GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis

Nils Constantin Hellwig
Media Informatics Group
University of Regensburg
Regensburg, Germany
nils-constantin.hellwig@ur.de

Jakob Fehle
Media Informatics Group
University of Regensburg
Regensburg, Germany
jakob.fehle@ur.de

Markus Bink
Media Informatics Group
University of Regensburg
Regensburg, Germany
markus.bink@student.ur.de

Christian Wolff
Media Informatics Group
University of Regensburg
Regensburg, Germany
christian.wolff@ur.de

Abstract

We present GERestaurant, a novel dataset consisting of 3,078 German language restaurant reviews manually annotated for Aspect-Based Sentiment Analysis (ABSA). All reviews were collected from Tripadvisor, covering a diverse selection of restaurants, including regional and international cuisine with various culinary styles. The annotations encompass both implicit and explicit aspects, including all aspect terms, their corresponding aspect categories, and the sentiments expressed towards them. Furthermore, we provide baseline scores for the four ABSA tasks Aspect Category Detection, Aspect Category Sentiment Analysis, End-to-End ABSA and Target Aspect Sentiment Detection as a reference point for future advances. The dataset fills a gap in German language resources and facilitates exploration of ABSA in the restaurant domain.

1 Introduction

Sentiment analysis (SA), also named opinion mining, is a research area in natural language processing (NLP) which involves the computational classification of individuals' sentiments, opinions and emotions. This usually involves categorizing sentiments into three polarities: positive, neutral and negative.

SA can be applied at both document- (Hellwig et al., 2023; Schmidt et al., 2022; Tripathy et al., 2017) and sentence-level (Liu, 2010). However, if a document or sentence comprises a mixture of different sentiments, it's often impossible to assign a solely positive, negative or neutral label. As an example, consider the sentence "The salad tasted wonderful, but was quite expensive." of a restaurant review wherein positive sentiment is expressed towards the food while, concurrently, negative sentiment is expressed when addressing the food's price. To overcome this issue, Aspect-Based Sentiment Analysis (ABSA) has been extensively studied as it goes beyond assessing general sentiment and instead delves into a more granular examination of sentiment by linking particular aspects with corresponding sentiment polarities (Liu et al., 2005; Pontiki et al., 2015).

In this work, we introduce GERestaurant, a novel dataset comprising 3,078 German language restaurant reviews annotated for ABSA. It's the first German language dataset of sentences from restaurant reviews for ABSA. The annotations included the aspect term (if available), an aspect category selected from a predefined set of categories, and the sentiment or polarity expressed towards the aspect. The dataset is provided as a benchmark dataset for future research and parallels the widely used SemEval 2015 and 2016 restaurant datasets in terms of annotation scheme and annotation guidelines (Pontiki et al., 2015, 2016). Thus, it not only contributes to the availability of German language resources but also enables the exploration of new ABSA methods in the restaurant domain in the German language. Additionally, we provide a baseline performance by fine-tuning state-of-the-art (SOTA) transformer-based language models on the annotated dataset for typical ABSA tasks: Aspect Category Detection (ACD), Aspect Category Sentiment Analysis (ACSA), End-to-End ABSA

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 123–133

(E2E-ABSA) and Target Aspect Sentiment Detection (TASD).

2 Related Work

ABSA has attracted increasing attention, in part due to benchmark datasets and shared tasks from various domains that facilitated the development of machine learning approaches for solving ABSA tasks. For instance, various datasets from different domains frequently employed in ABSA research include:

- Ganu et al. (2009): A dataset comprising restaurant reviews in English, annotated with six pre-defined aspect categories assigned to sentiment polarities positive, neutral, negative, and conflict.
- Saeidi et al. (2016): SentiHood, a dataset of English sentences extracted from a question answering (QA) platform discussing urban neighbourhoods. Annotations for aspect terms, their associated aspect categories, and the sentiment expressed towards them were provided.
- Jiang et al. (2019): MAMS, a dataset of English Tweets on celebrities, products, and companies. All aspect terms were annotated, along with the sentiment polarity expressed towards them.

However, the development of methods addressing the subtasks in ABSA was particularly driven by the SemEval shared task workshop in the years from 2014 to 2016 and the associated publishing of human-annotated datasets for ABSA. These comprised sentences from reviews of laptops and restaurants.

SemEval-2014 Task 4 (Pontiki et al., 2014) was dedicated to ABSA and included annotations of aspect terms and the sentiment polarity expressed towards them. In addition, annotations of the aspect categories and the sentiment polarity expressed towards them are part of the provided dataset.

In the subsequent year, SemEval-2015 Task 12 (Pontiki et al., 2015) was published, which included annotations of all aspect terms, their corresponding aspect category and the sentiment polarity expressed towards the aspect terms. Moreover, annotations of implicit aspects were provided, meaning cases where a sentiment was expressed towards an aspect category, without the presence of an aspect term. In such cases, the aspect term was annotated as `"NULL"`.

SemEval-2016 Task 5 (Pontiki et al., 2016) encompassed the same three sentiment elements as SemEval-2015 Task 12 (Pontiki et al., 2015). In addition, subsets containing annotated sentences of hotel reviews and reviews in languages other than English were provided for each domain (Pontiki et al., 2015).

When examining datasets in German language, there is a scarcity of annotated datasets. The most prominent dataset in German language is the dataset published as part of the GermEval 2017 shared task (Wojatzki et al., 2017), which includes customer reviews concerning the "Deutsche Bahn", the German public train operator (Wojatzki et al., 2017). Reviews were annotated as a whole, rather than individual sentences separately (Wojatzki et al., 2017). Similar to the datasets introduced by Pontiki et al. (2015, 2016), annotations were provided for all aspect terms, their associated aspect categories, and the sentiment expressed towards the aspect terms.

Gabryszak and Thomas (2022) introduced the German language dataset MobASA, which comprises tweets from public transportation companies and channels related to barrier-free travel for handicapped passengers (Gabryszak and Thomas, 2022). Annotations covered aspect terms, their associated aspect categories, and the sentiments expressed towards each aspect term (Gabryszak and Thomas, 2022).

In the realm of customer reviews, other notable resources include the SCARE corpus (Sänger et al., 2016), comprising annotated application reviews from the Google Play Store, alongside annotations for aspect terms and sentiment polarities. Similarly, Fehle et al. (2023) introduced a dataset consisting of sentences from hotel reviews on Tripadvisor, whereby annotations are provided for the sentiments expressed towards the considered aspect categories.

3 Methodology

3.1 Data Acquisition

To gather German language restaurant reviews, Tripadvisor was selected as the data source. The five restaurants with the most customer reviews in the 25 most densely populated German cities as of

2022[1] were considered, covering a wide spectrum of restaurant types, including regional and international cuisine with various culinary styles. In the course of the COVID-19 pandemic, restaurant reviews were influenced by the associated hygiene measures. To prevent sentiment bias introduced by hygiene regulations we included all reviews posted during a period without mandated COVID-19 hygiene restrictions, specifically reviews from October 15, 2022, to October 15, 2023, were taken into account.

Overall, a total of 3,212 user reviews with a German language label on Tripadvisor were collected. The reviews were segmented into 13,426 sentences using the NLTK Tokenizer (Loper and Bird, 2002). It was observed that, despite the German language code label, some sentences were in languages other than German. Due to this, *langdetect*[2] was employed to ascertain the language of each sentence, leading to the rejection of 631 sentences which resulted in a total of 12,795 sentences in German.

Ultimately, the sentences underwent an anonymization process. Named entity recognition (NER) was employed using *spaCy* (*de_core_news_lg* model) (Honnibal and Montani, 2017) to replace locations, personal names, and time-related references with anonymized placeholders *"LOC"*, *"PERSON"* and *"DATE"*. Subsequently, regular expressions were employed to substitute any mentions of the restaurant's name in the sentences with the placeholder *"RESTAURANT_NAME"*.

[1]German Federal Statistical Office, population and population density as of December 31, 2022: https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Administrativ/05-staedte.html

[2]https://pypi.org/project/langdetect

3.2 Data Annotation

From the complete set of 12,795 sentences, a subset of 5,000 sentences was randomly sampled for annotation. Care was taken to ensure an equal distribution of sentences from reviews with 1, 2, 3, 4 and 5-star ratings (1,000 sentences each) [3]. This distribution was established so that each sentiment polarity occurs as evenly as possible across all sentences.

3.2.1 Annotation Task

As proceeded for SemEval-2015 (Pontiki et al., 2015) and SemEval-2016 (Pontiki et al., 2016), for a given sentence *x*, one or multiple triplets *(a, c, p)* should be assigned, where *a* represents the aspect term, *c* denotes the aspect category, and *p* indicates the sentiment expressed towards the aspect. The annotations included the positional information of the aspect terms within the text. Multiple aspect terms could be assigned to the same aspect category. Similarly, an aspect term could be assigned to multiple aspect categories at once. Examples are presented in Table 1 and an English language translation of the table is provided in Appendix A.1.

The four aspect categories `FOOD`, `SERVICE`, `AMBIENCE` and `PRICE` were considered, similar to the rating categories of the Zagat Survey (Lee and Teng, 2007) for restaurants. These categories can also be found on Tripadvisor, allowing users to optionally assign one to five stars to each category in addition to an overall star rating.

However, in contrast to the categories from the Zagat Survey and as preceded by Pontiki et al. (2015), `AMBIENCE` was used as an aspect category instead of *"Decor"* as it encompasses a slightly

[3]A customer reviewing a restaurant on Tripadvisor is obligated to provide both a star rating and a textual assessment.

Aspect Category	Triplets	Sentence
GENERAL-IMPRESSION	[('Restaurant', 'GENERAL-IMPRESSION', 'POSITIVE')]	*"Sehr schönes Restaurant."*
FOOD	[('Bratwurst', 'FOOD', 'POSITIVE')]	*"Die Bratwurst war unglaublich lecker und perfekt gewürzt."*
SERVICE	[('Bedienung', 'SERVICE', 'NEGATIVE')]	*"Bedienung leider nicht aufmerksam."*
AMBIENCE	[('NULL', 'AMBIENCE', 'NEGATIVE')]	*"Es war viel zu laut, wie im Club."*
PRICE	[('NULL', 'PRICE', 'NEUTRAL')]	*"Preislich ist das ok gewesen."*
PRICE, SERVICE	[('Preise', 'PRICE', 'NEUTRAL') ('Service', 'SERVICE', 'NEGATIVE')]	*"Preise sind ok und Service auch."*
FOOD, AMBIENCE, SERVICE	[('Essen', 'FOOD', 'POSITIVE'), ('Atmosphäre', 'AMBIENCE', 'POSITIVE'), ('Service', 'SERVICE', 'POSITIVE')]	*"Tolles Essen, tolle Atmosphäre und ganz netter und aufmerksamer Service!"*

Table 1: Annotated examples for all aspect categories.

broader scope. Furthermore, a fifth category called GENERAL-IMPRESSION was introduced, which captured aspects that pertain to the restaurant in a general sense, similar to the datasets for ABSA published in the realm of SemEval-2015 (Pontiki et al., 2015) and SemEval-2016 (Pontiki et al., 2016), whereby an aspect category was introduced that encompassed general aspects related to a laptop or a restaurant for which a review was written.

Implicit addressing of an aspect category should be annotated as well. In this case, "NULL" was assigned as the aspect term. For each aspect term within these categories, one of the following sentiment polarity labels should be applied: POSITIVE, NEGATIVE, NEUTRAL (indicating mild positivity or mild negativity sentiment) or CONFLICT. The CONFLICT label was assigned in case both positive and negative sentiments are expressed towards an aspect term.

Furthermore, as preceded by Pontiki et al. (2015), aspects should only be annotated if a sentiment was expressed towards them. For instance, in the sentence *"You can eat pizza there"*, no sentiment is expressed towards the aspect *"Pizza"* (aspect category: FOOD), and thus, the aspect should not be annotated accordingly.

3.2.2 Data Labelling Procedure

Three persons were tasked with annotating sentences in order to establish the gold standard labels. Similar to the approach employed by Pontiki et al. (2014), the annotation process commenced with one annotator (annotator *A*, M.Sc. media computer science student) annotating all 5,000 sentences and subsequently, each of the annotations by annotator *A* underwent inspection and validation by a second annotator *B*.

For the second annotation, a PhD student and an M.Sc. student, both specializing in media computer science, were tasked to review 2,500 annotations by annotator *A* each. All annotators had prior experience in annotating textual data in the field of SA, with the PhD student having prior experience in annotating text for ABSA.

The annotation process was facilitated using *LabelStudio*[4]. All annotators were provided with a comprehensive annotation guideline[5], which explained the user interface in *LabelStudio* specifically created for this annotation task and included examples for sentences in German language closely aligned with those provided in the annotation guideline employed by Pontiki et al. (2015).

In addition to annotating all triplets *(a, c, p)*, annotators were tasked to tick a checkbox when they encountered two or more sentences in an example instead of one, since the NLTK Tokenizer employed for sentence segmentation could potentially introduce errors. Another checkbox was provided to mark examples where customers addressed an aspect without conveying any sentiment. This allowed for the possibility of annotating them at a later point in time for future studies.

In 113 out of 5,000 sentences, annotator *B* proposed a label different to that assigned by annotator *A*. Among these, Annotator *A* accepted the revised label suggested by annotator *B* in 81 sentences. The annotation of the remaining 32 sentences was decided in consensus of the two annotators. In 16 sentences, both annotators opted to adopt the annotations provided by annotator *A*, in seven instances, the annotation of annotator *B* was adhered to. For the remaining nine sentences, a consensus was reached on an annotation distinct from their initially proposed annotations.

Among the 5,000 examples, 589 were excluded since they consisted of more than one sentence. Subsequently, out of the remaining 4,411 sentences, 1,291 were omitted since no sentiment was expressed towards aspects of the considered aspect categories and 42 sentences were removed since they encompassed at least one triplet with a conflict polarity, resulting in a total of 3,078 sentences with a total of 3,149 explicit and 1,165 implicit aspects.

3.3 Baseline Models

For a total of four typical ABSA tasks, we provide transformer-based baseline models. All models were loaded using the Hugging Face *transformers* library[6] and trained on two NVIDIA RTX A5000 GPU with 24 GB VRAM each. The implementation was conducted using Python version 3.11.5. To assess the performance of each model, we conducted a random 70-30 train-test split. The models were trained on the training set, consisting of 2,154 examples, and evaluated on the test set, containing 924 examples.

[4]Label Studio - Open Source Data Labelling Tool: https://labelstud.io

[5]https://github.com/NilsHellwig/GERestaurant/blob/main/annotation_guideline.pdf

[6]https://pypi.org/project/transformers

3.3.1 Aspect Category Detection (ACD) and Aspect Category Sentiment Analysis (ACSA)

Similar to Fehle et al. (2023), the identification of aspect categories (ACD) and the identification of both aspect categories and the sentiment polarity expressed towards them (ACSA) was treated as a multi-label classification task. Two base models were fine-tuned in this study: gbert-large[7] (337 million parameters) and gbert-base[8] (111 million parameters) by *deepset*. Both models are based on the BERT architecture and are pre-trained on large amounts of German language texts (Chan et al., 2020).

For training and validation, a batch size of 16, an epoch-number of 3 and a learning rate of 2e-5 (c.f. Devlin et al. (2018)) was used. As proceeded by Fehle et al. (2023), a prediction was considered a true positive, if the predicted aspect(s) of a sentence (including the sentiment polarity for ACSA) occurred in the ground truth labels.

3.3.2 End-to-End ABSA (E2E-ABSA)

E2E-ABSA is the task that aims at simultaneously identifying aspect terms and determining the sentiment polarity expressed towards them in a given text. As proceeded by Li et al. (2019), E2E-ABSA was conducted employing a BERT model for token classification. gbert-large and gbert-base were employed for this task as well. The task involved predicting a tag sequence $y = \{y_1, \ldots, y_T\}$, with each tag corresponding to a token in the sentence. The potential values for y_t encompass B-$\{POS, NEG, NEU\}$, I-$\{POS, NEG, NEU\}$ or O. The tag denoted the beginning (B) and inside (I) of an aspect term, coupled with negative, neutral or positive sentiment and O, in case that a token was not a part of an aspect term.

For training, a binary cross-entropy loss was employed, and the sigmoid function was used as the activation function. Similar to the evaluations conducted by Li et al. (2019), learning rate was set to 2e-5, batch size was set to 16 and the model was trained for 1,500 steps. When calculating the evaluation metrics, the true positives included all correctly identified pairs of an aspect term and the sentiment polarity expressed towards it, similar to Zhang et al. (2023) and Li et al. (2019).

3.3.3 Target Aspect Sentiment Detection (TASD)

TASD is the task that leverages the full complexity of GERestaurants' annotations. Its objective is to identify all aspect terms, their associated aspect categories, and the sentiment expressed towards the aspect terms within a given text.

For the TASD task, the paraphrasing approach methodology introduced by Zhang et al. (2021) was employed. The paraphrase generation framework utilized is outlined in Appendix A.2. The polarity label `POSITIVE` was mapped to *"gut"* (Eng.: *"good"*) in the paraphrased label, `NEGATIVE` to *"schlecht"* (Eng.: *"bad"*) and `NEUTRAL` to *"ok"*. In the case of an implicit aspect, the aspect term was decoded as *"es"* (Eng.: *"it"*).

Both t5-large[9] (770 million parameters) and t5-base[10] (223 million parameters) were evaluated as the underlying seq2seq models. In terms of training parameters, batch size was set to 16, number of training epochs to 20 and learning rate to 3e-4, similar to Zhang et al. (2021). For evaluation, true positives encompassed all correctly identified triplets, meaning that all three sentiment elements (aspect term, aspect category and sentiment polarity) were identified correctly.

4 Results

4.1 Properties of the Annotated Dataset

Table 2 presents an overview of the frequency of triplets occurring with their respective aspect categories, reference types, and sentiment polarities in the overall dataset. The highest number of triplets was identified for the `FOOD` category (1,712 triplets), while the lowest count was observed for the `PRICE` category (255 triplets). Aspects were more frequently addressed explicitly (3,149 triplets) rather than implicitly (1,165 triplets). Positive sentiments were expressed towards the majority of identified aspects (2,339 triplets), followed by negative sentiments (1,795 triplets). A neutral sentiment was expressed towards 180 aspects.

Moreover, Table 3 presents the most frequently occurring aspect terms within each aspect category, and Table 4 shows the sample count for each triplet quantity. In the case of all aspect categories except for `GENERAL-IMPRESSION`, the most frequently occurring aspect term is equal to the name of the corresponding aspect category. Moreover, in more

[7] https://huggingface.co/deepset/gbert-large
[8] https://huggingface.co/deepset/gbert-base
[9] https://huggingface.co/t5-large
[10] https://huggingface.co/t5-base

	Positive		Negative		Neutral		Total	
Aspect Category	**Explicit**	**Implicit**	**Explicit**	**Implicit**	**Explicit**	**Implicit**	**Explicit**	**Implicit**
GENERAL-IMPRESSION	103	306	56	285	5	21	164	612
FOOD	880	83	532	98	109	10	1,521	191
SERVICE	514	69	316	177	10	0	840	246
AMBIENCE	312	26	99	42	6	0	417	68
PRICE	45	1	149	41	13	6	207	48
Total	1,854	485	1,152	643	143	37	3,149	1,165

Table 2: Aspect categories distribution per sentiment polarity and reference type for the annotated dataset.

Aspect Category	Description	Most Frequent Aspect Terms
GENERAL-IMPRESSION	Aspects related to the overall impression of the restaurant without focusing on the aforementioned aspect categories.	*Restaurant* (42) *RESTAURANT_NAME* (22) *LOC* (22) *Lokal* (12) *Brauhaus* (5)
FOOD	Aspects related to food in general or specific dishes and drinks.	*Essen* (302) *Bier* (46) *Speisen* (42) *Fleisch* (30) *Küche* (28)
SERVICE	Aspects related to service in general or the attitude and professionalism of staff, wait times, or service offerings such as takeout.	*Service* (209) *Bedienung* (125) *Personal* (90) *Kellner* (58) *Kellnerin* (17)
AMBIENCE	Aspects related to the ambiance and atmosphere in general or the environment of the restaurant's interior and exterior, including its decor and entertainment options.	*Ambiente* (103) *Atmosphäre* (51) *Lage* (13) *Lokal* (12) *Location* (10)
PRICE	Aspects related to price in general or the pricing of dishes, beverages, or other services offered by the restaurant.	*Preise* (30) *Preis* (25) *Essen* (14) *Preisen* (11) *Preis-Leistungsverhältnis* (10)

Table 3: Description of the aspect categories and their most frequent aspect terms.

# Triplets	1	2	3	4	5	6	7	8	9	16
# Sentences	2,236	590	168	57	14	7	3	1	1	1

Table 4: Sample count of each triplet quantity.

than two-thirds (2,236) of the 3,078 sentences, exactly one aspect or triplet was identified.

4.2 Comparison with the SemEval Datasets

As the dataset used in this work and the datasets from SemEval 2015 and 2016 are similar in terms of their domain and the type and depth of annotation, it is possible to compare dataset properties, such as their class distribution or language-specific features, such as the ratio of explicitly and implicitly expressed aspects. In order to ensure the comparability of the annotations of the GERestaurant dataset with the two SemEval datasets from 2015 and 2016, various adjustments had to be made, as although the datasets have undergone similar annotation procedures, the labels of the aspect categories are named and summarized differently: (1) The PRICES subcategories of the SemEval datasets were transformed to the PRICE aspect category; (2) the RESTAURANT category of the SemEval datasets was converted to the GENERAL-IMPRESSION category; (3) the LOCATION category of the SemEval datasets were integrated into the AMBIENCE category; and (4) The DRINKS category of the SemEval datasets was merged into the FOOD category.

Table 5 depicts the class balances of the five aspect categories as well as the polarity labels over the three datasets GERestaurant, SemEval 2015 and 2016. Subsequently, we consider a dataset as the combination of its train and test sets. The balance of the aspect classes of the SemEval datasets is almost identical, facilitated in part by the fact that almost the entire SemEval 2015 dataset, with

Dataset	Aspect Category					Polarity			Aspect Term Type	
	General Impression	Food	Service	Ambience	Price	Positive	Negative	Neutral	Implicit	Explicit
GERestaurant	18.0%	39.7%	25.2%	11.2%	5.9%	54.2%	41.6%	4.2%	27.0%	73.0%
SemEval 2015	20.6%	42.6%	17.7%	11.5%	7.5%	66.1%	30.0%	3.9%	24.9%	75.1%
SemEval 2016	20.6%	43.6%	17.9%	10.8%	7.1%	67.4%	28.3%	4.3%	24.8%	75.2%

Table 5: Comparison of the balances of the aspect category, the polarity labels and the ratio of implicitly and explicitly expressed aspect terms between the three ABSA datasets GERestaurant, SemEval 2015 and SemEval 2016.

1,700 of its 1,702 annotated examples, has been integrated into the SemEval 2016 dataset, which contains a total of 2,384 annotated examples. The overall class distribution of the GERestaurant dataset is also quite similar to that of the SemEval datasets and differs primarily in a 6.5 percentage point higher occurrence of the SERVICE aspect category, while all other aspect classes occur slightly less frequently. Considering the distributions of the polarity classes across all aspects, while the overall distributions of the polarity labels between the SemEval datasets are again very similar, bigger differences can be observed between the GERestaurant and SemEval datasets. The proportion of the neutral label remains comparably low between all datasets, but the negative polarity label was assigned up to 12 percentage points more frequently in the GERestaurant dataset at 41.6%, while the positive label was correspondingly annotated less frequently compared to the SemEval datasets, constituting only 54.5% of the total. Similar to the distribution of aspect classes, the ratio of implicitly and explicitly expressed aspects is very similar between all corpora. While the two SemEval datasets have an almost identical ratio, the GERestaurant dataset is only slightly above in terms of implicit aspects with an increase of about two percentage points, resulting in 27.0% implicitly expressed aspects and 73.0% explicitly expressed aspects.

4.3 Baseline Performance

The performance achieved in the four ABSA tasks under consideration are presented in Table 6. For predicting the five aspect classes (ACD task), gbert-large demonstrated the highest performance, achieving micro and macro F1 scores of 91.82 and 90.73, respectively, placing it approximately three percentage points ahead of gbert-base. Similarly, in the classification of aspects combined with their polarity (ACSA), the best performance was observed when employing gbert-large, which attained micro and macro F1 scores of 85.14 and 58.61, respectively. The micro-averaged F1 score surpassed that achieved with gbert-base by approximately 11 percentage points, while in the case of the macro-averaged F1 score, it exceeded it by around 22 percentage points.

Task	Language Model	F1 Micro	F1 Macro
ACD	gbert-large	91.82	90.73
	gbert-base	88.76	87.82
ACSA	gbert-large	85.14	58.61
	gbert-base	73.85	36.63
E2E-ABSA	gbert-large	81.61	77.28
	gbert-base	74.66	50.25
TASD	t5-large	68.86	59.03
	t5-base	64.74	54.32

Table 6: Performance for the baseline models per ABSA subtask.

For the E2E-ABSA task, gbert-large demonstrated the highest performance as well, achieving a micro F1 score of 81.61 and a macro F1 score of 77.28. This performance improvement over gbert-base, with a micro F1 score of 74.66 and a macro F1 score of 50.25.

Similarly to the previous tasks, again, the large model variant exceeded the performance of the base model by about four to five percentage points, resulting in a micro F1 score of 68.86 a macro F1 score of 59.03 for the t5-large model.

5 Limitations

While GERestaurant provides a valuable resource for studying ABSA in the German restaurant domain, it also comes with limitations. Firstly, the annotations are based on human judgments, which introduces subjectivity and potential inconsistencies. Furthermore, the quality of annotations is constrained by the fact that each example was not independently annotated by multiple annotators, but rather, one annotator annotated all sentences and their annotations were reviewed by another annotator.

Furthermore, the imbalance among the five aspect categories can be considered a limitation of this work. For instance, the fewest number of aspects (251) are assigned to the PRICE category, while the majority of aspects (1,676) are assigned to the FOOD category. Similar imbalances are observed in terms of sentiment polarities, with only 175 aspects toward which a neutral sentiment was expressed, compared to 2,283 aspects towards which a positive sentiment was expressed, which represents more than half of all aspects.

6 Discussion

GERestaurant offers a novel resource for ABSA research in the German language, specifically within the restaurant domain. Comprising 3,078 manually annotated sentences, GERestaurant encompasses both implicit and explicit aspects, annotated by human annotators. This is the third German language dataset besides GermEval 2017 (Wojatzki et al., 2017) and MobASA (Gabryszak and Thomas, 2022) to include annotations of aspect terms, aspect categories, and sentiment polarities of both implicit and explicit aspects.

The analysis of the class distributions of the aspect classes and the sentiment polarities between the German GERestaurant dataset and the English SemEval 2015 and 2016 datasets revealed a strong similarity of the ABSA-specific annotations of the datasets. The close correlation between the datasets opens up a variety of possibilities to compare the performance of ABSA methods on English and German datasets and could provide conclusions on how far methods can be used across languages despite language-specific differences in the datasets and methods.

Our provided baseline performance on all four ABSA tasks is in line with the performance reported in similar studies using transformer-based models for such tasks across various domains. However, it's important to acknowledge that the comparability of the results is limited due to variations in the number of aspect categories and the number of training examples across the datasets.

A micro-averaged F1 score of 91.82 was achieved in the ACD task, consistent with micro-averaged F1 scores obtained on other datasets, e.g. a micro-averaged F1 score of 90.89 on the restaurant dataset of SemEval from 2014 (Sun et al., 2019) or a micro-averaged F1 score of 90.6 on the dataset comprising hotel reviews presented by Fehle et al. (2023).

In the ACSA task, a micro-averaged F1 score of 85.14 was obtained, slightly exceeding the reported scores achieved on other datasets. Cai et al. (2020) reported micro-averaged F1 scores of 64.67 and 74.55 for the restaurant datasets of SemEval 2015 and 2016, respectively. Aßenmacher et al. (2021) reported a micro-averaged F1 score of 65.5 on GermEval 2017 and Fehle et al. (2023) reported a micro-averaged F1 score of 80.9 on the dataset comprising hotel reviews.

For the E2E-ABSA task, a micro-averaged F1 score of 81.61 was attained. Lower scores were reported for other domains, e.g. Li et al. (2019) reported a micro-averaged F1 score of 73.22 when considering the restaurant domain and 60.43 when considering the laptop domain, using datasets composed of examples from the SemEval datasets from 2014 to 2016.

The performance in the TASD task (micro-averaged F1 score of 68.86) falls within the spectrum of results observed by Zhang et al. (2021), who represented triplets as phrases, reporting a micro-averaged F1 score of 63.06 for the restaurant dataset of SemEval 2015 and a micro-averaged F1 score of 71.97 for the restaurant dataset of SemEval 2016.

7 Conclusion & Future Work

This work presents GERestaurant, a novel German language dataset comprising 3,078 restaurant reviews annotated for ABSA. The dataset covers implicit and explicit aspects, providing annotations for aspect terms, aspect categories, and sentiment polarities. Transformer-based SOTA models were fine-tuned on the training set provided by us for four common ABSA tasks, and subsequently evaluated on the test set.

In future work, GERestaurant could be utilized for developing improved machine learning models with focus on the German language for various ABSA tasks, building upon the methods introduced in this work and further improving the presented baseline values. Moreover, future work may involve expanding the aspect categories by incorporating fine-grained attributes, as in the SemEval datasets from 2015 and 2016, or including information about not only aspect phrases but also opinion phrases, in order to reflect the entire quadruple of an aspect-based annotation (Pontiki et al., 2015, 2016).

8 Ethical Considerations

The collection of our dataset adhered to strict privacy guidelines to safeguard the rights of users. Our primary objective was to extract reviews while avoiding the collection of personalized data that could potentially identify individual users or specific user groups. Furthermore, any direct references to individuals or restaurants were systematically anonymized to prevent indirect identification of individuals or establishments.

The dataset and its annotations are available upon request from the authors to ensure responsible usage for academic purposes, thus preserving the original intent of data collection. The Python code for data collection and data cleaning is accessible via GitHub[11].

Despite our meticulous data collection and anonymization procedures, inherent limitations and ethical considerations persist. Our dataset may not fully represent the spectrum of user sentiment due to potential bias in review writing, as reviewers may only represent a specific subset of the population. Furthermore, the transferability of knowledge about review semantics and characteristics across different rating platforms cannot be guaranteed.

References

Matthias Aßenmacher, Alessandra Corvonato, and Christian Heumann. 2021. Re-evaluating germeval17 using german pre-trained language models. In *Proceedings of the Swiss Text Analytics Conference 2021, Winterthur, Switzerland, June 14-16, 2021 (held online due to COVID19 pandemic)*, volume 2957 of *CEUR Workshop Proceedings*.

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th international conference on computational linguistics*, pages 833–843.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

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

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6.

Nils Constantin Hellwig, Markus Bink, Thomas Schmidt, Jakob Fehle, and Christian Wolff. 2023. Transformer-based analysis of sentiment towards german political parties on twitter during the 2021 election year. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 84–98.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6280–6285.

Hsin-Hsien Lee and Wei-Guang Teng. 2007. Incorporating multi-criteria ratings in recommendation systems. In *2007 IEEE International Conference on Information Reuse and Integration*, pages 273–278. IEEE.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 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)*, pages 34–41.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *Preprint*, arXiv:cs/0205028.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.

[11] https://github.com/NilsHellwig/GERestaurant

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

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

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 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*, pages 1546–1556.

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

Thomas Schmidt, Jakob Fehle, Maximilian Weissenbacher, Jonathan Richter, Philipp Gottschalk, and Christian Wolff. 2022. Sentiment analysis on twitter for the major german parties during the 2021 german federal election. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 74–87.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL-HLT*, pages 380–385.

Abinash Tripathy, Abhishek Anand, and Santanu Kumar Rath. 2017. Document-level sentiment classification using hybrid machine learning approach. *Knowledge and Information Systems*, 53:805–831.

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

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

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

A Appendix

A.1 Examples from the Annotated Dataset

Aspect Category	Triplets	Sentence
GENERAL-IMPRESSION	[('restaurant', 'GENERAL-IMPRESSION', 'POSITIVE')]	*"Very nice restaurant."*
FOOD	[('sausage', 'FOOD', 'POSITIVE')]	*"The sausage was incredibly delicious and perfectly seasoned."*
SERVICE	[('Service', 'SERVICE', 'NEGATIVE')]	*"Service unfortunately not attentive."*
AMBIENCE	[('NULL', 'AMBIENCE', 'NEGATIVE')]	*"It was much too loud, like in a club."*
PRICE	[('NULL', 'PRICE', 'NEUTRAL')]	*"Price-wise it was ok."*
PRICE, SERVICE	[('Prices', 'PRICE', 'NEUTRAL') ('service', 'SERVICE', 'NEGATIVE')]	*"Prices are ok and service as well."*
FOOD, AMBIENCE, SERVICE	[('food', 'FOOD', 'POSITIVE'), ('atmosphere', 'AMBIENCE', 'POSITIVE'), ('service', 'SERVICE', 'POSITIVE')]	*"Great food, great atmosphere and really nice and attentive service!"*

Table 7: Annotated examples for all aspect categories (English translation).

A.2 Paraphrase Generation Framework

A.2.1 Explicit Aspect

Sentence (Input)	Die Pasta war super, aber die Bedienung war unfreundlich!
Label	[('Pasta', 'FOOD', 'POSITIVE'), ('Bedienung', 'SERVICE', 'NEGATIVE')]
Paraphrased Label	Essen ist gut, weil Pasta gut ist [SSEP] Service ist schlecht, weil Bedienung schlecht ist [SSEP]

Table 8: Paraphrasing of an explicit aspect's label.

A.2.2 Implicit Aspect

Sentence (Input)	Es hat richtig gut geschmeckt!
Label	[('NULL', 'FOOD', 'POSITIVE')]
Paraphrased Label	Essen ist gut, weil es gut ist [SSEP]

Table 9: Paraphrasing of an implicit aspect's label.