were performed on a workstation with the following hardware and software configuration:

- **Operating System:** Ubuntu 22.04.5 LTS (x86_64)
- **CPU:** Intel Xeon W-2255 (20 cores) @ 4.50 GHz
- **GPU:** NVIDIA Quadro RTX 6000; 24 GB GDDR6
- **GPU Driver Version:** 560.35.03
- **CUDA Version:** 12.6
- **Memory:** 64 GB RAM; DDR4-3200 MHz

4.2. Annotation Evaluation

In this section, we evaluate the consistency of the annotation process across tasks and datasets. First, we examine IAA within the two annotation tasks, ACSA and TASD, to assess reliability at the task level. Second, we report the IAA obtained during the creation of the ground truth. Finally, we present a targeted comparison based on a shared sample of 100 sentences, where the ground truth annotations can be directly contrasted with those from the other datasets, allowing us to assess differences in annotation quality and style.

4.2.1. IAA for the ACSA Task

IAA is widely used to evaluate the reliability of annotations. Rather than serving as a guarantee that labels perfectly reflect the ground truth, IAA provides an indication of how consistently different annotators apply the same annotation guidelines. In this sense, it serves as a proxy for the reproducibility and stability of the annotation process (Klie et al., 2024; Monarch, 2021).

	Crowd		Students		LLMs	
	Pair-F1	α	Pair-F1	α	Pair-F1	α
Batch 1	66.75 \pm 11.31	54.36 \pm 30.14	85.11 \pm 1.43	71.91 \pm 22.77	98.12 \pm 0.55	97.82 \pm 2.58
Batch 2	84.57 \pm 1.43	60.43 \pm 36.86	81.55 \pm 1.01	60.11 \pm 29.68	96.46 \pm 1.10	93.28 \pm 10.57
Batch 3	78.94 \pm 2.19	55.88 \pm 34.13	50.65 \pm 25.84	37.42 \pm 48.93	96.23 \pm 1.49	84.30 \pm 26.51
Batch 4	84.24 \pm 1.60	64.67 \pm 29.96	52.03 \pm 27.54	38.65 \pm 48.27	97.32 \pm 1.27	96.16 \pm 5.06
Batch 5	83.54 \pm 2.64	67.90 \pm 27.89	81.56 \pm 1.51	60.75 \pm 31.32	97.86 \pm 0.69	94.82 \pm 6.57
Macro	79.61 \pm 7.54	60.65 \pm 5.72	70.18 \pm 17.27	53.77 \pm 15.12	97.20 \pm 0.83	93.28 \pm 5.29
Micro	78.95 \pm 2.27	60.99 \pm 28.47	63.38 \pm 16.75	49.61 \pm 39.48	97.20 \pm 0.88	94.27 \pm 4.77

Table 7.: Batch-wise evaluation metrics for the three datasets (ACSA). Values are per batch averages with the \pm standard deviation. At the bottom, average macro and micro scores across batches are provided. Note: The expert dataset includes only one annotator, so reliability metrics cannot be computed. Abbr.: Pair-F1 = average pairwise micro-F1, α = Krippendorff’s alpha.

For the ACSA task, IAA was computed using two complementary measures: the average pairwise micro-F1 between annotators (*Pair-F1*), and Krippendorff’s alpha (α). In the Crowd and Student datasets, each batch was annotated by three independent annotators. For the LLM dataset, five outputs per instance were generated using a temperature of 0.8, effectively treating them as five distinct annotators for IAA computation. For tasks with more than two annotators, the reported micro-F1 corresponds to the average over all possible pairwise comparisons, and the standard deviation is provided to indicate the variability across annotator pairs.

Table 7 presents IAA scores for the three datasets across batches. For the Crowd dataset, the average Pair-F1 is 79.61 \pm 7.54 and Krippendorff’s alpha is 60.65 \pm 5.72 (macro), and 78.95 \pm 2.27 and 60.99 \pm 28.47 (micro). The Student dataset achieves a Pair-F1 of 70.18 \pm 17.27 and alpha of 53.77 \pm 15.12 (macro), and 63.38 \pm 16.75 and 49.61 \pm 39.48 (micro). The LLM dataset reaches the highest agreement with a Pair-F1 of 97.20 \pm 0.83

and alpha of $93.28_{\pm 5.29}$ (macro), and $97.20_{\pm 0.88}$ and $94.27_{\pm 4.77}$ (micro).

4.2.2. IAA for the TASD Task

As with the ACSA task, we evaluated IAA for TASD to assess the consistency of annotations. The TASD task poses a greater challenge because it combines both aspect term extraction and sentiment classification, making exact agreement between annotators harder to achieve. To capture annotation reliability, we again report the average pairwise micro-F1 between annotators.

Table 8 presents IAA scores for the three datasets across batches. For the Crowd dataset, the macro Pair-F1 is $36.21_{\pm 16.80}$, and the micro Pair-F1 is $32.38_{\pm 10.18}$. Student annotators show higher agreement, with a macro Pair-F1 of $50.38_{\pm 9.07}$ and a micro Pair-F1 of $50.50_{\pm 5.84}$. The LLM dataset achieves the highest consistency, with a macro Pair-F1 of $90.17_{\pm 1.94}$ and a micro Pair-F1 of $90.22_{\pm 1.82}$.

Batch	Crowd	Students	LLMs
Batch 1	$44.47_{\pm 16.20}$	$41.29_{\pm 17.67}$	$90.20_{\pm 2.11}$
Batch 2	$61.55_{\pm 6.41}$	$63.81_{\pm 2.34}$	$87.66_{\pm 2.82}$
Batch 3	$28.78_{\pm 20.15}$	$45.85_{\pm 7.19}$	$90.74_{\pm 2.21}$
Batch 4	$19.91_{\pm 21.97}$	$45.71_{\pm 12.52}$	$89.30_{\pm 1.95}$
Batch 5	$26.33_{\pm 26.49}$	$55.26_{\pm 9.64}$	$92.95_{\pm 1.19}$
Macro	$36.21_{\pm 16.80}$	$50.38_{\pm 9.07}$	$90.17_{\pm 1.94}$
Micro	$32.38_{\pm 10.18}$	$50.50_{\pm 5.84}$	$90.22_{\pm 1.82}$

Table 8.: Batch-wise evaluation metrics for the three datasets (TASD). Average pairwise micro-F1 are per batches with the \pm standard deviation. At the bottom, average macro and micro scores across batches are provided. Note: The expert dataset includes only one annotator, so reliability metrics cannot be computed.

4.2.3. Reliability of Ground Truth Annotations

In contrast to the other datasets, the ground truth annotations were created by two expert annotators, following the procedure outlined in Section 3.2.1. During this process, the annotation guidelines were iteratively refined and improved.

Since only two annotators were involved, IAA was assessed using the average pairwise micro-F1 between them and Krippendorff’s alpha (additional for the ACSA

task). Table 9 presents IAA scores for both the ACSA and TASD tasks. For the ACSA task, which was derived from the same data by removing the aspect phrase before computing IAA, macro Pair-F1 was $87.20_{\pm 2.31}$ with Krippendorff’s alpha of $75.46_{\pm 7.50}$, and micro Pair-F1 was 87.22 with alpha $79.02_{\pm 13.98}$. For the TASD task, agreement reached an average macro Pair-F1 of $72.14_{\pm 5.54}$ and a micro Pair-F1 of 72.18. Since only two annotators were present for TASD (and ACSA), no standard deviation can be reported for the micro score.

Batch	ACSA		TASD
	Pair-F1	α	Pair-F1
Batch 1	83.93	$66.29_{\pm 31.15}$	63.33
Batch 2	88.38	$79.84_{\pm 25.41}$	70.25
Batch 3	89.66	$84.62_{\pm 15.47}$	75.78
Batch 4	88.25	$77.00_{\pm 24.68}$	74.41
Batch 5	85.78	$69.53_{\pm 34.28}$	76.95
Macro	$87.20_{\pm 2.31}$	$75.46_{\pm 7.50}$	$72.14_{\pm 5.54}$
Micro	87.22	$79.02_{\pm 13.98}$	72.18

Table 9.: Batch-wise evaluation metrics for the ground truth. Values are per batch averages with the \pm standard deviation. Abbr.: Pair-F1 = average pairwise micro-F1, α = Krippendorff’s alpha.

4.2.4. Cross-Dataset Comparison to the Ground Truth

To assess annotation consistency across datasets, we selected a shared subset of 100 texts from the training portion of each dataset. This subset allowed a direct comparison between the four datasets (Crowd, Students, LLMs, Experts) and a separate ground truth annotated by two independent expert annotators. For the ACSA task, both micro-F1 and Krippendorff’s alpha were computed, while for TASD only micro-F1 was evaluated. The ground truth values reflect the IAA directly between the two experts. The results are reported in Table 10, showing that micro-F1 for ACSA ranges from 89.68 to 91.87 and Krippendorff’s alpha from 75.42 to 81.76 across datasets, while TASD micro-F1 ranges from 61.22 to 81.88. Ground truth agreement is 87.50 (micro-F1) and 76.12 (alpha) for ACSA, and 65.38 (micro-F1) for TASD.

Dataset	ACSA		TASD
	Micro-F1	α	Micro-F1
Crowd	89.68	77.89	61.22
Students	91.87	81.76	71.49
LLM	90.76	76.38	77.44
Experts	91.87	75.42	81.88
Ground Truth	87.50	76.12	65.38

Table 10.: Comparison of IAA and model-independent metrics on predefined 100-text subsets drawn from each dataset’s training portion. For ACSA, both micro-F1 and Krippendorff’s alpha are reported. For TASD, micro-F1 is reported. Ground truth values correspond to agreement between two expert annotators.

4.3. Evaluation of Model Performance and Statistical Testing

This section presents the evaluation of models trained on the different datasets. The primary focus is on micro- and macro-F1 scores to assess model performance, while accuracy scores for all models and datasets are provided in the Appendix A.5 for completeness. In addition, Appendix A.6 reports the average training time and memory usage of the models. For the ACSA task, the models evaluated are *BERT-CLF* and *Hier-GCN*. For the TASD task, the models are *Paraphrase* and *MvP*. In addition, the LLMs *Gemma FS* (LLM-FS) and *LLaMA-FT* (LLM-FT) were used for both tasks.

The section is organized into two parts: first, the evaluation results for the ACSA task are presented, followed by the performance analysis for the TASD task. Statistical testing was conducted in two stages. For each task, we first assessed whether performance differed significantly between datasets across all models. We then analyzed each model individually to determine whether its performance was influenced by the dataset used.

4.3.1. Performance for the ACSA Task

The performance of the models on the ACSA task is summarized in Table 11. Across all datasets, the results show that models trained on expert-annotated data consistently achieve the highest micro- and macro-F1 scores, although differences to

other datasets are often small. For *BERT-CLF*, the best performance is reached on the expert dataset with a micro-F1 of 78.26 and a macro-F1 of 75.17. *Hier-GCN* also performs strongest on the expert data, with scores of 79.78 (micro-F1) and 78.13 (macro-F1). *Gemma FS* achieves its highest results with 86.43 (micro-F1) and 84.50 (macro-F1) on the students dataset. *LLaMA FT* obtains its best performance on the expert data with a micro-F1 of 86.39 and a macro-F1 of 83.12. Overall, scores are very similar across datasets, with the expert dataset showing slightly higher values. Depending on the model, however, one of the crowd-, student-, or LLM-annotated datasets may occasionally outperform the other two by a small margin.

Aspect Category Sentiment Analysis (ACSA)								
Dataset	Crowd		Students		LLMs		Experts	
Method	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
BERT-CLF	76.99	74.04	77.81	74.75	77.44	74.65	78.26	75.17
Hier-GCN	79.66	77.28	78.97	77.53	79.13	77.17	79.78	78.13
Gemma FS	86.03	84.63	86.43	84.50	85.60	83.83	86.29	84.62
LLaMA FT	85.64	83.46	85.71	84.33	84.85	82.02	86.39	83.12

Table 11.: Micro- and macro-F1 scores for ACSA, averaged over five seeds across datasets. Highest values are shown in bold.

4.3.2. Performance for the TASD Task

Table 12 reports the micro- and macro-F1 scores for the TASD task across all four datasets. Similar to the ACSA task, the expert-annotated dataset yields the strongest results overall, with the highest scores observed across most models. For the *Paraphrase model*, the best performance is achieved on the expert dataset with a micro-F1 of 61.65 and a macro-F1 of 56.24. *MvP* shows a similar trend, reaching 64.01 (micro-F1) and 59.42 (macro-F1) on expert data. The *Gemma FS* model performs strongest on the LLM dataset with a micro-F1 of 65.58 and a macro-F1 of 62.30. Finally, the *LLaMA FT* model achieves the overall best results, with its highest scores of 71.47 (micro-F1) and 67.40 (macro-F1) on the expert dataset.

Across models, the crowd dataset typically results in the lowest performance, while student and LLM datasets perform comparably, occasionally surpassing each

Target Aspect Sentiment Detection (TASD)								
Dataset	Crowd		Students		LLMs		Experts	
Method	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Paraphrase	52.77	47.97	57.33	52.10	57.37	53.06	61.65	56.24
MvP	51.29	50.58	56.83	55.89	60.65	56.95	64.01	59.42
Gemma FS	58.56	56.41	62.28	59.06	65.58	62.30	63.38	58.74
LLaMA FT	65.46	60.97	69.33	65.50	66.24	63.20	71.47	67.40

Table 12.: Micro- and macro-F1 scores for TASD, averaged over five seeds across datasets. Bold indicates the highest values.

other for specific models. Expert annotations consistently lead to the best results.

4.3.3. Statistical Evaluation of the ACSA Task

To assess whether differences in annotation style (Crowd, Students, LLMs, Experts) had a measurable influence on model performance, we conducted a series of statistical analyses on the micro-F1 scores. For each model, performance was obtained from five independent runs with different random seeds.

Overall Comparison Across Datasets

The Shapiro–Wilk normality test indicated that the model-averaged scores were not normally distributed ($W = 0.7988$, $p = 0.0026$). Accordingly, we applied non-parametric tests to assess differences between datasets. A Friedman test across all datasets did not reveal a significant overall effect on model performance ($p = 0.0752$). Subsequent pairwise comparisons using Holm–Bonferroni corrected Wilcoxon signed-rank tests also showed no significant differences between any dataset pairs (all adjusted $p \geq 0.7500$). The results are summarized in Figure 4.

Per-Model Analysis

To further investigate whether the choice of annotation dataset influenced performance at the individual model level, we conducted additional statistical analyses for each model separately. As before, performance scores were obtained from five in-

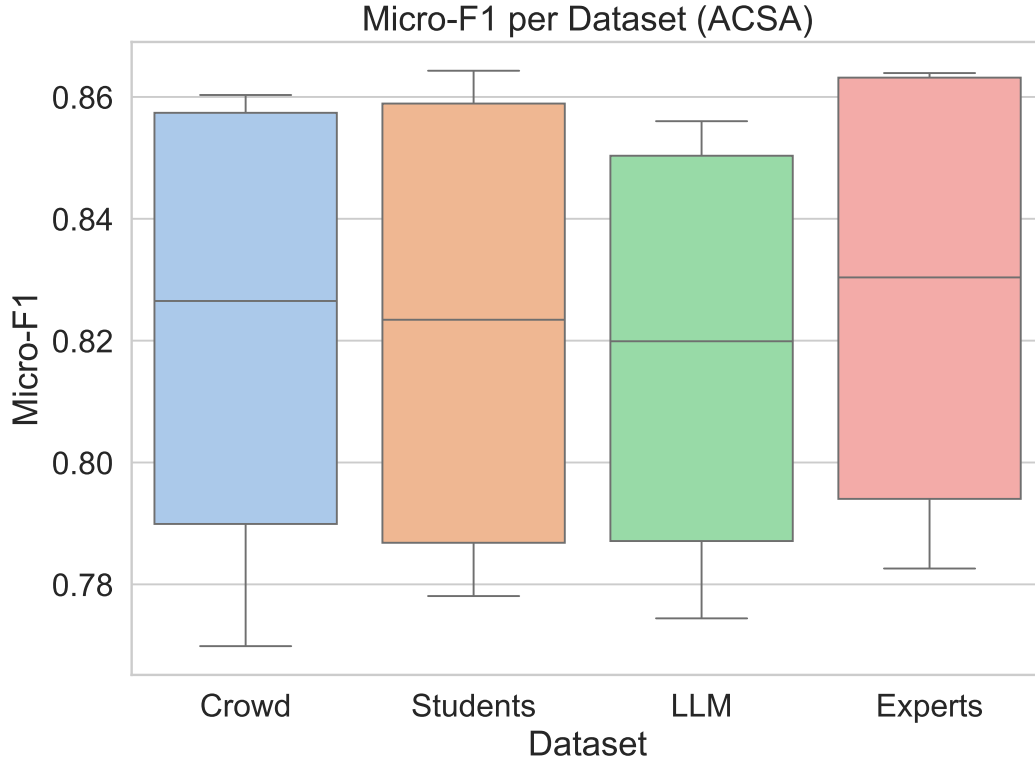


Figure 4: Micro-F1 scores for ACSA across annotation datasets. Each box shows the distribution of performance averaged over seeds for each model, combined across all models within the dataset.

dependent runs with different random seeds, which were treated as repeated measures in the analyses. Micro-F1 was used as the primary metric.

Shapiro–Wilk tests confirmed that all models satisfied the assumption of normality (*BERT-CLF*: $W = 0.9397$, $p = 0.2366$; *Hier-CGN*: $W = 0.9444$, $p = 0.2894$; *LLM-FS*: $W = 0.9752$, $p = 0.8581$; *LLM-FT*: $W = 0.9658$, $p = 0.6645$). Accordingly, repeated-measures one-way ANOVAs followed by Holm–Bonferroni corrected paired t-tests were applied to all four models.

For *BERT-CLF*, the ANOVA did not indicate a significant overall effect of dataset ($p = 0.3622$), and none of the post-hoc pairwise comparisons reached significance. For *Hier-CGN*, the ANOVA showed a borderline significant effect ($p = 0.0357$), with only the Experts vs Students comparison remaining significant after Holm–Bonferroni correction (adjusted $p = 0.0393$). *LLM-FS* exhibited a significant overall effect of dataset ($p = 0.0003$), with significant differences observed between Experts vs LLMs (adjusted $p = 0.0329$) and LLMs vs Students (adjusted $p = 0.0188$).

Finally, *LLM-FT* also showed a significant overall effect ($p = 0.0069$), but none of the pairwise comparisons remained significant after Holm–Bonferroni correction.

The results are summarized in Figure 5, and a detailed overview of all pairwise comparisons, including raw and adjusted p-values and significance per model, is provided in Appendix A.4.

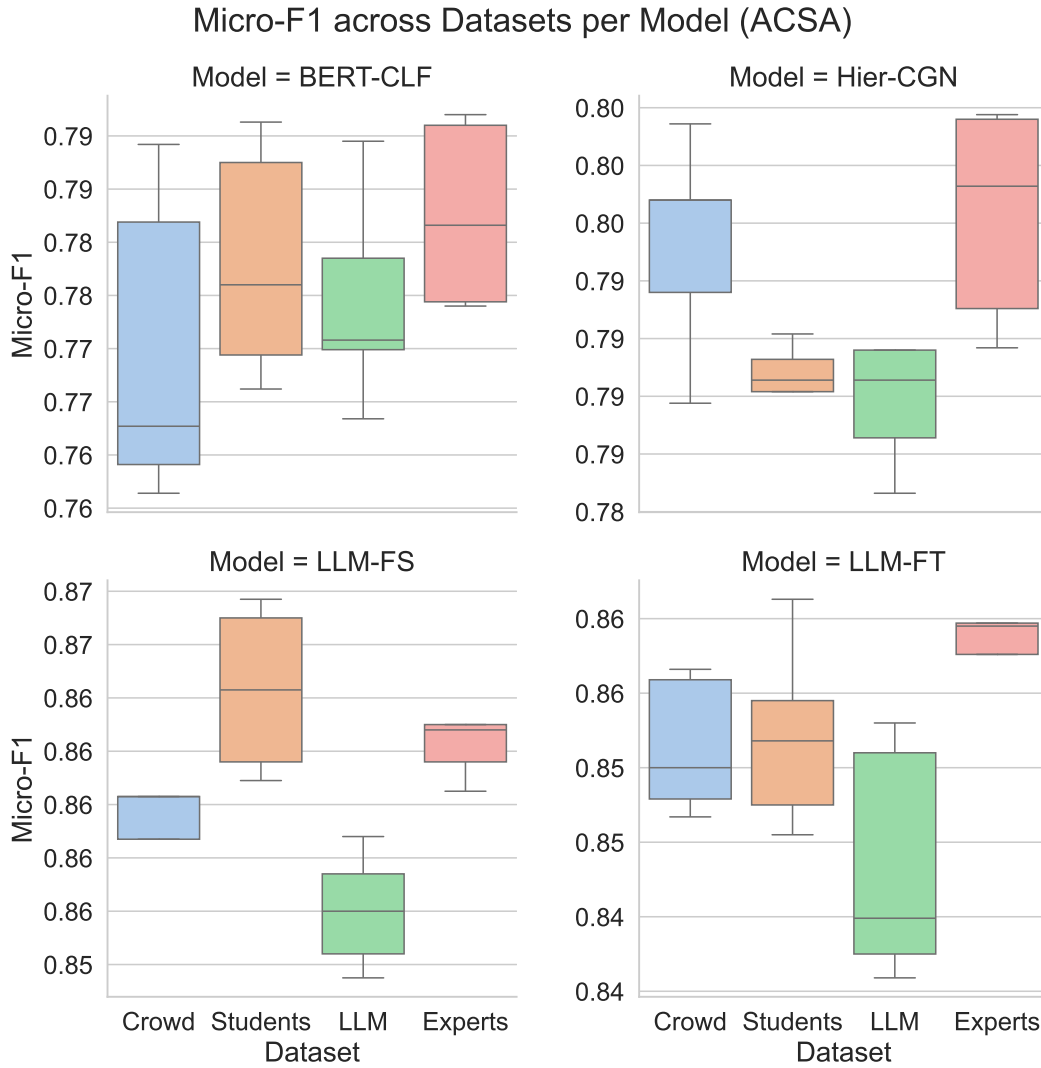


Figure 5: Micro-F1 scores for the ACSA task across annotation datasets. Each box represents the distribution of performance over five random seeds for a given model–dataset combination, shown separately per model. Note: the y-axis scale is adjusted individually for each model.

4.3.4. Statistical Evaluation of the TASD Task

For the TASD task, we applied the same statistical procedure as described for ACSA. Performance for each model was again obtained from five independent runs with different random seeds.

Overall Comparison Across Datasets

To determine the appropriate testing procedure, we first assessed the distribution of the model-averaged scores using the Shapiro–Wilk normality test, which indicated that the scores were normally distributed ($W = 0.9809$, $p = 0.9706$). Accordingly, we applied a repeated-measures one-way ANOVA to test for overall differences between datasets, which revealed no significant effect of dataset on performance ($p = 0.2348$). Post-hoc pairwise comparisons were conducted using Holm–Bonferroni corrected paired t-tests. Only the difference between Crowd and Students reached statistical significance (adjusted $p = 0.0106$), whereas all other pairwise comparisons were not significant (Crowd vs Experts: adjusted $p = 0.0960$; Crowd vs LLM: adjusted $p = 0.2367$; Experts vs LLM: adjusted $p = 0.4149$; Experts vs Students: adjusted $p = 0.2367$; LLM vs Students: adjusted $p = 0.5719$). The results are summarized in Figure 6.

Per-Model Analysis

To further investigate whether the choice of annotation dataset influenced individual model performance, we conducted additional statistical analyses for each model separately. For each model, performance scores were obtained from five independent runs with different random seeds, which were treated as repeated measures in the analyses. Micro-F1 was used as the primary metric.

Shapiro–Wilk normality tests indicated that *Paraphrase* ($W = 0.9601$, $p = 0.5468$) and *MvP* ($W = 0.9346$, $p = 0.1892$) scores were normally distributed, whereas *LLM-FS* ($W = 0.9042$, $p = 0.0495$) and *LLM-FT* ($W = 0.9034$, $p = 0.0477$) violated the normality assumption. Accordingly, repeated-measures one-way ANOVAs followed by Holm–Bonferroni corrected paired t-tests were applied for *Paraphrase* and *MvP*,

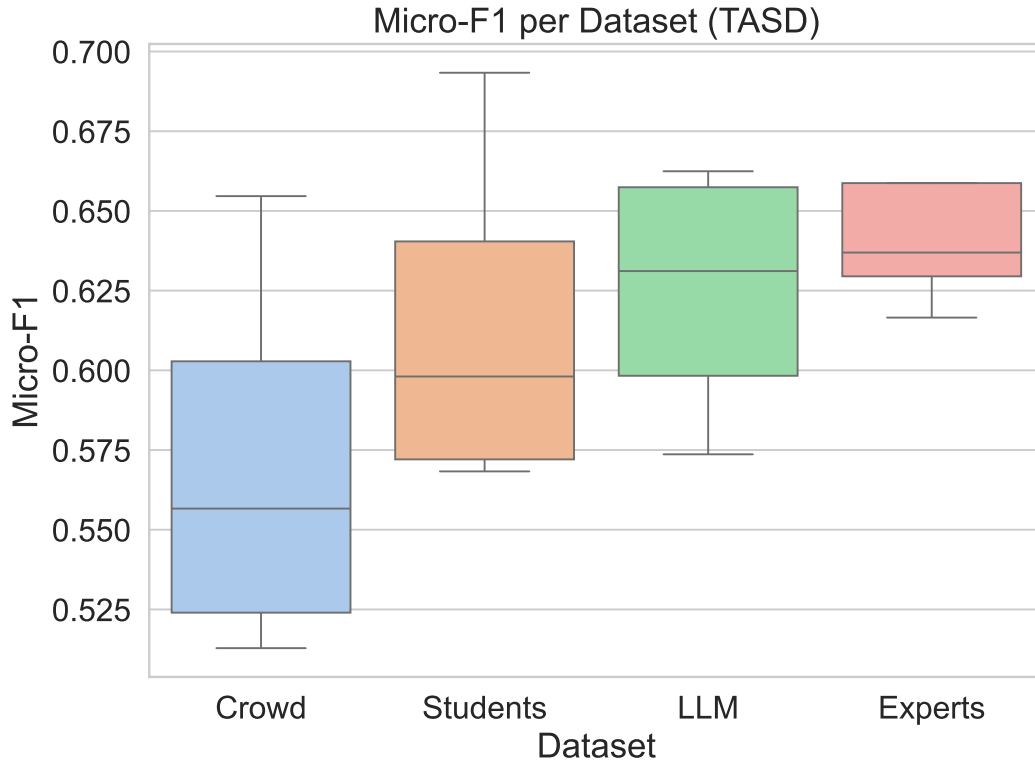


Figure 6: Micro-F1 scores for TASD across annotation datasets. Each box shows the distribution of performance averaged over seeds for each model, combined across all models within the dataset.

while non-parametric Friedman tests followed by Holm–Bonferroni corrected Wilcoxon signed-rank tests were used for *LLM-FS* and *LLM-FT*. For *Paraphrase* and *MvP*, the ANOVAs revealed a significant effect of dataset on performance ($p < 0.001$ for both), with most pairwise comparisons remaining significant after Holm–Bonferroni correction. The only exception was the LLM vs Student comparison for *Paraphrase*, which was not significant after correction (adjusted $p = 0.9602$). For *LLM-FS* and *LLM-FT*, the Friedman tests were significant ($p = 0.0029$ and $p = 0.0018$, respectively), but none of the pairwise comparisons were statistically significant after correction. The results are summarized in Figure 7. A detailed overview of all pairwise comparisons, including adjusted p-values and significance per model, is provided in Appendix A.4.

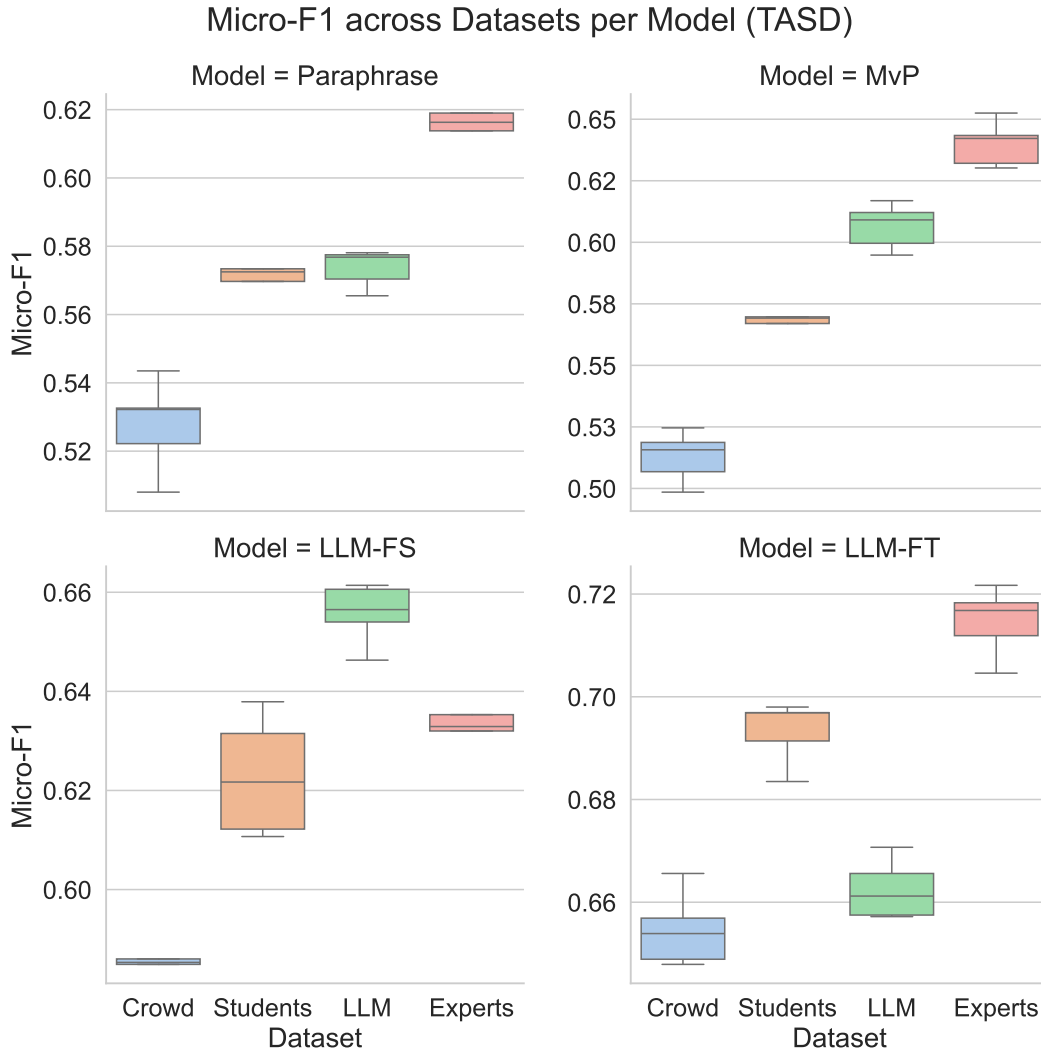


Figure 7: Micro-F1 scores for the TASD task across annotation datasets. Each box represents the distribution of performance over five random seeds for a given model–dataset combination, shown separately per model. Note: the y-axis scale is adjusted individually for each model.

4.4. Cost and Effort Analysis

This section provides an overview of the costs and effort involved in creating the datasets. Since not all annotation processes were systematically tracked, only partial information is available. The following results therefore focus on the measured and observed aspects of cost and time. For the crowd dataset, financial information is available for three Prolific studies. The ACSA study (15 participants) cost £202 in participant payment plus £81.02 in platform fees. The TASD pilot study (5 participants) amounted to £90 plus £35.99 in fees, and the final TASD study (10

participants) cost £180 plus £72 in fees. In total, across all crowd studies, 30 submissions were accepted with one rejection. Several returns ($n=22$) and one timeout occurred, but overall recruitment was fast and data collection was completed within two days for each study.

The creation of the student dataset required substantially more organizational effort. It took around three weeks to recruit participants, and although each had one week to complete the annotation, the process was slowed down by delayed responses and inactive participants. Participation was incentivized with *Versuchspersonenstunden* (2 per student), which proved necessary to ensure sufficient engagement. Without such incentives, recruitment would likely have been much more difficult. It should also be noted that the study took place during the semester break, which may have influenced participant recruitment.

The generation of the LLM dataset required only around ten hours in total. The main limitation here was computational: running Gemma 3 27B required relatively large GPU memory, even with optimizations through Ollama and quantization. Once sufficient resources were available, however, the process was straightforward and fully automated, with minimal manual effort.

For the expert dataset, annotation was based on an already existing dataset that one expert checked and refined. This setup made the process feasible: if no initial annotations had been available, multiple experts would have been required, making the task far more costly and time-consuming. Even under these conditions, the expert needed several hours to review and improve the annotations. In addition, suitable experts are generally scarce, making large-scale expert-based annotation particularly challenging.

4.5. Comparative Analysis of Dataset Variants

This section provides a comparison of the distributions across the five datasets, including the original *GERestaurant* corpus, the re-annotated versions (Crowd, Students, LLMs, Experts) and the new annotated ground truth. The goal is to assess the impact of the annotation processes on the resulting ACSA and TASD datasets.

Detailed tables showing the distribution of aspect categories and sentiment polarities for each dataset are provided in Appendix A.8.

4.5.1. ACSA

Figure 8 summarizes the distribution of aspect categories for the ACSA task across the re-annotated datasets. While the overall patterns remain comparable, noticeable shifts occur across categories. For example, the `FOOD` annotations vary between 432 (Students) and 458 (LLMs), while `SERVICE` ranges from 315 (LLMs) to 338 (Crowdworker). The `GENERAL` category shows stronger variation, with the highest number in the Students dataset (306) and the lowest in the Experts dataset (249). Similarly, the `AMBIENCE` category spans from 142 (Crowdworker) to 156 (Experts). Finally, `PRICE` remains the most stable category, ranging only from 66 (Students) to 72 (Crowdworker). Overall, the distributions are broadly consistent, but each annotation approach introduces characteristic shifts in category frequencies.

For the ACSA task, no implicit or explicit mentions are defined. However, polarity distributions provide additional insight. The distribution of sentiment polarity across the re-annotated datasets is largely consistent, with only minor variation. For example, *positive* cases range from 670 (LLMs) to 690 (Crowdworker), *negative* cases from 514 (Experts) to 541 (Students), and *neutral* cases from 56 (Experts) to 75 (LLMs). These small differences indicate that sentiment polarity is generally preserved across annotation sources.

4.5.2. TASD

Figure 9 visualizes the category distributions for the TASD task across the re-annotated datasets. Compared to the ACSA task, differences between datasets are more pronounced. The LLMs dataset shows the highest counts across most categories, for example 599 `FOOD` mentions. Experts annotations are slightly lower, while the Students and Crowdworker datasets generally contain fewer annotations. For instance, `SERVICE` ranges from 255 (Crowdworker) to 280 (Students), and `GENERAL` from 183 (Crowdworker) to 190 (Students). The reduction is most no-

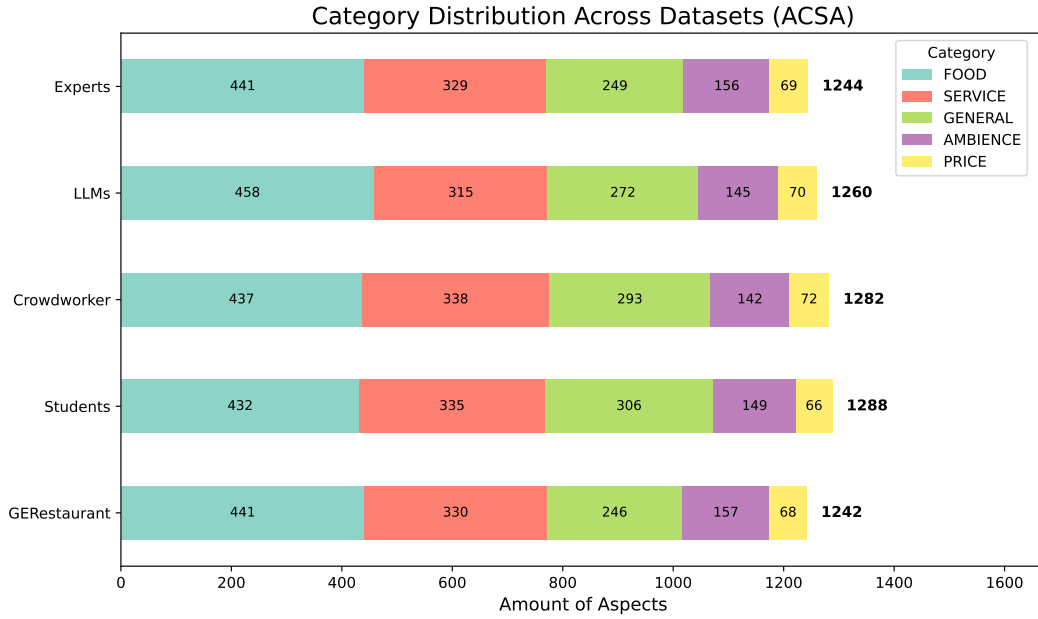


Figure 8: Category distribution across datasets for the ACSA datasets. The figure shows the number of annotated aspects per category across the different datasets. The total number of aspects per dataset is displayed at the end of each bar.

table in AMBIENCE and PRICE, with 118 and 49 cases for Crowdworke, compared to 156 and 69 for Experts. Overall, while the Experts and LLMs datasets maintain relatively high counts, the Students and Crowdworke annotations show considerably lower frequencies across categories.

Beyond category counts, the datasets also differ in their ratio of *explicit* and *implicit* mentions. The Experts and LLMs datasets maintain relatively high numbers, with 1,028/398 and 996/409 *explicit/implicit* annotations, respectively. In contrast, the Students and Crowdworke datasets show notably fewer annotations overall, with 851/241 and 777/263 *explicit/implicit* mentions.

The polarity distributions follow a similar pattern. The Experts and LLMs datasets maintain relatively high counts (Experts: 782/584/60; LLMs: 773/557/75 for *positive/negative/neutral*), while the Student and Crowdworke datasets record notably fewer instances (622/427/43 and 586/421/33, respectively).

Taken together, these results indicate that the Experts and LLMs datasets maintain relatively high counts and consistent distributions, whereas the Students and Crowdworke annotations systematically produce smaller datasets across categories,

polarity, and explicit/implicit mentions.



Figure 9: Category distribution across datasets for the TASD datasets. The figure shows the number of annotated aspects per category across the different datasets. The total number of aspects per dataset is displayed at the end of each bar.

4.5.3. Ground Truth

For the ground truth analysis, only two datasets are considered: the original *GERestaurant* corpus and the new expert re-annotation, both following the TASD schema (see Figure 10). Overall, the category distributions are very similar across the two versions. The original dataset contains 510 *FOOD* instances, compared to 479 in the re-annotated set. A similar small shift is visible for *SERVICE* (320 vs. 315) and *AMBIENCE* (130 vs. 124). In contrast, the *GENERAL* category increases from 236 to 263, while *PRICE* slightly decreases from 79 to 69. These differences indicate minor adjustments introduced during the re-annotation process, without substantially changing the overall balance of categories in the ground truth.

In addition the distinction between *explicit* and *implicit* mentions shows only a minor shift. The original dataset contains 914 *explicit* and 361 *implicit* annotations, while the re-annotated version has 920 *explicit* and 330 *implicit* cases. This indicates a slight reduction in implicit mentions, but overall the balance between explicit and

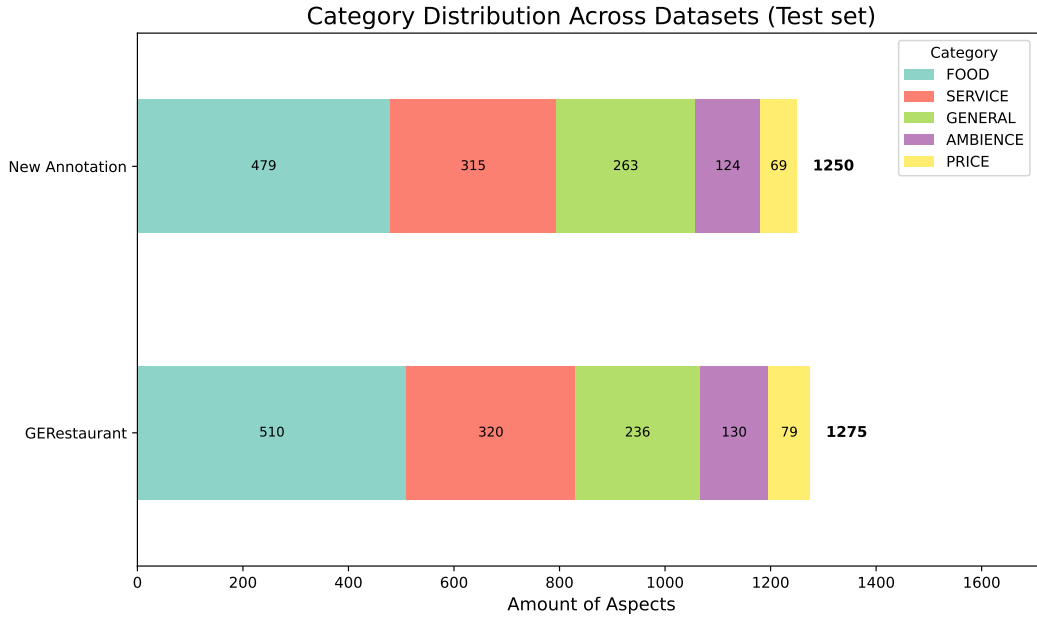


Figure 10: Category distribution across datasets for the ground truth datasets. The figure shows the number of annotated aspects per category across the different datasets. The total number of aspects per dataset is displayed at the end of each bar.

implicit annotations remains consistent.

Finally, the polarity balance was also compared across the two datasets. The overall distribution of *positive*, *negative*, and *neutral* instances remains nearly identical, with 672/539/64 in the original set and 654/532/64 in the re-annotation. This stability suggests that sentiment polarity was largely preserved during the expert revision.

4.6. Summary

The results reveal several consistent patterns across tasks and datasets. For IAA, the LLM-generated datasets achieve the highest reliability across tasks. For ACSA, the Student dataset shows slightly lower agreement than the Crowd dataset, due to two annotators misinterpreting the guidelines. For TASD, the Crowd dataset exhibits lower agreement than the Student dataset. Overall, differences in reliability between the non-LLM datasets are largely task-dependent.

Overall, scores are lower for the TASD task compared to ACSA, reflecting the

increased complexity of target aspect and phrase-level annotations.

Model performance follows a similar pattern: TASD results are generally lower than ACSA, with expert-annotated datasets providing the best performance in most cases. LLM fine-tuned models achieve the highest scores overall, with few exceptions where few-shot approaches occasionally outperform on specific datasets.

Cost and effort considerations differ across annotation methods. Crowd annotations are fast but costly and highly variable in quality, student annotations are slower but of moderate quality, LLM datasets are generated quickly but require substantial computational resources, and expert annotations deliver high-quality labels but are time-consuming and experts are rare.

Dataset composition also affects performance. In TASD, crowd and student datasets contain fewer aspect mentions, whereas all datasets are broadly similar for ACSA, a trend reflected in category and polarity distributions.

Statistical testing indicates that for ACSA, differences across datasets are generally not significant, although LLM-FS models show significant improvements between the Expert and LLM datasets and between the student and crowd datasets. For TASD, dataset differences are mostly non-significant, with some exceptions where MvP and Paraphrase models exhibit significant effects.

5. Discussion

This chapter reflects on the main findings of the study and situates them within the broader context of annotation practices and model evaluation for ABSA. The discussion is structured around two main perspectives: first, the creation of the different datasets and the resulting IAA for the ACSA and T ASD tasks as well as the ground truth; and second, the performance of models trained on these datasets. Following this, the limitations and ethical considerations of the study are outlined. The chapter concludes with a brief summary that highlights the key insights gained and their implications for future research.

5.1. Creating of the different Datasets

The creation of the different datasets revealed a range of practical challenges depending on the annotation approach. Crowdsourcing was relatively cost-intensive while still producing annotations of mixed quality, and student annotations, although more consistent, were often delayed as many participants tended to complete the task only towards the end of the allotted time. The LLM-based dataset required access to high-performance hardware and raised questions about potential biases inherited from the model’s training data. Expert annotations, while yielding the highest quality, were limited by the scarcity of qualified annotators and therefore restricted to the T ASD task, with the phrase component removed for the ACSA task to reduce workload.

Across all approaches, the importance of clear guidelines, well-structured instructions, and carefully designed annotation interfaces became evident. These factors were essential to minimize annotation errors, increase consistency, and ensure that annotators could reliably interpret the task requirements.

5.1.1. IAA on the ACSA Task

The IAA results for the ACSA task show relatively consistent patterns across datasets. Crowdfworker and student annotations achieve similar agreement levels, reflecting the simplicity of the annotation layout: annotators select from a predefined category–polarity tuple, which limits the potential for major mistakes. Nevertheless, some drops in IAA across batches indicate either that certain texts were more challenging to annotate or that individual annotators applied different interpretations. This is particularly noticeable in batches 3 and 4 of the student dataset, where the standard deviations for Pair-F1 are unusually high (Batch 3: $50.65_{\pm 25.84}$; Batch 4: $52.03_{\pm 27.54}$). These drops in both Pair-F1 and Krippendorff’s alpha indicate that two annotators misinterpreted the annotation task in these batches. This also inflates the overall macro-F1 standard deviation for the student dataset (± 17.27) compared to the crowdworker dataset (± 7.54).

For instance, in the Student dataset, two annotators misinterpreted the guidelines, leading to lower agreement. This example highlights that, even with clear instructions, errors can still occur during annotation. These observations align with Klie et al. (2024), who emphasize the importance of well-structured guidelines, comprehensive instructions, and thoughtfully designed annotation interfaces to reduce inconsistencies and improve reliability. To address these issues, an additional warning was introduced in the interface of the Crowdfworker study to prevent similar errors.

By contrast, the LLM annotations achieved very high agreement levels. Although a temperature of 0.8 was applied to encourage diversity in the outputs, the repeated prompts still produced highly similar responses, which likely explains the strong IAA observed for this dataset. These findings align also with Klie et al. (2024), who note that high agreement does not automatically guarantee high-quality labels.

5.1.2. IAA on the TASD Task

For the TASD task, the IAA patterns diverge more strongly across datasets, reflecting the higher complexity of the annotation task. Unlike ACSA, annotators were

required to freely select text spans within the given sentence, which increased variability in the results. The gap between student and crowdworker agreement was larger in this task: while student annotations were generally consistent with the guidelines, crowdworkers often produced long phrases or labeled sentence fragments containing the aspect instead of marking only the aspect itself. Other issues included inconsistent use of the implicit label and cases where annotators split longer aspects into multiple single-word annotations. This led to a wide spectrum of outcomes, ranging from careful, guideline-conform annotations to incomplete or clearly erroneous ones. Such outliers appeared far less frequently in the student dataset. These differences are also reflected in the standard deviations of the macro Pair-F1 scores, with students showing ± 9.07 and crowdworkers ± 16.80 . In particular, batches 4 and 5 of the crowdworker dataset exhibit high variance, as indicated by the combination of low macro-F1 scores and elevated standard deviations.

As in ACSA, the LLMs showed very high IAA due to the identical input prompt and fixed setup, despite the use of a temperature parameter intended to increase diversity. Overall, the IAA for the TASD task was lower than for ACSA, which can be attributed to the added difficulty of aspect phrase prediction. This observation aligns with Monarch (2021), who report that tasks challenging for human annotators tend to also pose difficulties for machine learning models (see Section 4.3.2 for the corresponding model results).

5.1.3. IAA on the Ground Truth

Inter-annotator agreement on the ground truth TASD dataset was consistently high, with an overall macro Pair-F1 of $72.14_{\pm 5.54}$, substantially higher than agreement levels observed in the student or crowdworker annotations.

Inspection of individual batches reveals a clear upward trend in agreement, with Pair-F1 rising from 63.33 to 76.95 and exhibiting low variance, suggesting that iterative discussions among annotators and guideline refinements foster more consistent labeling. This underscores the value of regular calibration sessions and collaborative review in achieving high-quality gold standard annotations.

These results, showing that iterative discussions and guideline refinements improve annotation consistency, align with previous findings that careful guideline design, regular feedback, and expert collaboration enhance annotation reliability for complex tasks such as TASD (Klie et al., 2024; Fehle et al., 2025).

5.2. Performance across Datasets

This section examines how model performance varies depending on the underlying training dataset. By comparing results across the different annotation sources, we can evaluate the influence of annotation quality and dataset characteristics on predictive outcomes. The analysis is presented separately for the ACSA and TASD tasks, highlighting both overall trends and category–polarity-specific patterns.

5.2.1. Performance on the ACSA Task

Across all models, performance differences between datasets remain relatively small, suggesting that each dataset is broadly suitable for the ACSA task. However, expert annotations consistently deliver the highest scores, confirming their value when the goal is to achieve the best possible performance.

The comparison of modeling approaches shows that LLM-based methods clearly outperform traditional baselines. Differences between few-shot prompting and fine-tuning are minor, although the few-shot setting slightly outperforms fine-tuning on most datasets, except for the expert-annotated corpus. This outcome may be partially explained by the underlying model sizes, since the few-shot experiments relied on a larger model (*Gemma 3 27B*) compared to the fine-tuned variant (*LLaMA 8B*).

Looking at category–polarity combinations, clear patterns emerge (compare Appendix A.9). Positive polarity is generally the easiest to predict, followed by negative polarity, whereas neutral polarity achieves the lowest scores. The `PRICE` category shows particularly low performance for the positive polarity, reflecting the limited number of such mentions in the training sets. For negative polarity, performance is relatively stable across categories with only a small drop in `AMBIENCE`. The most

notable decline appears for neutral polarity in `AMBIENCE`, where some models fail to predict any cases at all. Neutral polarity is also weaker in most categories except `FOOD`, which benefits from a comparatively higher number of neutral examples.

A detailed breakdown of these results is provided in Figure 14 in the Appendix.

5.2.2. Performance on the TASD Task

For the TASD task, differences in model performance across datasets are more pronounced compared to the ACSA task. The crowdworker dataset consistently achieves the lowest scores for all models, likely reflecting both the task complexity and lower inter-annotator agreement. Student annotations generally perform better than crowdworkers but still fall short of LLM-generated datasets, except in the case of the LLM fine-tuning approach, highlighting the capability of LLMs to produce high-quality annotations in complex tasks.

Expert annotations provide the best results across nearly all models, except for the LLM few-shot approach, where the LLM-generated dataset achieves a slightly higher score. This is likely because both the training examples and the prediction method come from the same model (*Gemma 3 27B*), allowing the few-shot method to generalize more effectively from familiar patterns. In line with W. Zhang et al. (2024), *LLM-FS* achieves higher performance than *MvP* and *Paraphrase* in the few-shot TASD setting, underlining the superior generalization capabilities of LLMs compared to smaller models.

Category-polarity analysis reveals patterns similar to those observed for ACSA (compare Appendix A.9). Positive polarity is generally the easiest to predict, followed by negative polarity, while neutral polarity achieves the lowest scores. Drops in performance are notable for positive mentions in the `PRICE` category, due to its low frequency in the training data, and for negative sentiment in `AMBIENCE`. Neutral polarity is particularly challenging, with some categories, especially `AMBIENCE`, showing no correct predictions. These patterns highlight that models struggle with infrequent classes and with the neutral polarity, which is inherently difficult to distinguish from positive or negative sentiment even for expert annotators.

For triplet prediction, similar trends are observed, with `PRICE` being the most challenging category due to its low representation in the training sets. A detailed breakdown of category–polarity F1 scores is presented in Figure 15 and Figure 13 in the Appendix.

5.3. Limitations & Ethical Considerations

This study has several limitations that should be taken into account. First, while crowd annotations enabled rapid data collection, they come with a financial cost that scales with dataset size. Student-based annotations were particularly time-consuming: recruitment and completion often took the entire allocated period, with many participants submitting their work only at the end of the one-week window. The LLM-generated dataset, on the other hand, may inherit biases present in the model’s training data and required substantial computational resources, including a high-performance GPU with considerable memory, which could limit reproducibility. Expert annotations produced the highest-quality data, yet they are time-intensive and rely on rare expertise, making large-scale collection challenging. Finally, all datasets and analyses are restricted to the restaurant review domain, which limits the generalization of our findings to other domains or languages. Taken together, these factors highlight practical and methodological constraints that should be considered when interpreting model performance, dataset comparisons, and the broader applicability of the results.

The dataset and its annotations are available upon request from the authors to ensure responsible academic use, and the associated Python code for data collection and cleaning is publicly accessible via GitHub.¹⁰

For the crowdworker and student annotations conducted in this study, no additional demographic information was collected from participants, minimizing potential privacy risks. The annotation procedures followed ethical guidelines for data protection and were reviewed to avoid potential harms.

We recognize that all datasets inherently involve some degree of personal judg-

¹⁰GitHub:<https://github.com/ValdrDarmir/Annotation-Quality-and-Its-Influence-on-ABSA-A-Case-Study-on-German-Restaurant-Reviews>

ment and potential bias. For the crowdworker, student, and LLM datasets, this was mitigated through majority voting, which reduces the influence of individual annotator preferences. Bias is particularly critical in the ground truth, as it serves as the reference for model evaluation; to minimize subjectivity here, two expert annotators independently annotated the data and then reconciled any differences to reach a consensus. Despite these measures, some residual influence of individual judgment cannot be entirely eliminated.

5.4. Summary

The discussion highlights several key insights regarding dataset creation, annotation quality, and model performance for the aspect-based sentiment analysis. First, clear guidelines and carefully designed annotation interfaces are essential for minimizing errors and ensuring consistent annotations across annotators. IAA provides a useful indication of task complexity and dataset reliability; however, high IAA does not automatically guarantee superior model performance, as observed for LLM annotations, which achieved high agreement levels, but model performance was similar to the student datasets.

Each annotation approach presents its own challenges: crowdsourcing is cost-intensive and can produce variable-quality annotations, student-based annotations are slower and sometimes inconsistent, LLM-generated datasets require computational resources and may reflect biases from training data, and expert annotations are time-consuming but deliver the highest quality. LLM-based models achieved very competitive performance across tasks, likely benefiting from the extensive examples seen during pretraining. Expert annotations remain the most reliable means of achieving the best predictive performance.

For ABSA research, these findings suggest that LLMs offer a fast, scalable, and generally high-performing annotation solution, with only minor limitations, while expert annotation remains the gold standard for maximizing model accuracy and reliability.

6. Conclusion & Future Work

The influence of annotation quality and annotator type on both dataset and model performance for complex tasks such as ABSA has not been systematically analyzed. This work addressed this gap by conducting five annotation studies, including the creation of an independent ground truth with two expert annotators, and four training datasets annotated by crowdworkers, students, LLMs, and experts.

Model performance was evaluated using SOTA methods for two ABSA tasks: ACSA and TASD. In addition, IAA was analyzed across the different datasets, and statistical tests were conducted to assess significant differences between methods and annotation sources.

The results indicate that expert-annotated datasets consistently provide the highest performance across nearly all models. For ACSA, the expert dataset yielded the best results for almost every method, although the differences between the datasets were relatively small. LLM-based models generally outperformed classical approaches such as BERT-CLF or Hier-GCN. For TASD, the expert dataset again achieved almost the best results across methods, with LLM fine-tuned models yielding the highest performance, particularly on the expert dataset. IAA analysis showed that for ACSA, crowdworker annotations had the lowest F1 scores, followed by student annotations, while LLM annotations achieved the highest agreement. For TASD, agreement scores were lower overall and varied considerably across annotation batches. These findings highlight several important insights. Structured annotation guidelines and well-designed interfaces are crucial to prevent erroneous or inconsistent annotations. LLM-generated annotations are close to expert-level quality for the ACSA task and achieve similar performance to student or crowdworker datasets for TASD. Expert annotations, while time-intensive, consistently improve dataset quality, with every method achieving its best score on the expert dataset

(except for LLM-FS, which performed best on the LLM dataset). Across both tasks, LLM-based models generally outperform classical methods in terms of Micro- and Macro-F1 scores.

Several avenues for future research emerge from this study. First, evaluating model performance and dataset quality on a different domain would allow an assessment of domain transfer and generalization beyond restaurant reviews. Second, combining annotation approaches could be explored, for instance by having LLM-generated annotations reviewed and refined by experts, potentially balancing efficiency and quality. Finally, increasing the number of annotators in the crowd-worker and student datasets, as well as experimenting with alternative label aggregation strategies, could further improve dataset reliability. Additionally, relaxing strict phrase boundaries may increase coverage and provide more robust training data for aspect-based sentiment analysis models.

Bibliography

- Ara, J., Hasan, M. T., Al Omar, A., & Bhuiyan, H. (2020, June). Understanding Customer Sentiment: Lexical Analysis of Restaurant Reviews. In *2020 IEEE Region 10 Symposium (TENSYP)* (pp. 295–299). Retrieved 2025-09-25, from <https://ieeexplore.ieee.org/abstract/document/9230712> (ISSN: 2642-6102) doi: 10.1109/TENSYP50017.2020.9230712
- Bai, Y., Han, Z., Zhao, Y., Gao, H., Zhang, Z., Wang, X., & Hu, M. (2024, November). Is Compound Aspect-Based Sentiment Analysis Addressed by LLMs? In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 7836–7861). Miami, Florida, USA: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2024.findings-emnlp.460/> doi: 10.18653/v1/2024.findings-emnlp.460
- Basile, P., Croce, D., Basile, V., & Polignano, M. (2018). Overview of the EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA). In *EVALITA Evaluation of NLP and Speech Tools for Italian* (pp. 1–10). CEUR. Retrieved 2024-11-12, from <https://iris.unito.it/bitstream/2318/1759769/1/paper003.pdf>
- Bhoi, A., & Joshi, S. (2018, May). *Various Approaches to Aspect-based Sentiment Analysis*. arXiv. Retrieved 2025-09-26, from <http://arxiv.org/abs/1805.01984> (arXiv:1805.01984 [cs]) doi: 10.48550/arXiv.1805.01984
- Brauwers, G., & Frasincar, F. (2022, November). A Survey on Aspect-Based Sentiment Classification. *ACM Comput. Surv.*, 55(4), 65:1–65:37. Retrieved 2024-11-27, from <https://dl.acm.org/doi/10.1145/3503044> doi: 10.1145/3503044
- Brun, C., & Nikoulina, V. (2018, October). Aspect Based Sentiment Analysis into the Wild. In A. Balahur, S. M. Mohammad, V. Hoste, & R. Klinger (Eds.), *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 116–122). Brussels, Belgium: Association for Computational Linguistics. Retrieved 2025-09-25, from <https://aclanthology.org/W18-6217/> doi: 10.18653/v1/W18-6217
- Bu, J., Ren, L., Zheng, S., Yang, Y., Wang, J., Zhang, F., & Wu, W. (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* (pp. 2069–2079).
- Cai, H., Tu, Y., Zhou, X., Yu, J., & Xia, R. (2020, December). Aspect-Category based Sentiment Analysis with Hierarchical Graph Convolutional Network. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 833–843). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved 2025-09-16, from <https://aclanthology.org/2020.coling-main.72/> doi: 10.18653/v1/2020.coling-main.72
- Cai, H., Xia, R., & Yu, J. (2021, August). Aspect-Category-Opinion-Sentiment Quadruple Extraction with Implicit Aspects and Opinions. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 340–350). Online: Association for Computational Linguistics. Retrieved 2025-06-04, from <https://aclanthology.org/2021.acl-long.29/> doi: 10.18653/v1/2021.acl-long.29
- Chauhan, G. S., Nahta, R., Meena, Y. K., & Gopalani, D. (2023, August). Aspect based sentiment analysis using deep learning approaches: A survey. *Computer Science Review*, 49, 100576. Retrieved 2024-11-27, from <https://www.sciencedirect.com/science/article/pii/S1574013723000436> doi: 10.1016/j.cosrev.2023.100576
- Chebolu, S. U. S., Derroncourt, F., Lipka, N., & Solorio, T. (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)* (pp. 611–628). Retrieved 2025-03-25, from <https://aclanthology.org/2023.ijcnlp-main.41.pdf>
- Cheng, J., Zhao, S., Zhang, J., King, I., Zhang, X., & Wang, H. (2017, November). Aspect-level Sentiment Classification with HEAT (HiErarchical ATtention) Network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 97–106). New York, NY, USA: Association for Computing Machinery. Retrieved 2025-09-24, from <https://dl.acm.org/doi/10.1145/3132847.3133037> doi: 10.1145/3132847.3133037
- Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., ... Wiegand, M. (2012). Mlsa — a multi-layered reference corpus for german sentiment analysis. In *Proceedings of the eight international conference on*

- language resources and evaluation (lrec'12)*. (Retrieved from <http://www.lrec-conf.org/proceedings/lrec2012/index.html>)
- Cohen, J. (1960, April). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. Retrieved 2025-09-22, from <https://journals.sagepub.com/doi/10.1177/001316446002000104> doi: 10.1177/001316446002000104
- Colucci Cante, L., D'Angelo, S., Di Martino, B., & Graziano, M. (2024). Text Annotation Tools: A Comprehensive Review and Comparative Analysis. In L. Barolli (Ed.), *Complex, Intelligent and Software Intensive Systems* (pp. 353–362). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-70011-8_33
- de França Costa, D., & da Silva, N. F. F. (2018, April). INF-UFG at FiQA 2018 Task 1: Predicting Sentiments and Aspects on Financial Tweets and News Headlines. In *Companion Proceedings of the The Web Conference 2018* (pp. 1967–1971). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. Retrieved 2024-11-12, from <https://dl.acm.org/doi/10.1145/3184558.3191828> doi: 10.1145/3184558.3191828
- Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., ... Solti, I. (2012). Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings* (Vol. 2012, p. 144). Retrieved 2025-09-29, from <https://pmc.ncbi.nlm.nih.gov/articles/PMC3540456/>
- De Mattei, L., De Martino, G., Iovine, A., Miaschi, A., Polignano, M., & Rambelli, G. (2020). ATE ABSITA @ EVALITA2020: Overview of the aspect term extraction and aspect-based sentiment analysis task. CEUR-WS. Retrieved 2024-11-12, from <https://cris.unibo.it/handle/11585/938097> (Accepted: 2023-08-31T12:49:36Z)
- Ding, X., Liu, B., & Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 231–240). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-11-12, from <https://dl.acm.org/doi/10.1145/1341531.1341561> doi: 10.1145/1341531.1341561
- Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., & Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 49–54). Retrieved 2024-11-12, from <https://aclanthology.org/P14-2009.pdf>
- Fan, Z., Wu, Z., Dai, X.-Y., Huang, S., & Chen, J. (2019, June). Target-oriented Opinion Words Extraction with Target-fused Neural Sequence Labeling. In J. Burstein,

- C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2509–2518). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved 2024-11-12, from <https://aclanthology.org/N19-1259> doi: 10.18653/v1/N19-1259
- Fehle, J., Donhauser, N., Kruschwitz, U., Hellwig, N. C., & Wolff, C. (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)* (Vol. 9, p. 213). Retrieved 2025-09-14, from <https://serwiss.bib.hs-hannover.de/files/3678/978-3-69018-015-3.pdf#page=227>
- Fehle, J., Münster, L., Schmidt, T., & Wolff, C. (2023). Aspect-Based Sentiment Analysis as a Multi-Label Classification Task on the Domain of German Hotel Reviews. In *KONVENS*. Retrieved 2025-05-05, from <https://openreview.net/forum?id=XgadCAfUhd>
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r*. Sage Publications. Retrieved from <https://www.torrossa.com/gs/resourceProxy?an=4913501&publisher=FZ7200>
- Fisher, R. A. (1992). Statistical Methods for Research Workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 66–70). New York, NY: Springer. Retrieved 2025-09-22, from https://doi.org/10.1007/978-1-4612-4380-9_6 doi: 10.1007/978-1-4612-4380-9_6
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378. Retrieved 2025-09-22, from <https://psycnet.apa.org/record/1972-05083-001> (Publisher: American Psychological Association)
- Friedman, M. (1937, December). 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. Retrieved 2025-09-04, from <http://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522> doi: 10.1080/01621459.1937.10503522
- Gabryszak, A., & Thomas, P. (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* (pp. 35–39). Retrieved 2025-03-25, from <https://aclanthology.org/2022.csrnlp-1.5/>
- Gou, Z., Guo, Q., & Yang, Y. (2023, July). MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction. In A. Rogers, J. Boyd-Graber, & N. Okazaki

- (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4380–4397). Toronto, Canada: Association for Computational Linguistics. Retrieved 2025-09-16, from <https://aclanthology.org/2023.acl-long.240/> doi: 10.18653/v1/2023.acl-long.240
- Grandini, M., Bagli, E., & Visani, G. (2020, August). *Metrics for Multi-Class Classification: an Overview*. arXiv. Retrieved 2025-09-23, from <http://arxiv.org/abs/2008.05756> (arXiv:2008.05756 [stat]) doi: 10.48550/arXiv.2008.05756
- Greve, L. D., Singh, P., Hee, C. V., Lefever, E., & Martens, G. (2021, December). 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. Retrieved 2025-04-23, from <https://clinjournal.org/clinj/article/view/142>
- Hamborg, F., Donnay, K., & Merlo, P. (2021). NewsMTSC: a dataset for (multi-) target-dependent sentiment classification in political news articles. Association for Computational Linguistics (ACL). Retrieved 2024-11-12, from <https://www.zora.uzh.ch/id/eprint/207183/>
- Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (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)* (pp. 123–133).
- Hellwig, N. C., Fehle, J., Kruschwitz, U., & Wolff, C. (2025, May). *Do we still need Human Annotators? Prompting Large Language Models for Aspect Sentiment Quad Prediction*. arXiv. Retrieved 2025-06-13, from <http://arxiv.org/abs/2502.13044> (arXiv:2502.13044 [cs]) doi: 10.48550/arXiv.2502.13044
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70. Retrieved 2025-09-04, from <https://www.jstor.org/stable/4615733> (Publisher: JSTOR)
- Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3), 296–298. Retrieved 2025-09-29, from <https://academic.oup.com/jamia/article-abstract/12/3/296/812057> (Publisher: BMJ Group BMA House, Tavistock Square, London, WC1H 9JR)
- Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168–177). New York, NY, USA: Association for Com-

- puting Machinery. Retrieved 2024-11-12, from <https://dl.acm.org/doi/10.1145/1014052.1014073> doi: 10.1145/1014052.1014073
- Hua, Y. C., Denny, P., Wicker, J., & Taskova, K. (2024, September). A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57(11), 296. Retrieved 2025-02-19, from <https://doi.org/10.1007/s10462-024-10906-z> doi: 10.1007/s10462-024-10906-z
- Jiang, Q., Chen, L., Xu, R., Ao, X., & Yang, M. (2019, November). A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 6280–6285). Hong Kong, China: Association for Computational Linguistics. Retrieved 2024-11-12, from <https://aclanthology.org/D19-1654> doi: 10.18653/v1/D19-1654
- Jun, Y., & Lee, H. (2025, July). Dynamic Order Template Prediction for Generative Aspect-Based Sentiment Analysis. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 614–626). Vienna, Austria: Association for Computational Linguistics. Retrieved 2025-09-10, from <https://aclanthology.org/2025.acl-short.48/> doi: 10.18653/v1/2025.acl-short.48
- Kessler, J. S., Eckert, M., Clark, L., & Nicolov, N. (2010). The ICWSM 2010 JDPA sentiment corpus for the automotive domain. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC)*. Retrieved 2024-11-12, from https://www.icwsml.org/2010/papers/icwsml0dcw_8.pdf
- Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018, August). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In D. Zhao (Ed.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 5–9). Santa Fe, New Mexico: Association for Computational Linguistics. Retrieved 2025-06-03, from <https://aclanthology.org/C18-2002/>
- Klie, J.-C., Castilho, R. E. d., & Gurevych, I. (2024). Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3), 817–866. Retrieved 2025-05-08, from https://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00516/2351693/coli_a_00516.pdf (Publisher: MIT Press 255 Main Street, 9th Floor, Cambridge, Massachusetts 02142, USA ...)

- Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. University of Pennsylvania, Department of Communication.
- Lee, L.-H., Yu, L.-C., Wang, S., & Liao, J. (2024, August). Overview of the SIGHAN 2024 shared task for Chinese dimensional aspect-based sentiment analysis. In K.-F. Wong et al. (Eds.), *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)* (pp. 165–174). Bangkok, Thailand: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2024.sighan-1.19/>
- Li, X., Bing, L., Zhang, W., & Lam, W. (2019). Exploiting bert for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th workshop on noisy user-generated text (w-nut 2019)*.
- Li, Y., Wang, F., & Zhong, S.-h. (2023). A more fine-grained aspect-sentiment-opinion triplet extraction task. *Mathematics*, 11(14), 3165.
- Liu, B. (2022). *Sentiment analysis and opinion mining*. San Rafael, California: Morgan & Claypool. Retrieved from <https://www.doi.org/10.1007/978-3-031-02145-9>
- Liu, Q., Gao, Z., Liu, B., & Zhang, Y. (2015). Automated rule selection for aspect extraction in opinion mining. In *Twenty-Fourth international joint conference on artificial intelligence*. Retrieved 2024-11-12, from <https://www.ijcai.org/Proceedings/15/Papers/186.pdf>
- Lv, H., Liu, J., Wang, H., Wang, Y., Luo, J., & Liu, Y. (2023, May). Efficient Hybrid Generation Framework for Aspect-Based Sentiment Analysis. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1007–1018). Dubrovnik, Croatia: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2023.eacl-main.71/> doi: 10.18653/v1/2023.eacl-main.71
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024, February). *Large Language Models: A Survey*. arXiv. Retrieved 2024-04-18, from <http://arxiv.org/abs/2402.06196> (arXiv:2402.06196 [cs]) doi: 10.48550/arXiv.2402.06196
- Monarch, R. M. (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.
- Mubarok, M. S., Adiwijaya, & Aldhi, M. D. (2017, August). Aspect-based sentiment analysis to review products using Naïve Bayes. *AIP Conference Proceedings*,

- 1867(1), 020060. Retrieved 2025-09-26, from <https://doi.org/10.1063/1.4994463> doi: 10.1063/1.4994463
- Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., & Daudpota, S. M. (2024). Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access*, 12, 60943-60959. doi: 10.1109/ACCESS.2024.3386969
- Orr, W., & Crawford, K. (2024, September). The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media & Society*, 26(9), 4955-4972. Retrieved 2025-09-25, from <https://doi.org/10.1177/14614448241251797> (Publisher: SAGE Publications) doi: 10.1177/14614448241251797
- Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., & Si, L. (2020, April). Knowing What, How and Why: A Near Complete Solution for Aspect-Based Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 8600-8607. Retrieved 2024-11-12, from <https://ojs.aaai.org/index.php/AAAI/article/view/6383> (Number: 05) doi: 10.1609/aaai.v34i05.6383
- Pham, D.-H., & Le, A.-C. (2018, December). Exploiting multiple word embeddings and one-hot character vectors for aspect-based sentiment analysis. *International Journal of Approximate Reasoning*, 103, 1-10. Retrieved 2025-09-24, from <https://www.sciencedirect.com/science/article/pii/S0888613X17304139> doi: 10.1016/j.ijar.2018.08.003
- Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., ... De Clercq, O. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation* (pp. 19-30). Retrieved 2024-11-08, from <https://hal.science/hal-01838537/>
- Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015, June). SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In P. Nakov, T. Zesch, D. Cer, & D. Jurgens (Eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 486-495). Denver, Colorado: Association for Computational Linguistics. Retrieved 2024-11-13, from <https://aclanthology.org/S15-2082> doi: 10.18653/v1/S15-2082
- Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014, August). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In P. Nakov & T. Zesch (Eds.), *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 27-35). Dublin, Ireland: Association for Computational Linguistics. Retrieved 2024-11-13, from <https://aclanthology.org/S14-2004> doi: 10.3115/v1/S14-2004

- Rahman, M. A., & Kumar Dey, E. (2018, June). Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation. *Data*, 3(2), 15. Retrieved 2024-11-12, from <https://www.mdpi.com/2306-5729/3/2/15> (Number: 2 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/data3020015
- Regatte, Y. R., Gangula, R. R. R., & Mamidi, R. (2020, May). Dataset Creation and Evaluation of Aspect Based Sentiment Analysis in Telugu, a Low Resource Language. In N. Calzolari et al. (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 5017–5024). Marseille, France: European Language Resources Association. Retrieved 2024-11-12, from <https://aclanthology.org/2020.lrec-1.617>
- Sadia, A., Khan, F., & Bashir, F. (2018). An Overview of Lexicon-Based Approach For Sentiment Analysis. In *2018 3rd international electrical engineering conference (ieec 2018)* (pp. 1–6).
- Saeidi, M., Bouchard, G., Liakata, M., & Riedel, S. (2016). Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 1546–1556).
- Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, 321–325. Retrieved 2025-09-26, from <https://www.jstor.org/stable/2746450> (Publisher: JSTOR)
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591–611. Retrieved 2025-09-04, from <https://academic.oup.com/biomet/article-pdf/52/34/591/962907/52-3-4-591.pdf> (Publisher: Oxford University Press)
- Sidarenka, U. (2016, May). PotTS: The Potsdam Twitter Sentiment Corpus. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1133–1141). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved 2025-03-26, from <https://aclanthology.org/L16-1181/>
- Simmering, P. F., & Huoviala, P. (2023, October). *Large language models for aspect-based sentiment analysis*. arXiv. Retrieved 2025-09-16, from <http://arxiv.org/abs/2310.18025> (arXiv:2310.18025 [cs]) doi: 10.48550/arXiv.2310.18025
- Singhi, V., Chauhan, C., & Soni, P. K. (2024, April). Exploring Progress in Aspect-based Sentiment Analysis: An In-depth Survey. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1–10). Retrieved 2024-11-05, from <https://ieeexplore.ieee.org/abstract/document/10543612> doi: 10.1109/I2CT61223.2024.10543612

- Steinberger, J., Brychcín, T., & Konkol, M. (2014). Aspect-level sentiment analysis in czech. In *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 24–30). Retrieved 2024-11-12, from <https://aclanthology.org/W14-2605.pdf>
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012, April). brat: a Web-based Tool for NLP-Assisted Text Annotation. In F. Second (Ed.), *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102–107). Avignon, France: Association for Computational Linguistics. Retrieved 2025-06-05, from <https://aclanthology.org/E12-2021/>
- Student. (1908). The probable error of a mean. *Biometrika*, 1–25. Retrieved 2025-09-04, from <https://www.jstor.org/stable/2331554> (Publisher: JSTOR)
- Sänger, M., Kemmerer, S., Adolphs, P., Klinger, R., & Leser, U. (2016). SCARE : the Sentiment Corpus of App Reviews with Fine-grained Annotations in German. Otto-Friedrich-Universität. Retrieved 2025-05-08, from <https://fis.uni-bamberg.de/handle/uniba/96474>
- Tong, Z., & Wei, W. (2024, August). CCIPLab at SIGHAN-2024 dimABSA Task: Contrastive Learning-Enhanced Span-based Framework for Chinese Dimensional Aspect-Based Sentiment Analysis. In K.-F. Wong et al. (Eds.), *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)* (pp. 102–111). Bangkok, Thailand: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2024.sighan-1.12/>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved 2025-09-24, from <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Wan, H., Yang, Y., Du, J., Liu, Y., Qi, K., & Pan, J. Z. (2020). Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 9122–9129). Retrieved 2025-09-25, from <https://aaai.org/ojs/index.php/AAAI/article/view/6447> (Issue: 05)
- Wang, H., Lu, Y., & Zhai, C. (2011, August). Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 618–626). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-11-12, from

- <https://dl.acm.org/doi/10.1145/2020408.2020505> doi: 10.1145/2020408.2020505
- Wang, Z., Xie, Q., & Xia, R. (2023, July). 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* (pp. 1765–1770). New York, NY, USA: Association for Computing Machinery. Retrieved 2025-09-24, from <https://dl.acm.org/doi/10.1145/3539618.3591940> doi: 10.1145/3539618.3591940
- Wankhade, M., Kulkarni, C., & Rao, A. C. S. (2024, December). A survey on aspect base sentiment analysis methods and challenges. *Applied Soft Computing*, 167, 112249. Retrieved 2024-11-05, from <https://www.sciencedirect.com/science/article/pii/S1568494624010238> doi: 10.1016/j.asoc.2024.112249
- Wiegand, M., Bocionek, C., Conrad, A., Dembowski, J., Giesen, J., Linn, G., & Schmeling, L. (2014). Saarland university's participation in the german sentiment analysis shared task (gestalt). In *Workshop proceedings of the 12th edition of the konvens conference, hildesheim, germany, october 8-10, 2014* (pp. 174–184).
- Wilcoxon, F. (1992). Individual Comparisons by Ranking Methods. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 196–202). New York, NY: Springer New York. Retrieved 2025-09-04, from http://link.springer.com/10.1007/978-1-4612-4380-9_16 (Series Title: Springer Series in Statistics) doi: 10.1007/978-1-4612-4380-9_16
- Wogenstein, F., Drescher, J., Reinel, D., Rill, S., & Scheidt, J. (2013, August). Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (pp. 1–8). Chicago Illinois: ACM. Retrieved 2025-09-26, from <https://dl.acm.org/doi/10.1145/2502069.2502074> doi: 10.1145/2502069.2502074
- Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., & Biemann, C. (2017). GermEval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*, 1–12.
- Wu, C., Ma, B., Liu, Y., Zhang, Z., Deng, N., Li, Y., ... Xue, Y. (2025, February). M-ABSA: A Multilingual Dataset for Aspect-Based Sentiment Analysis. arXiv. Retrieved 2025-04-23, from <http://arxiv.org/abs/2502.11824> (arXiv:2502.11824 [cs]) doi: 10.48550/arXiv.2502.11824

- Wu, C., Ma, B., Zhang, Z., Deng, N., He, Y., & Xue, Y. (2025, June). Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models. *International Journal of Machine Learning and Cybernetics*. Retrieved 2025-09-29, from <https://link.springer.com/10.1007/s13042-025-02711-z> doi: 10.1007/s13042-025-02711-z
- Xu, H., Zhang, D., Zhang, Y., & Xu, R. (2024, August). HITSZ-HLT at SIGHAN-2024 dimABSA Task: Integrating BERT and LLM for Chinese Dimensional Aspect-Based Sentiment Analysis. In K.-F. Wong et al. (Eds.), *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)* (pp. 175–185). Bangkok, Thailand: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2024.sighan-1.20/>
- Xu, T., Yang, H., Wu, Z., Chen, J., Zhao, F., & Dai, X. (2023). Measuring your aste models in the wild: A diversified multi-domain dataset for aspect sentiment triplet extraction. In *The 61st annual meeting of the association for computational linguistics*.
- Yang, S., Cho, H., Lee, J., Yoon, S., Choi, E., Choo, J., & Cho, W. I. (2025, April). Single Ground Truth Is Not Enough: Adding Flexibility to Aspect-Based Sentiment Analysis Evaluation. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 12071–12096). Albuquerque, New Mexico: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2025.naacl-long.603/> doi: 10.18653/v1/2025.naacl-long.603
- Yimam, S. M., Biemann, C., De Castilho, R. E., & Gurevych, I. (2014). Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 91–96). Retrieved 2025-05-08, from <https://aclanthology.org/P14-5016.pdf>
- Yin, Y., Song, Y., & Zhang, M. (2017, September). Document-Level Multi-Aspect Sentiment Classification as Machine Comprehension. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2044–2054). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved 2024-11-12, from <https://aclanthology.org/D17-1217> doi: 10.18653/v1/D17-1217
- Zainuddin, N., Selamat, A., & Ibrahim, R. (2016). Twitter Feature Selection and Classification Using Support Vector Machine for Aspect-Based Sentiment Analysis. In H. Fujita, M. Ali, A. Selamat, J. Sasaki, & M. Kurematsu (Eds.), *Trends in*

- Applied Knowledge-Based Systems and Data Science* (pp. 269–279). Cham: Springer International Publishing. doi: 10.1007/978-3-319-42007-3_23
- Zhang, L., Liu, B., Lim, S. H., & O’Brien-Strain, E. (2010). Extracting and Ranking Product Features in Opinion Documents. In *Coling 2010: posters* (pp. 1462–1470).
- Zhang, M., & Qian, T. (2020, November). Convolution over Hierarchical Syntactic and Lexical Graphs for Aspect Level Sentiment Analysis. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3540–3549). Online: Association for Computational Linguistics. Retrieved 2025-09-24, from <https://aclanthology.org/2020.emnlp-main.286/> doi: 10.18653/v1/2020.emnlp-main.286
- Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., & Lam, W. (2021, November). Aspect Sentiment Quad Prediction as Paraphrase Generation. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 9209–9219). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved 2025-09-16, from <https://aclanthology.org/2021.emnlp-main.726/> doi: 10.18653/v1/2021.emnlp-main.726
- Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024, June). Sentiment Analysis in the Era of Large Language Models: A Reality Check. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3881–3906). Mexico City, Mexico: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2024.findings-naacl.246/> doi: 10.18653/v1/2024.findings-naacl.246
- Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2023, November). A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11), 11019–11038. Retrieved 2024-10-24, from https://ieeexplore.ieee.org/abstract/document/9996141?casa_token=3a7CivxoMLMAAAA:9aT2RjIW9830-IQ8bdWuD84WqCoIFSnn8qzd26OvVG3ETsp3neSuI1VjRwqF-1stSkZHcUSRuNsP8w (Conference Name: IEEE Transactions on Knowledge and Data Engineering) doi: 10.1109/TKDE.2022.3230975
- Zhou, C., Song, D., Tian, Y., Wu, Z., Wang, H., Zhang, X., ... Zhang, S. (2024, December). *A Comprehensive Evaluation of Large Language Models on Aspect-Based Sentiment Analysis*. arXiv. Retrieved 2025-09-24, from <http://arxiv.org/abs/2412.02279> (arXiv:2412.02279 [cs]) doi: 10.48550/arXiv.2412.02279
- Zhu, S., Zhao, H., Wang, X., Liu, S., Jia, Y., & Zan, H. (2024, August). ZJU-NLP at SIGHAN-2024 dimABSA Task: Aspect-Based Sentiment Analysis with Coarse-

- to-Fine In-context Learning. In K.-F. Wong et al. (Eds.), *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)* (pp. 112–120). Bangkok, Thailand: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2024.sighan-1.13/>
- Šmíd, J., Priban, P., & Kral, P. (2024, August). LLaMA-Based Models for Aspect-Based Sentiment Analysis. In O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, & S. Tafreshi (Eds.), *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 63–70). Bangkok, Thailand: Association for Computational Linguistics. Retrieved 2025-09-09, from <https://aclanthology.org/2024.wassa-1.6/> doi: 10.18653/v1/2024.wassa-1.6

A. Appendix

A.1. Prompts Examples 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 'Sentiment-Polarität' 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', 'Sentiment-Polarität', 'Aspektbegriff'), ...].
```

`[[examples]]`

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

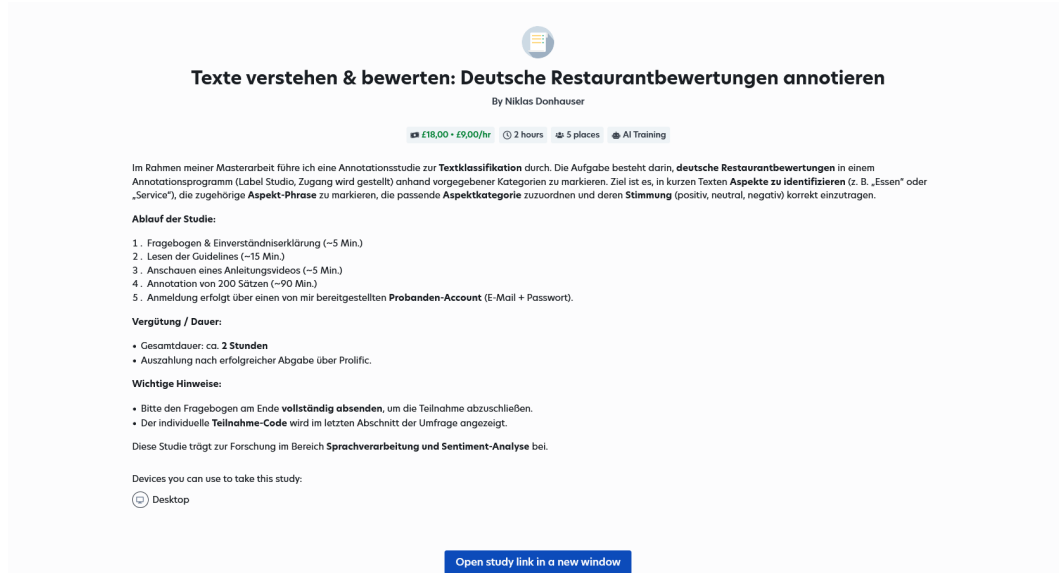

```

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.

A.2. Study Description on Prolific



Texte verstehen & bewerten: Deutsche Restaurantbewertungen annotieren
By Niklas Donhauser

£18.00 • £9.00/hr • 2 hours • 5 places • AI Training

Im Rahmen meiner Masterarbeit führe ich eine Annotationsstudie zur **Textklassifikation** durch. Die Aufgabe besteht darin, **deutsche Restaurantbewertungen** in einem Annotationsprogramm (Label Studio, Zugang wird gestellt) anhand vorgegebener Kategorien zu markieren. Ziel ist es, in kurzen Texten **Aspekte zu identifizieren** (z. B. „Essen“ oder „Service“), die zugehörige **Aspekt-Phrase** zu markieren, die passende **Aspektkategorie** zuzuordnen und deren **Stimmung** (positiv, neutral, negativ) korrekt einzutragen.

Ablauf der Studie:

1. Fragebogen & Einverständniserklärung (~5 Min.)
2. Lesen der Guidelines (~15 Min.)
3. Anschauen eines Anleitungsvideos (~5 Min.)
4. Annotation von 200 Sätzen (~90 Min.)
5. Anmeldung erfolgt über einen von mir bereitgestellten **Probanden-Account** (E-Mail + Passwort).

Vergütung / Dauer:

- Gesamtdauer: ca. **2 Stunden**
- Auszahlung nach erfolgreicher Abgabe über Prolific.

Wichtige Hinweise:

- Bitte den Fragebogen am Ende **vollständig absenden**, um die Teilnahme abzuschließen.
- Der individuelle **Teilnahme-Code** wird im letzten Abschnitt der Umfrage angezeigt.

Diese Studie trägt zur Forschung im Bereich **Spracheverarbeitung und Sentiment-Analyse** bei.

Devices you can use to take this study:

☒ Desktop

[Open study link in a new window](#)

Figure 11: Study description shown to participants on Prolific.

A.3. Annotation Example

Aspect Category	ID	Triplets	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 13.: Examples of ground truth annotations showing all aspect categories, IDs, extracted triplets, and their corresponding sentences.

A.4. Statistical Testing

The following section discusses the adjusted p-values from Holm-Bonferroni correction in more detail.

ACSA

For each model, the Shapiro–Wilk test was used to assess normality of the seed-averaged micro-F1 scores. *BERT-CLF* ($W = 0.9397$, $p = 0.2366$), *Hier-CGN* ($W = 0.9444$, $p = 0.2894$), *LLM-FS* ($W = 0.9752$, $p = 0.8581$) and *LLM-FT* ($W = 0.9658$, $p = 0.6645$) met the normality assumption.

Accordingly, repeated-measures one-way ANOVAs were applied for all models. The overall test results indicated no significant effect of dataset on performance for *BERT-CLF* (ANOVA $p = 0.3622$) and *Hier-CGN* (ANOVA $p = 0.0357$), whereas significant effects were observed for *LLM-FS* (ANOVA $p = 0.0003$) and *LLM-FT* (ANOVA $p = 0.0069$). These overall test results are not shown in Table 14, which reports only the Holm–Bonferroni corrected pairwise comparisons.

Model	Crowd Experts	Crowd LLM	Crowd Students	Experts LLM	Experts Students	LLM Students
BERT-CLF	0.40	0.40	0.18	0.40	0.42	0.40
Hier-CGN	1.00	0.57	0.46	0.59	0.04	1.00
LLM-FS	0.43	0.24	0.43	0.03	0.43	0.03
LLM-FT	0.22	0.56	0.65	0.11	0.56	0.56

Table 14.: Adjusted p-values (Holm–Bonferroni) for pairwise comparisons of micro-F1 across datasets per model for the ACSA task. Significant values ($p < 0.05$) are in bold.

TASD

For each model, the Shapiro–Wilk test was used to assess normality of the seed-averaged micro-F1 scores. *Paraphrase* ($W = 0.9601$, $p = 0.5468$) and *MvP* ($W = 0.9346$, $p = 0.1892$) met the normality assumption, whereas *LLM-FS* ($W = 0.9042$, $p = 0.0495$) and *LLM-FT* ($W = 0.9034$, $p = 0.0477$) violated it.

Accordingly, repeated-measures one-way ANOVAs were applied for *Paraphrase* and *MvP*, while Friedman tests were used for *LLM-FS* and *LLM-FT*. The overall test results indicated a significant effect of dataset on performance for *Paraphrase* (ANOVA $p < 0.001$), *MvP* (ANOVA $p < 0.001$), *LLM-FS* (Friedman $p = 0.0029$), and *LLM-FT* (Friedman $p = 0.0018$). These overall test results are not shown in Table 15, which reports only the Holm–Bonferroni corrected pairwise comparisons.

Model	Crowd Experts	Crowd LLM	Crowd Students	Experts LLM	Experts Students	LLM Students
Paraphrase	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	0.96
MvP	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
LLM-FS	0.38	0.38	0.38	0.38	0.38	0.38
LLM-FT	0.38	0.38	0.38	0.38	0.38	0.38

Table 15.: Adjusted p-values (Holm–Bonferroni) for pairwise comparisons of micro-F1 across datasets per model for the TASD task. Significant values ($p < 0.05$) are in bold.

A.5. Accuracy Scores for ACSA and TASD

Aspect Category Sentiment Analysis (ACSA)				
Method / Dataset	Crowd	Students	LLM	Experts
BERT-CLF	62.60	63.69	63.20	64.29
Hier-GCN	66.20	65.26	65.48	66.37
Gemma FS	75.49	76.11	74.83	75.89
LLaMA FT	74.90	75.00	73.69	76.05

Table 16.: Accuracy scores for ACSA, averaged over five seeds across datasets. Bold indicates the highest values

Target Aspect Sentiment Detection (TASD)				
Method / Dataset	Crowd	Students	LLM	Experts
Paraphrase	35.85	40.19	40.22	44.57
MvP	34.49	39.70	43.53	47.07
Gemma FS	41.40	45.23	48.79	46.69
LLaMA FT	48.66	53.07	49.53	55.60

Table 17.: Accuracy scores for TASD, averaged over five seeds across datasets. Bold indicates the highest values.

A.6. Training Time and Memory Usage

Task	Model	Training Time	GPU Memory
ACSA	BERT-CLF	06 m	2.7 GB
	Hier-GCN	10 m	3.2 GB
	LLaMA Fine-Tune	20 m	10.2 GB
TASD	Paraphrase	06 m	11.4 GB
	MvP	58 m	11.8 GB
	LLaMA Fine-Tune	38 m	14.3 GB

Table 18.: Training times and GPU memory usage per model and task on the GER-restaurant (LLM) dataset, averaged over five seeds. Because of the similar size of the training sets and the shared test set, we only report the values for one dataset.

A.7. Label Interface ACSA

Essen durchschnittlich, Preisleistung passt nicht.

Aspekt-Label

⚠ Hinweis: Wähle nur Kategorien, die im Text tatsächlich erwähnt oder angesprochen werden. Wenn eine Kategorie im Text nicht vorkommt, vergebe kein Label (auch kein „Neutral“).

Essen 🍽 Essen-Positiv¹¹ ☐ Essen-Negativ¹² ☐ Essen-Neutral¹³

Service 🍽 Service-Positiv¹⁴ ☐ Service-Negativ¹⁵ ☐ Service-Neutral¹⁶

Ambiente 🍽 Ambiente-Positiv¹⁷ ☐ Ambiente-Negativ¹⁸ ☐ Ambiente-Neutral¹⁹

Gesamteindruck 🍽 Gesamteindruck-Positiv²⁰ ☐ Gesamteindruck-Negativ²¹ ☐ Gesamteindruck-Neutral²²

Preis 💰 Preis-Positiv²³ ☐ Preis-Negativ²⁴ ☐ Preis-Neutral²⁵

Metadaten

☐ Die Annotation war schwierig.²⁶

☐ Fehlender Kontext aus der Bewertung könnte die Annotation beeinflussen.²⁷

Zusätzliche Kommentare...

Die Annotationsanleitung kann hier abgerufen werden: https://drive.google.com/file/d/1DDu1H_b6d1UNfeoltzT-QDDRMzUHN6p6/view?usp=sharing

Figure 12: Label interface for the ACSA task in Label Studio.

A.8. Dataset Statistics

Original Test Set								
Aspect Category	Positive		Negative		Neutral		Total	
	Expli.	Impli.	Expli.	Impli.	Expli.	Impli.	Expli.	Impli.
AMBIENCE	80	5	27	11	1	0	108	16
FOOD	244	22	157	19	37	0	438	41
GENERAL	29	100	26	98	3	7	58	205
PRICE	11	0	40	5	9	4	60	9
SERVICE	145	18	108	41	3	0	256	59
Total	509	145	358	174	53	11	920	330

Table 19.: Ground truth distribution of aspect categories across sentiment polarities and reference types for the original GERestaurant test sets.

New Annotated Test Set								
Aspect Category	Positive		Negative		Neutral		Total	
	Expli.	Impli.	Expli.	Impli.	Expli.	Impli.	Expli.	Impli.
AMBIENCE	83	7	24	14	2	0	109	21
FOOD	260	22	159	29	37	3	456	54
GENERAL	23	96	18	89	1	9	42	194
PRICE	14	0	45	11	5	4	64	15
SERVICE	146	21	94	56	3	0	243	77
Total	526	146	340	199	48	16	914	361

Table 20.: Ground truth distribution of aspect categories across sentiment polarities and reference types for the new annotated test sets.

Crowd Train Set								
Aspect Category	Positive		Negative		Neutral		Total	
	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.
AMBIENCE	81	3	26	8	0	0	107	11
FOOD	232	19	134	26	21	3	387	48
GENERAL	26	77	16	62	1	1	43	140
PRICE	8	2	20	16	2	1	30	19
SERVICE	128	10	79	34	3	1	210	45
Total	475	111	275	146	27	6	777	263

Table 21.: Counts of polarity triplets by category, with explicit/implicit split for the crowd dataset.

Student Train Set								
Aspect Category	Positive		Negative		Neutral		Total	
	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.
AMBIENCE	91	0	33	2	1	0	125	2
FOOD	238	14	135	18	32	1	405	33
GENERAL	35	74	13	67	1	0	49	141
PRICE	14	0	28	10	5	0	47	10
SERVICE	140	16	82	39	3	0	225	55
Total	518	104	291	136	42	1	851	241

Table 22.: Counts of polarity triplets by category, with explicit/implicit split for the student dataset

LLM Train Set								
Aspect Category	Positive		Negative		Neutral		Total	
	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.
AMBIENCE	104	7	33	9	5	0	142	16
FOOD	311	19	182	40	45	2	538	61
GENERAL	5	122	3	98	0	9	8	229
PRICE	19	1	30	12	6	0	55	13
SERVICE	157	28	88	62	8	0	253	90
Total	596	177	336	221	64	11	996	409

Table 23.: Counts of polarity triplets by category, with explicit/implicit split for the LLM dataset.

Experts Train Set								
Aspect Category	Positive		Negative		Neutral		Total	
	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.
AMBIENCE	110	15	40	10	2	0	152	25
FOOD	288	26	167	40	38	4	493	70
GENERAL	33	103	17	90	2	5	52	198
PRICE	15	1	42	12	6	0	63	13
SERVICE	167	24	98	68	3	0	268	92
Total	613	169	364	220	51	9	1028	398

Table 24.: Counts of polarity triplets by category, with explicit/implicit split for the Experts dataset.

Original Train Set								
Aspect Category	Positive		Negative		Neutral		Total	
	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.
AMBIENCE	110	15	46	7	2	0	158	22
FOOD	296	25	175	36	37	4	508	65
GENERAL	33	99	17	89	3	6	53	194
PRICE	16	1	42	12	6	0	64	13
SERVICE	165	23	108	62	3	0	276	85
Total	620	163	388	206	51	10	1059	379

Table 25.: Counts of polarity triplets by category, with explicit/implicit split for the original GERestaurant train set.

A.9. Category and Polarity F1-scores

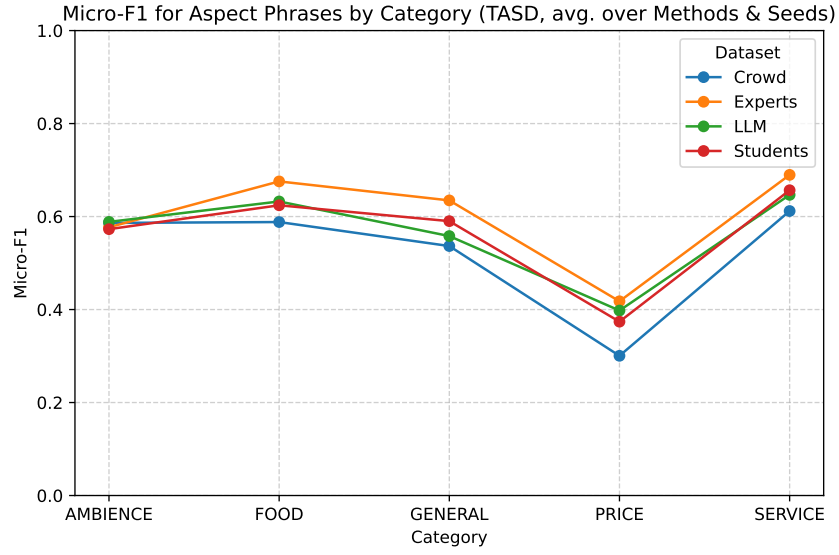


Figure 13: Micro-F1 scores for the TASD task across categories, phrase prediction. Results are averaged over methods and seeds for each dataset.

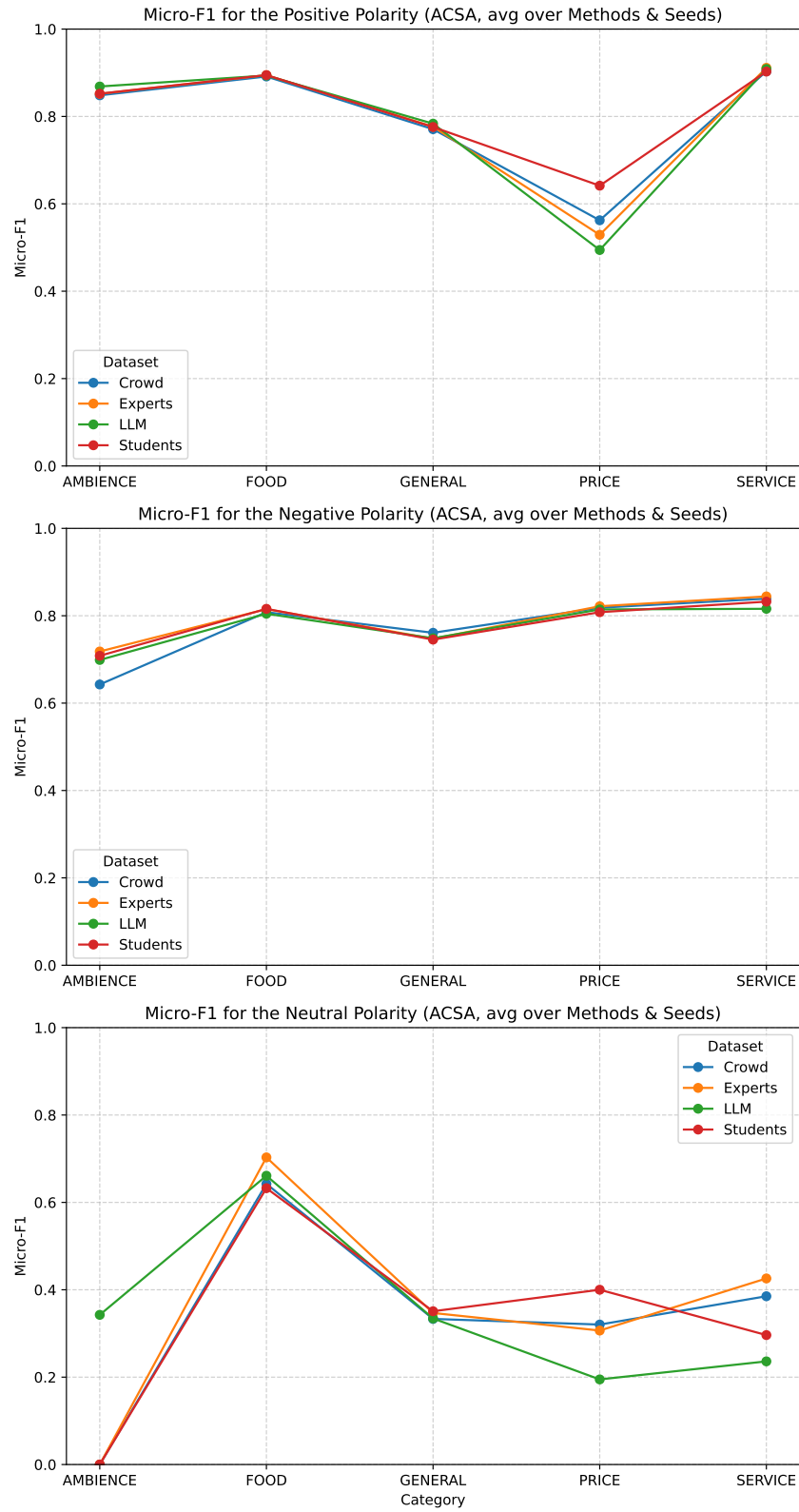


Figure 14: Micro-F1 scores for the ACSA task across categories, separated by polarity (positive, negative, neutral). Results are averaged over methods and seeds for each dataset.

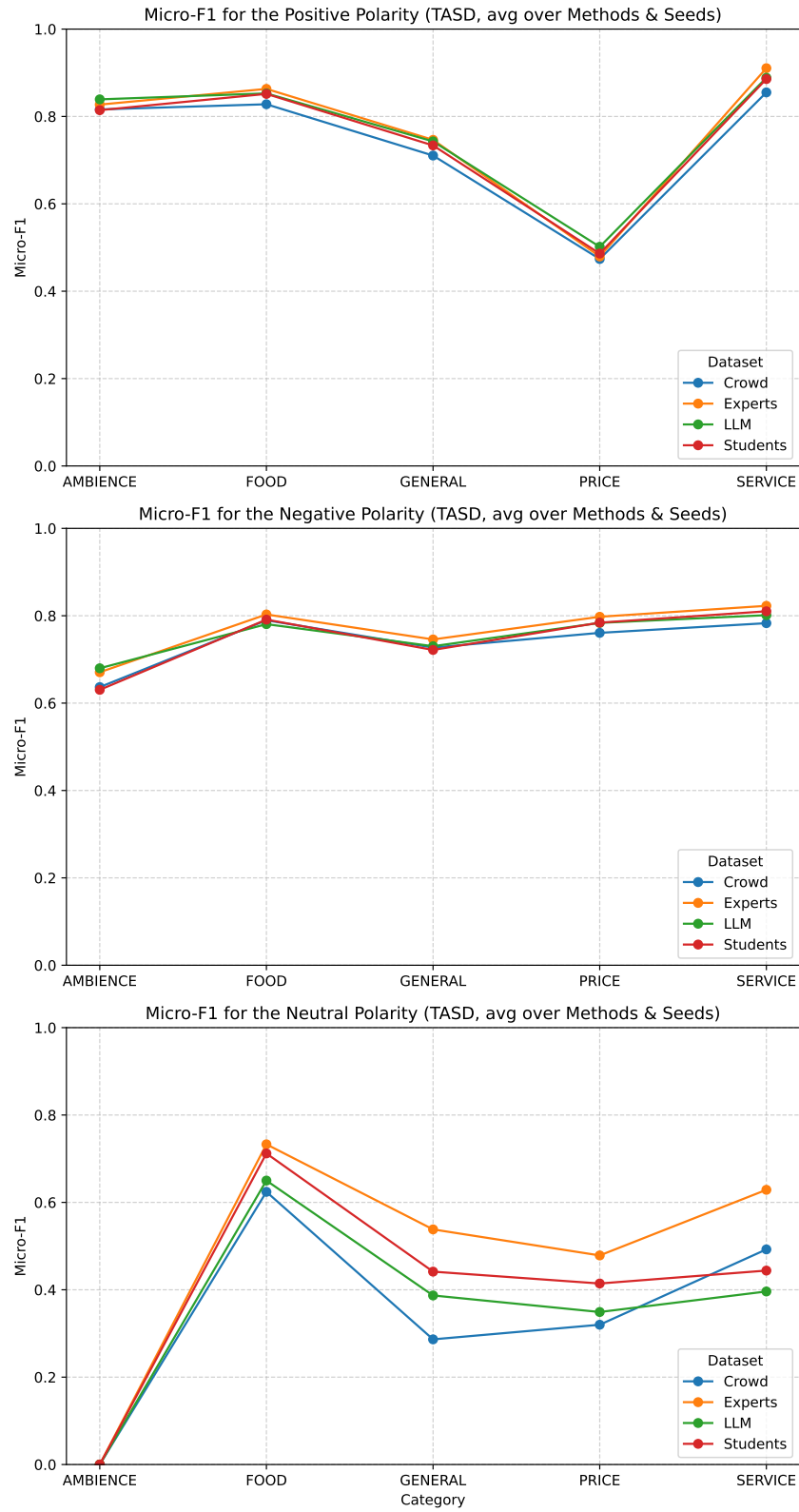


Figure 15: Micro-F1 scores for the TASD task across categories, separated by polarity (positive, negative, neutral). Results are averaged over methods and seeds for each dataset.

Erklärung zur Urheberschaft

Die vorgelegten Druckexemplare sowie die vorgelegte digitale Version der Arbeit sind identisch.

Ich habe die Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit nicht bereits an einer anderen Hochschule zur Erlangung eines akademischen Grades eingereicht.

Von den zu § 27 Abs. 5 der Prüfungsordnung vorgesehenen Rechtsfolgen habe ich Kenntnis.

Regensburg, 30.09.2025

Signature

Erklärung zur Lizenzierung und Publikation dieser Arbeit

Name: Niklas Donhauser

Titel der Arbeit: *Annotation Quality and Its Influence on Aspect-Based Sentiment Analysis: A Case Study on German Restaurant Reviews*

Hiermit gestatte ich die Verwendung der schriftlichen Ausarbeitung zeitlich unbegrenzt und nicht-exklusiv unter folgenden Bedingungen:

- ☐ Nur zur Bewertung dieser Arbeit
- ☐ Nur innerhalb des Lehrstuhls im Rahmen von Forschung und Lehre
- ☒ Unter einer Creative-Commons-Lizenz mit den folgenden Einschränkungen:
 - ☒ BY – Namensnennung des Autors
 - ☐ NC – Nichtkommerziell
 - ☐ SA – Share-Alike, d.h. alle Änderungen müssen unter die gleiche Lizenz gestellt werden.

(An Zitaten und Abbildungen aus fremden Quellen werden keine weiteren Rechte eingeräumt.)

Außerdem gestatte ich die Verwendung des im Rahmen dieser Arbeit erstellten Quellcodes unter folgender Lizenz:

- ☐ Nur zur Bewertung dieser Arbeit
- ☐ Nur innerhalb des Lehrstuhls im Rahmen von Forschung und Lehre
- ☐ Unter der CC-0-Lizenz (= beliebige Nutzung)
- ☒ Unter der MIT-Lizenz (= Namensnennung)
- ☐ Unter der GPLv3-Lizenz (oder neuere Versionen)

(An explizit mit einer anderen Lizenz gekennzeichneten Bibliotheken und Daten werden keine weiteren Rechte eingeräumt.)

Ich willige ein, dass der Lehrstuhl für Medieninformatik diese Arbeit – falls sie besonders gut ausfällt - auf dem Publikationsserver der Universität Regensburg veröffentlichen lässt.

Ich übertrage deshalb der Universität Regensburg das Recht, die Arbeit elektronisch zu speichern und in Datennetzen öffentlich zugänglich zu machen. Ich übertrage der Universität Regensburg ferner das Recht zur Konvertierung zum Zwecke der Langzeitarchivierung unter Beachtung der Bewahrung des Inhalts (die Originalarchivierung bleibt erhalten).

Erklärung zur Lizenzierung und Publikation dieser Arbeit

Ich erkläre außerdem, dass von mir die urheber- und lizenzrechtliche Seite (Copyright) geklärt wurde und Rechte Dritter der Publikation nicht entgegenstehen.

- ☒ Ja, für die komplette Arbeit inklusive Anhang
- ☐ Ja, für eine um vertrauliche Informationen gekürzte Variante (auf dem Datenträger beigefügt)
- ☐ Nein
- ☐ Sperrvermerk bis (Datum):

Regensburg, 30.09.2025

Signature

Inhalt des beigefügten Datenträgers

01_Paper	Final written thesis as a PDF file, including all figures used.
02_Code	Source code for model implementations, dataset generation, preprocessing, and evaluation.
03_Datasets	All created datasets and the individual annotations from participants and models.
04_Annotations	Annotation guidelines, interface materials, and instructional videos used in the annotation studies.
05_Results	Outputs of this work, including model evaluation results and inter-annotator agreement scores.
06_Presentations	Initial project presentation and final presentation.
07_Sources	Bibliographic sources and references cited in the thesis.
08_Miscellaneous	Additional artifacts and data generated during the research.
