Aspect-Based Sentiment Analysis as a Multi-Label Classification Task on the Domain of German Hotel Reviews

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

Leonie Münster
Media Informatics Group
University of Regensburg
Regensburg, Germany
leonie.muenster@stud.uni-regensburg.de

Thomas Schmidt
Media Informatics Group
University of Regensburg
Regensburg, Germany
thomas.schmidt@ur.de

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

Abstract

Aspect-Based Sentiment Analysis (ABSA) plays a crucial role in understanding fine-grained customer feedback, particularly in domains like hospitality where specific aspects of service often influence overall satisfaction. However, non-English languages such as German face a scarcity of readily available corpora and evaluated methods for ABSA, making it a challenging problem. This paper addresses this gap by utilizing BERT-based transformer models, known for their exceptional performance in context-sensitive natural language processing tasks, to perform ABSA in a multi-label classification setting. We demonstrate our approach on a novel dataset of German hotel reviews that we have collected and annotated from *TripAdvisor*, thus contributing a new resource to the field and proving the effectiveness of our methodology. With achieving a micro f1-score of up to 0.91 for aspect category classification and 0.81 for end-to-end ABSA, our approach aligns with the performance of similar methods on other German-language datasets and surpasses performance achieved on English-language datasets in the hotel domain.

1 Introduction

Sentiment analysis deals with the classification of attitudes, opinions, and sentiments and typically focuses on the three classes positive, neutral, and negative. The ever-increasing integration and presence of social media and the internet in everyday life is generating a huge amount of user-generated data that favors the use of sentiment analysis. As a result, it is nowadays used in various fields and domains, such as the analysis of political discourse (Xia et al., 2021; Schmidt et al., 2022), digital humanities (Schmidt and Burghardt, 2018; Schmidt et al., 2020), healthcare natural language processing (Moßburger et al., 2020), in improving products and services (Xu et al., 2019), and in the financial sector to predict stock market movement (Sousa et al., 2019). In recent years, sentiment analysis has also expanded its application areas to non-text-based media such as images and videos, e.g., in human-computer interaction (Halbhuber et al., 2019; Ortloff et al., 2019) or film analysis (Schmidt et al., 2021c; El-Keilany et al., 2022).

Assigning a positive or negative label to entirely positive or negative texts is usually straightforward. However, analyzing texts that contain a mixture of different sentiments in a single sentence or text quickly becomes a challenge. This is particularly the case when it is not about general trends or developments but about precise statements concerning different aspects or characteristics of products or services, where a rough estimate of sentiment is insufficient. For over a decade, Aspect Based Sentiment Analysis (ABSA) has gained popularity to solve this problem, whereby instead of determining an overall sentiment for a sentence or a document, the sentiment is determined in relation to individual aspects or entities occurring in the text, such as the battery life of a smartphone or the friendliness of a service employee (Liu et al., 2005).

As in other research fields of natural language processing (NLP), there is a clear imbalance in sentiment analysis in terms of available resources and evaluated techniques when looking at differ-

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 202–218
September 18–22, 2023. ©2023 Association for Computational Linguistics

ent languages and domains. While research has progressed a lot during recent years in the English language domain, the field of ABSA in German is still relatively unexplored. To our knowledge, only a small number of ready-to-use corpora exist and only a few methods have been evaluated (Fehle et al., 2021; Chebolu et al., 2022). Moreover, corpora that are needed for the training of aspect-based machine learning approaches or for the evaluation of ABSA methods are not compatible with corpora that can be used as resources for general sentiment analysis approaches that determine sentiment only at the document or sentence level. Since annotation of training data for ABSA usually involves working at the phrase or word level to establish complex relationships between phrases describing the aspect and phrases containing the sentiment, the annotation process is often highly time-consuming and difficult. To counteract this, there are approaches that handle datasets that have not been annotated manually or only in a less complex way (Chang et al., 2019; Kastrati et al., 2020). One promising example is the definition of ABSA as a multi-label classification problem (Tao and Fang, 2020; Jin et al., 2020). In this case, the classifiers are trained with texts annotated with aspects and polarities, albeit at the sentence level rather than the phrase level, thus decreasing complexity. The annotation contains information about the aspect occurring in the text as well as its assigned sentiment, but no information about where the aspect occurs or by which exact phrase it is composed. This approach has already achieved good results in the German language (Aßenmacher et al., 2021). Building on prior research, this work explores the potential of applying the multi-label classification method for ABSA to a different domain. Given the promising results this approach has yielded in the realm of customer reviews in context of public transportation (Aßenmacher et al., 2021), this work determines its effectiveness and the expected classification results when applied to other areas and domains for which ABSA is a relevant tool for extracting fine-grained opinions from user-generated content. For this purpose, a new corpus was created on a domain that is widely discussed in the English language (Akhtar et al., 2017; Abro et al., 2020), but to our knowledge has not yet been addressed in the German language: Online reviews of hotels and their services.

The contributions of this paper are as follows: (1) the creation of a dataset for ABSA in the domain of hotel reviews in the German language, (2) an evaluation of multiple pre-trained transformer-based models for ABSA as a multi-label classification task on hotel reviews in German and (3) a discussion about the performance of transformer-based models for ABSA at different tasks and various levels of annotation complexity.

2 Related Work

Over the last decade, ABSA has experienced significant growth through different shared task workshops, such as the SemEval Shared Tasks for the English language from 2014 to 2016 (Pontiki et al., 2014, 2015, 2016), stimulating the development of various methods addressing the three fundamental subtasks in aspect-based sentiment analysis: aspect term extraction, aspect category classification, and aspect sentiment classification. These tasks utilized datasets compiled from two domains: restaurant and laptop reviews. With each iteration of the SemEval Shared Task, the size of the dataset and the complexity of the annotations increased. Initially, only the specific aspect word, its aspect category, and the corresponding polarity were annotated. Later, however, the aspects were divided into entities/main aspect categories and attributes/sub-aspect categories (these terms are often used interchangeably), thus increasing the complexity of the datasets due to a large number of possible combinations between main and subcategories. Even after these workshops, the datasets continue to be used as a benchmark resource for the evaluation of newfound ABSA approaches (Brauwers and Frasincar, 2022; Nazir et al., 2020).

These datasets are far from being the only ones available in the English language. In particular, since the first SemEval workshop on ABSA in 2014, the number of accessible datasets for the English language has significantly increased, covering various domains with different levels of annotation complexity, such as hotel reviews (Yin et al., 2017), financial microblogs (Maia et al., 2018), and Amazon product reviews (Liu et al., 2015).

Approaches to determining aspect-based sentiment are diverse and have evolved over time. While earlier methods primarily relied on rules, word frequencies, or lexicon-based techniques and tackled only sub-tasks to the ABSA problem, contemporary approaches emphasize neural networks and

deep learning and try to solve ABSA as a one-in-all/end-to-end solution (Chen et al., 2022; Yan et al., 2021). Since their introduction in 2017, pre-trained transformer models have, together with deep learning neural networks, been recognized as state-of-the-art in the field. Nowadays, approaches achieve accuracy and f1-scores of over 80 % for subtasks of ABSA or complete ABSA solutions on various corpora. Notably, some transformer-based architectures attain scores exceeding 90 % on specific datasets (Brauwers and Frasincar, 2022; Do et al., 2019).

In the hotel domain, methods usually only deal with subtasks of ABSA. For that, star ratings of the review or individual aspects on rating portals (e.g. *Tripadvisor*) are often used to derive the polarity of individual reviews and aspects and to gain a ground truth dataset. Chang et al. (2019) builds on this method and classifies the individual aspect categories by using support vector machines and convolution tree kernels with good success on eight classes (macro f1-score: 0.80). Tran et al. (2019) uses the combination of a BiLSTM-CRF model for the extraction of aspect phrases and their polarity and LDA topic modeling for aspect category classification to capture the aspect category and the associated sentiment from hotel reviews and achieves a micro f1-score of 0.873 for the extraction of aspect phrases and their polarity and an accuracy of 0.800 for the determination of the associated aspect category. Qiang et al. (2020) tackled the aggregation of aspect-sentiment information and used a Multi-Attention-Network BiLSTM to capture the fine-grained statements regarding individual aspects in hotel reviews in order to infer the overall sentiment of a review and achieved a micro f1-score of 0.798 on a custom generated dataset.

In German, the largest available dataset was published as part of the GermEval Shared Task Workshop in 2017, which contains more than 26,000 annotations consisting of entity-attribute-polarity tuples related to the German transportation service provider *Deutsche Bahn* (Wojatzki et al., 2017). However, the dataset was evaluated only based on main aspects and their corresponding polarities, with the category of attributes being omitted.

Other datasets in German language include the SCARE corpus, consisting of 1,760 Google Play Store reviews with 2,487 aspect-polarity annotations (Sänger et al., 2016); the USAGE corpus, comprising 611 Amazon reviews with more than 5,000 aspect-polarity annotations (Klinger and Cimiano, 2014); the PotTS dataset, made of 7,992 Twitter messages on political topics with annotations for sentiment targets and their sentiment phrases (Sidarenka, 2016); a corpus in the domain of German historical plays consisting of around 6,500 sentiment/emotion and 12,000 source and target annotations (Schmidt et al., 2021a,b); and the TDDL corpus, consisting of 4,521 tweets about the "Tage der deutschsprachigen Literatur" (Engl.: "Days of German Literature") with 8,264 main aspect-attribute-polarity annotations (De Greve et al., 2021). As with the English-language datasets, these German datasets also vary in quality and have been annotated using different levels of complexity and granularity in their annotation schemes.

ABSA approaches have been evaluated on German-language datasets only to a limited extent. While earlier approaches were mainly based on classical machine learning like conditional random fields or neural networks with pre-trained word-embeddings, more recent methods focus on recent advances in NLP like deep learning and pre-trained transformer architectures (Sänger et al., 2016; Schmitt et al., 2018; Akhtar et al., 2019). Aßenmacher et al. (2021) were able to significantly improve the performance for classifying aspects and their polarities on the GermEval dataset. They achieved this by treating ABSA as a multi-label classification problem and employed a BERT-transformer model instead of the CNN+FastText model used by Schmitt et al. (2018). This led to a significant improvement of the model's accuracy with a rise of micro-averaged f1-scores from 0.54 and 0.44 to 0.78 and 0.67 for aspect and aspect-polarity classification respectively. De Greve et al. (2021) also addressed the subtasks of aspect-term classification and aspect-sentiment classification using a BERT architecture. They achieved macro and weighted F1 scores of 0.69 and 0.83 for the classification of the six main aspects on the TDDL dataset, as well as macro and weighted f1-scores of 0.54 and 0.73 for the classification of all 48 combinations of main and sub-aspects while using the gold annotations of the aspect terms as input. The authors were also able to achieve a macro f1-score of 0.72 for aspect-polarity classification by implementing a context window of five words before and after the aspect phrase.

3 Methods

3.1 Creation of a Dataset of German Hotel Reviews

3.1.1 Dataset Generation

The foundation of the dataset are 1,512 user reviews about a selection of hotels in the city of Regensburg (situated in the south of Germany) in German language. The reviews were acquired with the web scraping application Parsehub[1] from the site *TripAdvisor*.[2] The selection process focused on five mid-class hotels, chosen specifically for their substantial number of user reviews and diverse proximity to the city center. In this way, we were able to capture a range of perspectives related to the location of the hotels. Furthermore, attention was paid to ensure that the selected hotels had comparable features (e.g. restaurants and parking) to facilitate consistent topics across the user reviews.

In order to annotate the dataset with aspects and polarities contained at the sentence level, the 1,512 user reviews were split into sentences with the online sentence splitter tool TextConverter.[3] Subsequently, we manually inspected the splits and made any necessary corrections, in case the user's statement was otherwise no longer comprehensible.[4] This results in a dataset of 21,182 sentences. For the annotation process, the dataset was divided into chunks of 200 units and randomly distributed to the participants. This resulted in a subset of 5,000 sentences, with each sentence annotated by two different annotators as part of the annotation study.

3.1.2 Data Annotation

The goal of the study was the annotation of three-part tuples consisting of an aspect, an attribute or sub-aspect (a specific facet of an aspect), and the associated polarity, following the approach of previous work (Pontiki et al., 2015, 2016; Wojatzki et al., 2017). The aspect (e.g. hotel) and the attribute (e.g. price) are combined to form the aspect category pair. For the determination of the aspect categories of our dataset, the four predefined rating categories of each *TripAdvisor* review - location (Ger.: Lage), price (Ger.: Preis), cleanliness (Ger.: Sauberkeit), and service (Ger.: Service) - were taken into account, as we assumed that, at least to some extent, these categories were used by the users as reference for their written reviews. Furthermore, we also took into account findings from related work in the same domain, in which additional aspects and attributes such as ambience (ger. Ambiente), restaurant (Ger.: Restaurant), rooms (Ger.: Zimmer), general (Ger.: Haupt) and quality (Ger.: Qualität) were used (Abro et al., 2020; Chang et al., 2019). On the basis of this information, the five aspects hotel (Ger.: Hotel), food & drink (Ger.: Essen & Trinken), location, service, and rooms were selected for annotation. General, price, quietness (Ger.: Ruhe), cleanliness, and style (Ger.: Style) were selected as attributes, which could be annotated in different combinations with the main aspects. The annotation of the polarity of the aspects was carried out using the three classes positive, neutral, and negative. All possible annotations of aspects and attributes can be seen in Table 1, furthermore all possible combinations of aspect categories are depicted in Table 5 in the appendix. It was possible to annotate one or more tuples of entities, attributes, and polarities per sentence. If no aspect could be identified in the sentence, it was also possible to skip the sentence and omit it from the annotation.

Category	Possible Class Labels
Aspect	Hotel, Location, Food & Drinks, Service, Rooms
Attribute	General, Price, Quietness, Cleanliness, Style
Polarity	Positive, Neutral, Negative

Table 1: All possible class labels of the annotation.

The annotation was carried out in the web tool INCEpTION (Klie et al., 2018) which is a more advanced version of its predecessor WebAnno (Yimam et al., 2014). All study participants received detailed annotation guidelines with an explanation of the background of the study, an introduction to the topic, a list of all possible combinations of aspects and attributes with example annotations, and an introduction on how to operate the annotation tool INCEpTION. The selection of the different aspects, attributes, and polarities was predetermined by the annotation tool in order to prevent incorrect annotations. The annotation study was carried

[1] https://www.parsehub.com/

[2] https://www.tripadvisor.de/

[3] https://textconverter.com/

[4] In rare cases, the manual correction resulted in a sample comprising up to two sentences. However, for ease of understanding, we refer to one sample as a sentence in the remainder of the text.

Aspect	Count	Percentage
Hotel	1,477	26.3 %
Rooms	1,457	25.9 %
Location	963	17.1 %
Service	907	16.2 %
Food & Drinks	813	14.5 %

Attribute	Count	Percentage
General	3,326	59.2 %
Style	1,201	21.4 %
Cleanliness	405	7.2 %
Price	396	7.1 %
Quietness	289	5.1 %

Polarity	Count	Percentage
Positive	4,032	71.8 %
Neutral	957	17.0 %
Negative	628	11.2 %

Table 2: Amount of samples per aspect, attribute, and polarity class, ordered by the respective portions.

out by 27 students, with each participant annotating a subset of either 200 or 400 sentences. Each sentence was annotated by two annotators. The agreement between the annotators is visualized in Table 6 in the appendix. Due to the possibility of assigning none, one or more aspects to a sentence, Krippendorff's α[5] is a suitable metric for agreement. The metric is calculated using the masi distance (Passonneau, 2006). The agreement can be examined at different levels of complexity: (1) an isolated view on the aspects, (2) the combination of either aspects and attributes or (3) aspects and polarities, and (4) all metrics together - the aspects, attributes as well as their polarities. If only the aspects are considered, the average agreement of the annotators is 0.61, if the attributes are included, the average value drops to 0.48 and if the whole tuple is considered, the average agreement goes down to 0.43. If the complexity of the attribute is removed from the tuple and only the aspect and its polarity are considered, the average agreement is 0.54. These agreement values are considered to be of moderate agreement (Hayes and Krippendorff, 2007; Landis and Koch, 1977).

Subsequently, to increase the quality of the dataset, all 5,000 sentences were manually curated. First, all the annotations were approved where both annotators assigned the same aspect tuple. If sentences were annotated by only one annotator, it was decided individually whether to accept or discard the annotation. For sentences with different annotations in terms of entity, attribute or polarity, it was individually decided which annotation should be classified as correct or not.

3.1.3 Dataset Characteristics

After curation, the dataset consists of 4,254 sentences (746 sentences did not contain clearly discernible aspects) and 5,617 annotations of aspect tuples (see Figure 2 in the appendix for an excerpt of the dataset). Table 2 contains the frequency distribution of the dataset at the level of the aspects, attributes, and polarities in an isolated view. The frequencies of the aspects are slightly unbalanced, with the most frequent aspect "hotel" occurring almost twice as often as the least frequent aspect "food & drinks". The distributions for the attributes, as well as the polarities, are strongly unbalanced. Thus, about 59 % of all attributes are assigned to the general class, while the three least represented attributes cleanliness, price, and quietness take up less than 20 % of the total amount. A similar picture emerges for the polarities. Thus, almost 72 % of all labels are assigned to the polarity positive, while the classes neutral and negative are only represented with around 17 % and 11 % respectively. The distributions for different combinations of aspects, attributes, and polarities in the data set are also strongly unbalanced (see Figure 1 and additionally Tables 7, 8 and 9 in the appendix). The frequency distributions for the combinations of multiple classes are depicted in Figure 1. For example, for the aspect-polarity combinations, 1/3 of the most frequently occurring combinations account for more than 2/3 of the total dataset; this value is significantly higher for the aspect-attribute combinations with about 83 % and is still topped off by the aspect-attribute-polarity combinations, where 1/3 of all combinations account for more than 87 % of all samples in the dataset.

3.2 Dataset Evaluation with Pre-Trained Transformer-Models

3.2.1 Data Preprocessing

In multi-label classification, one or more classes are assigned to each sample, which requires remodeling the dataset structure. For each class, each sample is given a binary truth value about whether the class is present in the sample or not, resulting in a one-hot-encoded sequence. The number of classes is determined by the level of annotation granularity. For instance, classifying only the aspects results in 5 classes, considering both aspects and attributes leads to 18 classes, and incorporating aspects, attributes, and polarity results in 54 classes. It is

[5] https://pypi.org/project/krippendorff/

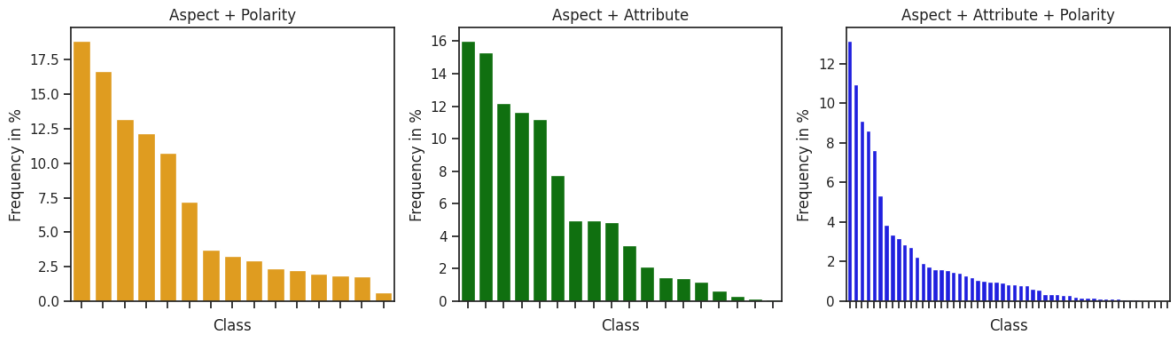

Figure 1: Frequency distributions for all combinations of classes.

important to note that out of these 54 classes, one class, namely "Food & Drinks#Quietness:Neutral" did not occur in the dataset and was therefore unintentionally omitted during the annotation process. However, through the conversion into a binary statement regarding the occurrence of a class, a maximum of one occurrence of the same class/same combination of classes can be included. Thus, the information of identical classes occurring several times in the same sequence is considered as one.

As an example, the class labels of the sentence "The service staff was very nice, but I think the location of the hotel is inconvenient." are depicted in Table 3. Here, the aspect "Location" was annotated as negative and the aspect "Service" was annotated as positive, resulting in the one-hot-encoded labeling sequence [0,0,1,0,0,0,0,0,0,1,0,0,0,0,0] which serves as input for our classifier.

3.2.2 Metrics

In a multi-label classification setting commonly used metrics are hamming loss, accuracy, precision, recall, and f1-score (Zhang and Zhou, 2013; Tsoumakas and Katakis, 2007).

Similar to the metrics used in SemEval 2014, 2015, and 2016, and GermEval 2017 Shared Tasks, we use a micro-averaged f1-score as the primary evaluation metric. However, since the balancing of the created dataset tends to be skewed depending on the level of detail of the annotation, we also provide a macro f1-score, averaged over the individual class f1-scores. Thus, this value also takes into account the prediction performance of the underrepresented classes.

3.2.3 Evaluation Procedure

We tested three different pre-trained BERT transformer models publicly available: (1) one of the largest transformer-based BERT language models for German, *gbert-large* by *Deepset* (Chan et al., 2020), and two of the best-performing BERT-based models in similar studies, (2) *bert-base-german-uncased* by DBMDZ[6] and (3) the comparatively lightweight model *distilbert-base-german-cased* (Sanh et al., 2019), pre-trained on the same dataset as (2).[7] All models were acquired via the *Hugging Face* platform and implemented using the Python libraries *Pytorch* (Paszke et al., 2019) and *Transformers* (Wolf et al., 2020). Evaluation metrics were calculated using *scikit-learn* (Pedregosa et al., 2011).

For increased validity, the dataset was cross-evaluated with stratified 4-split kfold, alternating 3 parts of the dataset for training and one part of the dataset for evaluation. Each model was evaluated based on four different tasks, split into two categories of subtasks of ABSA: (1) aspect category classification and (2) aspect sentiment classification. For each subtask, we evaluated on different

[6]Munic Digitalization Centre Digital Library team at the Bavarian State Library, see https://github.com/dbmdz.

[7]In text further referenced as *deepset-gbert-large*, *dbmdz-bert-base* and *distilbert-base*.

Aspect-Polarity-Combinations

Hotel			Location			Food & Drinks			Service			Rooms		
Pos	Neut	Neg	Pos	Neut	Neg	Pos	Neut	Neg	Pos	Neut	Neg	Pos	Neut	Neg
0	0	1	0	0	0	0	0	0	1	0	0	0	0	0

Table 3: Example labels of the input for the model of an aspect polarity classification. A '1' means that this class occurs in the text, a '0' the opposite.

sets of data, once with information about attributes and once without. Thus, both subtasks differ in complexity of the ground truth data used: classification of the aspect class, classification of the aspect class and its associated polarity, and both in combination with the attribute class. This resulted in classification tasks with 5 and 18 classes for task 1 and 15 and 53 classes for task 2.

Training was done using an AdamW-optimizer (Loshchilov and Hutter, 2017) and a binary cross entropy loss function with sigmoid activation, which is mandatory for multi-label classification. Since finding the right hyperparameters is a crucial component in every deep learning-based classification task, we performed systematic hyperparameter tuning for 20 trials per evaluation run with *Optuna* (Akiba et al., 2019) while trying to minimize the value of hamming loss with a *Tree-structured Parzen Estimator (TPE)*. The pre-selection of hyperparameters is based on Devlin et al. (2019) and own pre-experiments:

- Learning rate $\in [2e-5, 5e-5]$
- Batch size $\in \{8, 16, 32\}$
- Number of epochs $\in \{2, 3, 4\}$

Hyperparameter optimization showed that for 11 out of the 12 runs the best configuration comprised a batch size of 8 and 3 or 4 epochs. The only exception was the aspect class determination by *deepset-gbert-large*, which achieved the best result with a batch size of 32. It's worth noting that all models struggled significantly with classifying the aspect-attribute-polarity tuple when using a batch size of 32, frequently failing to predict any class. Regarding the learning rate, no clear trend is discernible, although often the best results were achieved with values just at the specified minimum or maximum, which indicates that the actual optimum of the parameter might lie outside the limits we had defined.

The training and evaluation were done on a workstation setup with an Intel Xeon W-2275 CPU, 128 GB of Ram, and 2x NVIDIA RTX A5000 GPUs.

4 Results

The evaluation results for all four subtasks are depicted in Table 4, divided in subtasks and models. In addition, we also included values obtained by Aßenmacher et al. (2021) which implemented multi-label classification with BERT on the GermEval 2017 dataset.[8]

4.1 Evaluation of Aspect Category Classification

The three BERT models for classifying aspects and aspects & attributes differ only slightly in terms of performance. In predicting the five aspect classes, *deepset-gbert-large* performs best with micro and macro f1-scores of 0.906 and 0.910, placing it about one percentage point ahead of both *dbmdz-bert-base* and *distilbert-base*. Further analysis showed that for the best performing model *deepset-gbert-large* the individual classes could be predicted almost equally well with an f1-score of approximately 0.92, the only outlier being the aspect "Hotel" with 0.86. Furthermore, when the attribute classes are included, *deepset-gbert-large* also performed best in the classification of the 18 aspect combinations, achieving micro and macro f1 scores of 0.797 and 0.542, but this time by a margin of between about 2 and 6 percentage points over the other models. Upon further analysis of the individual aspect-attribute class combinations, it's obvious that the prediction performance of all models correlates with the frequency of occurrence of the class

[8]The GermEval dataset was published along with two datasets for evaluation, each collected at different points in time. When referring to the results of the GermEval dataset throughout this paper, we report the average of both eval datasets.

Language Model	Aspect		Aspect + Attribute		Aspect + Polarity		Aspect + Attribute + Polarity	
	F1 Micro	F1 Macro	F1 Micro	F1 Macro	F1 Micro	F1 Macro	F1 Micro	F1 Macro
deepset-gbert-large	**0.906**	**0.910**	**0.797**	**0.542**	**0.809**	**0.659**	**0.651**	**0.173**
dbmdz-bert-base-german-uncased	0.891	0.895	0.774	0.504	0.779	0.599	0.592	0.119
distilbert-base-german-cased	0.880	0.886	0.744	0.432	0.741	0.490	0.561	0.107
Multi-label BERT on GermEval2017 (Aßenmacher et al., 2021)	0.776		0.776		0.672		0.672	

Table 4: Results for the 4 subtasks of the evaluation. Best values are depicted in bold.

samples. In terms of the *deepset-gbert-large* model, this means that the four least frequently occurring classes are not detected by the model, while the four most frequently occurring classes are among the top 5 predicted classes in terms of classification results.

4.2 Evaluation of Aspect Sentiment Classification

The classification of aspects in combination with polarity gives a similar picture as in chapter 4.1. Again, *deepset-gbert-large* achieves the best results both with and without consideration of the attribute class. Thus, *deepset-gbert-large* achieves micro f1- and macro f1-scores of 0.809 and 0.659 for the classification of the 15 classes from aspect & polarity. The model obtains relatively good classification results for most of the 15 individual classes, up to an f1-score of 0.931. However, once more, the performance drops off with the decrease in frequency of the class in the dataset, whereby the rarest combination "Service - Neutral" with only 36 occurrences cannot be predicted at all. Aspects related to positive polarity labels are recognized best, followed by negative and eventually neutral polarity labels.

Taking the attribute category into account, thus predicting the whole aspect-attribute-polarity tuple, *deepset-gbert-large* achieves a micro f1-score of 0.651 and is, therefore, at least five percentage points ahead of the other models. Since *deepset-gbert-large* can only make a correct prediction for 17 of the total 53 classes, the macro f1-score drops significantly, down to 0.173. The model almost completely fails to recognize combinations with the neutral polarity class, while aspects & attributes in combination with the positive polarity class work best.

5 Discussion

5.1 Aspect Category and Aspect Sentiment Classification

In this work, we investigated the adaptation of ABSA as a multi-label classification for the domain of hotel reviews and compared its performance in the context of previous methods. However, comparing values between corpora and approaches should be done with caution, given the considerable disparities in the origin, quality, depth, and size of the datasets that most approaches rely on. Based on the fact that *deepset-gbert-large* was pre-trained on ten times the amount of raw data and at the same time has more than three times as many parameters and more than twice as many layers as the other two models, it is plausible that this model also achieves the best classification results. Nevertheless, the results in some categories (e. g. aspect classification) are sufficiently close to each other that it can be considered that the significantly smaller model size and the much faster fine-tuning phase could outweigh the disadvantages in classification accuracy (see Table 10 for model parameters and Table 11 for training times).

With regard to the subtask of aspect category classification, the best transformer model we evaluated, *deepset-gbert-large*, achieves micro and macro f1-scores of 0.906 and 0.910 for the classification of the 5 aspect classes, outperforming values achieved in the domain of English hotel reviews, such as Andono et al. (2022) with a micro f1-score of 0.89 on 5 aspects, Chang et al. (2019) with a macro f1-score of 0.80 on 8 aspect categories or Afzaal et al. (2019) with 0.85 on an unknown number of aspects, and in the domain of social media comments about German literary prize winners with a macro f1-score 0.79 on 7 aspects (De Greve et al., 2021).

In terms of the end-to-end approach which combines aspect category classification and aspect sentiment classification, all tested BERT models delivered convincing results. Among them, the highest f1-scores were obtained by *deepset-gbert-large* with micro- and macro-averages of 0.809 and 0.651 on aspects and their polarities. Our results surpass those achieved in comparable settings, such as the results reported by Tran et al. (2019) and Afzaal et al. (2019) on the domain of hotel reviews in the English language. Notably, the approach by Tran et al. (2019) achieved an f1-score of 0.873 for aspect term extraction and binary polarity classification, as well as an accuracy of 0.80 for aspect category classification, while Afzaal et al. (2019) managed to achieve f1-scores of 0.85 and 0.91 for aspect category and aspect sentiment classification, respectively. However, two key considerations highlight the differences between their works and ours: (1) Their approach relied exclusively on binary polarity labels, which inherently simplified the sentiment analysis process compared to our approach and (2) they concatenated both subtasks, which could potentially compound error propagation throughout their pipeline and, thus, lower the overall classification performance. In

contrast, our approach produced superior results while combining both subtasks, likely due to the individual strengths of transfer-learning and our chosen BERT models.

However, it must be noted that our results have shown that the classification performance can decrease significantly when additional aspect classes are added, which is in line with results obtained in current research (Aßenmacher et al., 2021). Therefore the number of classified aspects can be decisive for a comparison between different methods and datasets.

Additionally, our results for the aspect classification subtask on 18 aspect categories (micro f1-score: 0.797) are slightly better than the results achieved by Aßenmacher et al. (2021) on 20 aspect categories (micro f1-score: 0.776), which followed the same approach as we did, a (BERT-based) multi-label classification, but on a German dataset of user ratings (GermEval 2017). If polarity is taken into account for the end-to-end overall ABSA solution, here again, *deepset-gbert-large* achieves comparable classification results with a micro f1-score of 0.651 on 53 classes (aspect-attribute-polarity combinations) to Aßenmacher et al. (2021) on the GermEval corpus with an f1-score of 0.672 on 60 classes by their best-performing model *dbmdz-bert-base*. Although the classification results for the aspect-polarity classification case are slightly worse than the results obtained on GermEval 2017 by Aßenmacher et al. (2021), *deepset-gbert-large* performs better than *dbmdz-bert-base* in the direct comparison on the domain of hotel reviews, suggesting that the performance difference may not be due to the model itself, but to the underlying language-specific differences of the domain or the dataset. Nevertheless, it can be observed with both approaches on both domains that the classification of strongly underrepresented classes is significantly worse than that of frequently occurring classes. This suggests that this is not a domain-specific problem, but could be due to the implementation of our approach or the underlying datasets, which needs to be taken into account when developing future multi-label classification approaches.

In summarization, our results allow the conclusion that (BERT-based) multi-label classification is a valid method for aspect classification and end-to-end ABSA on domains other than user ratings on social media, and should be extended to other domains as it is already the case for the English language (Kumar et al., 2019).

5.2 Limitations & Ethical Considerations

As the selection of the right dataset is an essential component for any classification task, the quality of its (manual) annotations may also reflect on the classification results of machine learning approaches. The agreement of the participants regarding the annotation of the dataset of this work indicates a low to moderate agreement. Considering the fact that a large number of combinations of different classes can be annotated in ABSA, this is usually presented as an acceptable result (Moreno-Ortiz et al., 2019), even though it is reasonable that a lower level of agreement and thus a debatable lower quality of the dataset is likely to affect the classification performance of the methods applied to it (Mozetič et al., 2016). Since Krippendorff's α varies considerably between individual annotator-pairings (see Table 6 in the appendix), it is possible that demographic characteristics, such as previous experience with annotation studies or the subject of the sentiment analysis, could have an influence on the quality of the annotations. However, the imbalance of the annotated classes does not seem to be a rare phenomenon and often occurs in context of ABSA in connection with user reviews in general or reviews from the hotel domain or *Tripadvisor* in particular (Risch et al., 2021; Tran et al., 2019; Pontiki et al., 2015).

The process of gathering our dataset followed strict privacy guidelines to protect the rights of users. The primary aim was to extract reviews or texts, while carefully avoiding the collection of personalized data that could potentially identify individual users or specific user groups. By doing so, we aimed to mitigate the risk of drawing unwarranted or ethically questionable conclusions from our analyses. Additionally, any direct references to individuals or hotels were systematically anonymized. This was done to prevent indirect identification of individuals or establishments.

The dataset and its annotations are available upon request from the authors, to ensure that the dataset is used responsibly and for academic purposes only, thus, respecting the original intent of the data collection. The Python code for the implementation of this evaluation and the documentation about the evaluation process is accessible via

GitHub.[9]

Despite our thorough data collection and anonymization procedures, some inherent limitations and ethical considerations persist. Our dataset may not capture the full spectrum of user sentiment due to potential bias in review writing, as those who write reviews may only represent a certain subset of the population. The ability to transfer knowledge about semantics and characteristics of reviews across different rating platforms cannot be guaranteed either. This inherent bias may be unintentionally perpetuated by BERT-based models used in our ABSA, despite their general effectiveness in NLP. In addition, our dataset was composed of reviews in German, which may include the bias of different language characteristics that might not be transferable to other languages.

5.3 Future Work

Our work provides valuable insight into the implementation and expected performance of a multi-label classification approach for detecting aspect categories and their associated polarities in reviews about the hotel industry. Importantly, we demonstrate that this methodology can be applied beyond social media to other domains in the German language. However, several potential directions for future work emerge from this study.

Foremost, we want to improve our dataset both in terms of size and annotation quality. Increasing the number of sentences in the dataset will provide a more robust representation of reviews, while a refined curation process ensures higher accuracy of labels. Currently, our dataset exhibits class imbalance, which presents challenges to the ABSA methods applied and can distort classification performance, particularly for underrepresented aspect categories.

From a methodological perspective, despite our results outperforming comparable ABSA approaches in both German and English languages, there is still room for improvement. We observed that the classification performance for severely underrepresented classes tends to decline significantly. To mitigate this, future efforts could involve optimizing training data balance via class weighting or subsampling, coupled with a more thorough hyperparameter tuning process.

Furthermore, we see great potential in further investigating the performance of large language models in the scenario of zero- or few-shot learning in the context of ABSA, which has already yielded remarkable results in the field of (aspect-based) sentiment analysis (Zhang et al., 2023; Qin et al., 2023).

[9] https://github.com/JakobFehle/absa-hotel-reviews

References

Sindhu Abro, Sarang Shaikh, Rizwan Ali Abro, Sana Fatima Soomro, and Hafiz Mehmood Malik. 2020. Aspect based sentimental analysis of hotel reviews: A comparative study. *Sukkur IBA Journal of Computing and Mathematical Sciences*, 4(1):11–20.

Muhammad Afzaal, Muhammad Usman, and Alvis Fong. 2019. Tourism mobile app with aspect-based sentiment classification framework for tourist reviews. *IEEE Transactions on Consumer Electronics*, 65(2):233–242.

Md Shad Akhtar, Abhishek Kumar, Asif Ekbal, Chris Biemann, and Pushpak Bhattacharyya. 2019. Language-agnostic model for aspect-based sentiment analysis. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 154–164.

Nadeem Akhtar, Nashez Zubair, Abhishek Kumar, and Tameem Ahmad. 2017. Aspect based sentiment oriented summarization of hotel reviews. *Procedia computer science*, 115:563–571.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Pulung Nurtantio Andono, Sunardi, Raden Arief Nugroho, and Budi Harjo. 2022. Aspect-based sentiment analysis for hotel review using lda, semantic similarity, and bert. *International Journal of Intelligent Engineering and Systems*, 15.

Matthias Aßenmacher, Alessandra Corvonato, and Christian Heumann. 2021. Re-evaluating germeval17 using german pre-trained language models. *arXiv preprint arXiv:2102.12330*.

Gianni Brauwers and Flavius Frasincar. 2022. A survey on aspect-based sentiment classification. *ACM Computing Surveys*, 55(4):1–37.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *arXiv preprint arXiv:2010.10906*.

Yung-Chun Chang, Chih-Hao Ku, and Chun-Hung Chen. 2019. Social media analytics: Extracting and visualizing hilton hotel ratings and reviews from tripadvisor. *International Journal of Information Management*, 48:263–279.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2022. Survey of aspect-based sentiment analysis datasets. *arXiv preprint arXiv:2204.05232*.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985.

Lore De Greve, Pranaydeep Singh, Cynthia Van Hee, Els Lefever, and Gunther Martens. 2021. Aspect-based sentiment analysis for german: analyzing'talk of literature'surrounding literary prizes on social media. *Computational Linguistics in the Netherlands Journal*, 11:85–104.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299.

Alina El-Keilany, Thomas Schmidt, and Christian Wolff. 2022. Distant Viewing of the Harry Potter Movies via Computer Vision. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022).*, pages 33–49, Uppsala, Sweden.

Jakob Fehle, Thomas Schmidt, and Christian Wolff. 2021. Lexicon-based sentiment analysis in German: Systematic evaluation of resources and preprocessing techniques. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 86–103, Düsseldorf, Germany. KONVENS 2021 Organizers.

David Halbhuber, Jakob Fehle, Alexander Kalus, Konstantin Seitz, Martin Kocur, Thomas Schmidt, and Christian Wolff. 2019. The mood game-how to use the player's affective state in a shoot'em up avoiding frustration and boredom. In *Proceedings of Mensch Und Computer 2019*, pages 867–870.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Zeyi Jin, Xin Lai, and Jingjig Cao. 2020. Multi-label sentiment analysis base on bert with modified tf-idf. In *2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN)*, pages 1–6. IEEE.

Zenun Kastrati, Ali Shariq Imran, and Arianit Kurti. 2020. Weakly supervised framework for aspect-based sentiment analysis on students' reviews of moocs. *IEEE Access*, 8:106799–106810.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

Roman Klinger and Philipp Cimiano. 2014. The usage review corpus for fine-grained, multi-lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Citeseer.

J Ashok Kumar, S Abirami, and Tina Esther Trueman. 2019. Multilabel aspect-based sentiment classification for abilify drug user review. In *2019 11th International Conference on Advanced Computing (ICoAC)*, pages 376–380. IEEE.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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.

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *Twenty-Fourth international joint conference on artificial intelligence*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Antonio Moreno-Ortiz, Soluna Salles-Bernal, and Aroa Orrequia-Barea. 2019. Design and validation of annotation schemas for aspect-based sentiment analysis in the tourism sector. *Information Technology & Tourism*, 21:535–557.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Luis Moßburger, Felix Wende, Kay Brinkmann, and Thomas Schmidt. 2020. Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 70–81, Barcelona, Spain (Online). Association for Computational Linguistics.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Anna-Marie Ortloff, Lydia Güntner, Maximiliane Windl, Thomas Schmidt, Martin Kocur, and Christian Wolff. 2019. Sentibooks: Enhancing audiobooks via affective computing and smart light bulbs. In *Proceedings of Mensch Und Computer 2019*, MuC'19, page 863–866, New York, NY, USA. Association for Computing Machinery.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

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.

Yao Qiang, Xin Li, and Dongxiao Zhu. 2020. Toward tag-free aspect based sentiment analysis: A multiple attention network approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose natural language processing task solver?

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand, editors. 2021. *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*. Association for Computational Linguistics, Duesseldorf, Germany.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Thomas Schmidt and Manuel Burghardt. 2018. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021a. Towards a Corpus of Historical German Plays with Emotion Annotations. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 9:1–9:11, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 2190-6807.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021b. Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays. In Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis, and Ulrike Wuttke, editors, *Fabrikation von Erkenntnis: Experimente in den Digital Humanities*. Melusina Press, Esch-sur-Alzette, Luxembourg.

Thomas Schmidt, Alina El-Keilany, Johannes Eger, and Sarah Kurek. 2021c. Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical

Movies. In *2nd International Conference of the European Association for Digital Humanities (EADH 2021)*, Krasnoyarsk, Russia.

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.

Thomas Schmidt, Florian Kaindl, and Christian Wolff. 2020. Distant reading of religious online communities: A case study for three religious forums on reddit. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*, pages 157–172, Riga, Latvia.

Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. *arXiv preprint arXiv:1808.09238*.

Uladzimir Sidarenka. 2016. Potts: the potsdam twitter sentiment corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1133–1141.

Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. 2019. Bert for stock market sentiment analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1597–1601.

Jie Tao and Xing Fang. 2020. Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7:1–26.

Thang Tran, Hung Ba, and Van-Nam Huynh. 2019. Measuring hotel review sentiment: An aspect-based sentiment analysis approach. In *Integrated Uncertainty in Knowledge Modelling and Decision Making: 7th International Symposium, IUKM 2019, Nara, Japan, March 27–29, 2019, Proceedings 7*, pages 393–405. Springer.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Ethan Xia, Han Yue, and Hongfu Liu. 2021. Tweet sentiment analysis of the 2020 us presidential election. In *Companion proceedings of the web conference 2021*, pages 367–371.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Hang Yan, Junqi Dai, Xipeng Qiu, Zheng Zhang, et al. 2021. A unified generative framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2106.04300*.

Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. 2014. Automatic annotation suggestions and custom annotation layers in webanno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96.

Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2044–2054.

Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

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.

A Appendix

A.1 Possible Combinations for Aspects and Attributes

Aspect	Sub-Aspect/Attribute
Food & Drinks	General (Universal Assessments)
	Price (Restaurant, Bar, Minibar)
	Style (Food Options, Extras)
	Quietness (Loudness in the Dining Area, Privacy)
Hotel	General (Universal Assessments)
	Price (Spa, Wellness, Fitness, Parking)
	Cleanliness
	Style (Furniture, Products, Convenience)
Location	General (Universal Assessments)
	Quietness (Traffic Noise)
	Price (Public Transport, Taxi)
Service	General (Universal Assessments, Friendliness, Helpfullness)
	Cleanliness
Rooms	General (Universal Assessments)
	Price (Stay)
	Quietness (Sleep, Noise)
	Cleanliness
	Style (Furniture, Size, Comfort)

Table 5: All possible combinations of aspects and their attributes.

A.2 Annotators Agreement for the Dataset Annotation

Ann. 1	Ann. 2	Size	Asp	Asp + Attr	Asp + Pol	Asp + Attr + Pol
2	9	200	0.74	0.59	0.66	0.53
4	7	400	0.76	0.61	0.66	0.53
3	13	400	0.72	0.52	0.65	0.49
15	17	400	0.65	0.56	0.62	0.54
5	18	400	0.63	0.5	0.60	0.48
24	25	400	0.66	0.51	0.58	0.47
14	16	400	0.62	0.49	0.56	0.45
8	10	400	0.66	0.55	0.55	0.46
1	20	400	0.58	0.46	0.54	0.43
11	23	400	0.61	0.51	0.52	0.44
12	21	200	0.54	0.42	0.47	0.35
6	22	200	0.55	0.39	0.4	0.29
26	27	400	0.46	0.36	0.37	0.28
2	19	400	0.29	0.25	0.27	0.22
Total/Mean	14	5000	0.61	0.48	0.54	0.43

Table 6: Krippendorf's α values with different levels of granularity for the 14 annotator pairings, sorted by α values of aspect-polarity combinations.

A.3 Dataset Excerpt

```
<documents>

    ...

    <document id="159">
        <opinions>
            <opinion category="SERVICE#HAUPT" polarity="positiv" />
        </opinions>
        <text>Der Service war sehr nett und es hat alles unkompliziert funktioniert</text>
    </document>
    <document id="160">
        <opinions>
            <opinion category="ZIMMER#STYLE" polarity="positiv" />
            <opinion category="ZIMMER#SAUBERKEIT" polarity="positiv" />
        </opinions>
        <text>Das Zimmer war schön eingerichtet, modern und sauber</text>
    </document>

        ...

</documents>
```

Figure 2: Example snippet of the dataset with two entries.

A.4 Class Frequencies for Aspect-Attribute Combinations

Aspect#Attribute	Count	Percentage
Service#General	901	16.0 %
Location#General	861	15.3 %
Rooms#Style	684	12.2 %
Food&Drinks#General	654	11.6 %
Hotel#General	629	11.2 %
...	...	...
Food&Drinks#Price	69	1.2 %
Rooms#Price	37	0.7 %
Location#Price	18	0.3 %
Food&Drinks#Quietness	10	0.2 %
Service#Cleanliness	6	0.1 %

Table 7: Amount of samples per aspect-attribute combination.

A.5 Class Frequencies for Aspect-Polarity Combinations

Aspect + Polarity	Count	Percentage
Hotel - Positive	1,062	18.9 %
Rooms - Positive	937	16.7 %
Service - Positive	743	13.2 %
Location - Positive	685	12.2 %
Food & Drinks - Positive	605	10.8 %
Rooms - Negative	405	7.2 %
Hotel - Negative	209	3.7 %
Hotel - Neutral	186	3.3 %
Location - Neutral	166	3.0 %
Rooms - Neutral	135	2.4 %
Service - Negative	128	2.3 %
Location - Negative	112	2.0 %
Food & Drinks - Neutral	105	1.9 %
Food & Drinks - Negative	103	1.8 %
Service - Neutral	36	0.6 %

Table 8: Amount of samples per aspect-polarity combination.

A.6 Class Frequencies for Aspect-Attribute-Polarity Combinations

Aspect#Attribute:Polarity	Count	Percentage
Service#General:Positive	740	13.1 %
Location#General:Positive	615	10.9 %
Food&Drinks#General:Positive	513	9.1 %
Hotel#General:Positive	485	8.6 %
Rooms#Style:Positive	428	7.6 %
...	...	...
Food&Drinks#Quietness:Positive	3	<0.1 %
Service#Cleanliness:Positive	3	<0.1 %
Service#Cleanliness:Neural	2	<0.1 %
Location#Price:Negative	1	<0.1 %
Service#Cleanliness:Negative	1	<0.1 %

Table 9: Amount of samples per aspect-attribute-polarity tuple.

A.7 Model Parameters and Characteristics of the Pre-Trained BERT models

Model	Parameters	Layers	Attention Heads	Training Data	Hidden States
deepset-gbert-large	335 M	24	16	161 GB	768
dbmdz-bert-base-german-uncased	110 M	12	12	16 GB	768
distilbert-base-german-cased	66 M	12	12	16 GB	1024

Table 10: Model parameters and characteristics for each of the 3 pre-trained BERT models.

A.8 Hyperparameter Configurations for the Best Runs

Task	Language Model	Learning Rate	Batch Size	Epochs	Runtime
Aspect	deepset-gbert-large	2.01 E-05	32	4	3 m 53 s
	dbmdz-bert-base-german-uncased	3.90 E-05	8	4	4 m 00 s
	distilbert-base-german-cased	5.00 E-05	8	3	1 m 51 s
Aspect + Attribute	deepset-gbert-large	2.06 E-05	8	4	10 m 07 s
	dbmdz-bert-base-german-uncased	2.82 E-05	8	3	3 m 04 s
	distilbert-base-german-cased	4.83 E-05	8	3	1 m 51 s
Aspect + Polarity	deepset-gbert-large	3.50 E-05	8	3	7 m 36 s
	dbmdz-bert-base-german-uncased	4.66 E-05	8	4	3 m 56 s
	distilbert-base-german-cased	3.97 E-05	8	4	2 m 26 s
Aspect + Attribute + Polarity	deepset-gbert-large	2.28 E-05	8	4	10 m 07 s
	dbmdz-bert-base-german-uncased	4.42 E-05	8	4	3 m 58 s
	distilbert-base-german-cased	4.99 E-05	8	3	1 m 51 s

Table 11: Best hyperparameter configuration for each model per task. Average runtime is given for a single train-eval run.