Re-Evaluating GermEval17 Using German Pre-Trained Language Models

Matthias Aßenmacher[1]♠ **Alessandra Corvonato**[1]♣ **Christian Heumann**[1]♠

[1] Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany

♠{matthias,chris}@stat.uni-muenchen.de, ♣alessandracorvonato@yahoo.de

Abstract

The lack of a commonly used benchmark data set (collection) such as (Super) GLUE (Wang et al., 2018, 2019) for the evaluation of non-English pre-trained language models is a severe shortcoming of current English-centric NLP-research. It concentrates a large part of the research on English, neglecting the uncertainty when transferring conclusions found for the English language to other languages. We evaluate the performance of German and multilingual BERT models currently available via the huggingface `transformers` library on four subtasks of Aspect-based Sentiment Analysis (ABSA) from the GermEval17 workshop. We compare them to pre-BERT architectures (Wojatzki et al., 2017; Schmitt et al., 2018; Attia et al., 2018) as well as to an ELMo-based architecture (Biesialska et al., 2020) and a BERT-based approach (Guhr et al., 2020). The observed improvements are put in relation to those for a similar ABSA task (Pontiki et al., 2014) and similar models (pre-BERT vs. BERT-based) for the English language and we check whether the reported improvements correspond to those we observe for German.

1 Introduction

(Aspect-based) Sentiment Analysis is often used to transform reviews into helpful information on how a product or service of a company is perceived among the customers. Until recently, Sentiment Analysis was mainly conducted using traditional machine learning and recurrent neural networks, like LSTMs (Hochreiter and Schmidhuber, 1997) or GRUs (Cho et al., 2014). Those models have been practically replaced by language models relying on (parts of) the Transformer architecture, a novel framework proposed by Vaswani et al. (2017). Devlin et al. (2019) developed a Transformer-encoder-based language model called BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransfomers), achieving state-of-the-art (SOTA) performance on several benchmark tasks - mainly for the English language - and becoming a milestone in the field of NLP.

Up to now, only a few researchers have focused on sentiment related problems for German reviews, despite language-specific evaluation is a crucial driving force for a more universal model development and improvement. Unique characteristics of the different languages present different challenges to the models, which is why sole evaluation on English data is a severe shortcoming.

The first shared task on German ABSA, which provides a large annotated data set for training and evaluation, is the *GermEval17 Shared Task* (Wojatzki et al., 2017). The participating teams back then analyzed the data using mostly standard machine learning techniques such as SVMs, CRFs, or LSTMs. In contrast to 2017, today, different pre-trained BERT models are available for a variety of different languages, including German. We re-analyzed the complete GermEval17 Task using seven pre-trained BERT models suitable for German provided by the huggingface `transformers` library (Wolf et al., 2020). We evaluate which one of the models is best suited for the different GermEval17 subtasks by comparing their performance values. Furthermore, we compare our findings on whether (and how much) BERT-based models are able to improve the pre-

BERT SOTA in German ABSA with the SOTA developments for English ABSA by the example of SemEval-2014 (Pontiki et al., 2014).

We first give an overview on the GermEval17 tasks (cf. Sec. 2) and on related work (cf. Sec. 3). Second, we present the data and the models (cf. Sec. 4), while Section 5 holds the results of our re-evaluation. Sections 6 and 7 conclude our work by stating our main findings and drawing parallels to the English language.

2 The GermEval17 Task(s)

The GermEval17 Shared Task (Wojatzki et al., 2017) is a task on analyzing aspect-based sentiments in customer reviews about "Deutsche Bahn" (DB) - the German public train company. The main data was crawled from various social media platforms such as Twitter, Facebook and Q&A websites from May 2015 to June 2016. The documents were manually annotated, and split into a training (**train**), a development (**dev**) and a synchronic (**test**$_{syn}$) test set. A diachronic test set (**test**$_{dia}$) was collected the same way from November 2016 to January 2017 in order to test for temporal robustness. The task comprises four subtasks representing a complete classification pipeline. Subtask A is a binary Relevance Classification task which aims at identifying whether the feedback refers to DB. Subtask B aims at classifying the Document-level Polarity ("negative", "positive" and "neutral"). In Subtask C, the model has to identify all the aspect categories with associated sentiment polarities in a relevant document. This multi-label classification task was divided into Subtask C1 (*Aspect-only*) and Subtask C2 (*Aspect+Sentiment*). For this purpose, the organizers defined 20 different aspect categories, e.g. `Allgemein` (*General*), `Sonstige Unregelmäßigkeiten` (*Other irregularities*). Finally, Subtask D refers to the Opinion Target Extraction (OTE), i.e. a sequence labeling task extracting the linguistic phrase used to express an opinion. We differentiate between exact match (Subtask D1) and overlapping match, tolerating errors of $+/-$ one token (Subtask D2).

3 Related Work

Already before BERT, many researchers focused on (English) Sentiment Analysis (Behdenna et al., 2018). The most common architectures were traditional machine learning classifiers and recurrent neural networks (RNNs). SemEval14 (Task 4; Pontiki et al., 2014) was the first workshop to introduce Aspect-based Sentiment Analysis (ABSA) which was expanded within SemEval15 Task 12 (Pontiki et al., 2015) and SemEval16 Task 5 (Pontiki et al., 2016). Here, restaurant and laptop reviews were examined on different granularities. The best model at SemEval16 was an SVM/CRF architecture using GloVe embeddings (Pennington et al., 2014). However, many works recently focused on re-evaluating the SemEval Sentiment Analysis task using BERT-based language models (Hoang et al., 2019; Xu et al., 2019; Sun et al., 2019; Li et al., 2019; Karimi et al., 2020; Tao and Fang, 2020).

In comparison, little research deals with German ABSA. For instance, Barriere and Balahur (2020) trained a multilingual BERT model for German Document-level Sentiment Analysis on the SB-10k data set (Cieliebak et al., 2017). Regarding the GermEval17 Subtask B, Guhr et al. (2020) considered both FastText (Bojanowski et al., 2017) and BERT, achieving notable improvements. Biesialska et al. (2020) made use of ensemble models: One is an ensemble of ELMo (Peters et al., 2018), GloVe and a bi-attentive classification network (BCN; McCann et al., 2017), achieving a score of 0.782, and the other one consists of ELMo and a Transformer-based Sentiment Analysis model (TSA), reaching a score of 0.789 for the synchronic test data set. Moreover, Attia et al. (2018) trained a convolutional neural network (CNN), achieving a score of 0.7545 on the synchronic test set. Schmitt et al. (2018) advanced the SOTA for Subtask C by employing biLSTMs and CNNs to carry out end-to-end Aspect-based Sentiment Analysis. The highest score was achieved using an end-to-end CNN architecture with FastText embeddings, scoring 0.523 and 0.557 on the synchronic and diachronic test data set for Subtask C1, respectively, and 0.423 and 0.465 for Subtask C2.

4 Materials and Methods

Data The GermEval17 data is freely available in `.xml`- and `.tsv`-format[1]. Each data split (train, validation, test) in `.tsv`-format contains the following variables:

- document id (URL)
- document text
- relevance label (`true`, `false`)

[1]The data sets (in both formats) can be obtained from http://ltdata1.informatik.uni-hamburg.de/germeval2017/.

- document-level sentiment label (`negative`, `neutral`, `positive`)
- aspects with respective polarities (e.g. `Ticketkauf#Haupt:negative`)

For documents which are annotated as irrelevant, the sentiment label is set to `neutral` and no aspects are available. Visibly, the `.tsv`-formatted data does not contain the target expressions or their associated sequence positions. Consequently, Subtask D can only be conducted using the data in `.xml`-format, which additionally holds the information on the starting and ending sequence positions of the target phrases.

The data set comprises $\sim$ 26k documents in total, including the diachronic test set with around 1.8k examples. Further, the main data was randomly split by the organizers into a train data set for training, a development data set for validation and a synchronic test data set. Table 1 displays the number of documents for each split.

train	**dev**	**test**$_{syn}$	**test**$_{dia}$
19,432	2,369	2,566	1,842

Table 1: Number of documents per split of the data set.

While roughly 74% of the documents form the train set, the development split and the synchronic test split contain around 9% and around 10%, respectively. The remaining 7% of the data belong to the diachronic set (cf. Tab. 1). Table 2 shows the relevance distribution per data split. This unveils a pretty skewed distribution of the labels since the relevant documents represent the clear majority with over 80% in each split.

Relevance	**train**	**dev**	**test**$_{syn}$	**test**$_{dia}$
true	16,201	1,931	2,095	1,547
false	3,231	438	471	295

Table 2: Relevance distribution for Subtask A.

The distribution of the sentiments is depicted in Table 3, which shows that between 65% and 69% (per split) belong to the neutral class, 25–31% to the negative and only 4–6% to the positive class.

Table 4 holds the distribution of the 20 different aspect categories assigned to the documents[2]. It shows the number of documents containing certain categories without differentiating between how often a category appears within a given document.

Sentiment	**train**	**dev**	**test**$_{syn}$	**test**$_{dia}$
negative	5,045	589	780	497
neutral	13,208	1,632	1,681	1,237
positive	1,179	148	105	108

Table 3: Sentiment distribution for Subtask B.

The relative distribution of the aspect categories is similar between the splits. On average, there are $\sim$ 1.12 different aspects per document. Again, the label distribution is heavily skewed, with `Allgemein` (*General*) clearly representing the majority class, as it is present in 75.8% of the documents with aspects. The second most frequent category is `Zugfahrt` (*Train ride*) appearing in around 13.8% of the documents. This strong imbalance in the aspect categories leads to an almost Zipfian distribution (Wojatzki et al., 2017).

Category	**train**	**dev**	**test**$_{syn}$	**test**$_{dia}$
Allgemein	11,454	1,391	1,398	1,024
Zugfahrt	1,687	177	241	184
Sonstige Unregelmäßigkeiten	1,277	139	224	164
Atmosphäre	990	128	148	53
Ticketkauf	540	64	95	48
Service und Kundenbetreuung	447	42	63	27
Sicherheit	405	59	84	42
Informationen	306	28	58	35
Connectivity	250	22	36	73
Auslastung und Platzangebot	231	25	35	20
DB App und Website	175	20	28	18
Komfort und Ausstattung	125	18	24	11
Barrierefreiheit	53	14	9	2
Image	42	6	0	3
Toiletten	41	5	7	4
Gastronomisches Angebot	38	2	3	3
Reisen mit Kindern	35	3	7	2
Design	29	3	4	2
Gepäck	12	2	2	6
QR-Code	0	1	1	0
total	18,137	2,149	2,467	1,721
# documents with aspects	16,200	1,930	2,095	1,547
∅ different aspects/document	1.12	1.11	1.18	1.11

Table 4: Aspect category distribution for Subtask C. Multiple mentions of the same aspect category in a document are only considered once.

Pre-trained architectures BERT was initially introduced in a `base` (110M parameters) and a `large` (340M) variant, Sanh et al. (2019) proposed an even smaller BERT model (DistilBERT, 60M parameters) trained via knowledge distillation

[2]Multiple annotations per document are possible; for a detailed category description see https://sites.google.com/view/germeval2017-absa/data.

Model variant	Pre-training corpus	Properties
`bert-base-german-cased`	12GB of German text (deepset.ai)	L=12, H=768, A=12, 110M parameters
`bert-base-german-dbmdz-cased`	16GB of German text (dbmdz)	L=12, H=768, A=12, 110M parameters
`bert-base-german-dbmdz-uncased`	16GB of German text (dbmdz)	L=12, H=768, A=12, 110M parameters
`bert-base-multilingual-cased`	Largest Wikipedias (top 104 languages)	L=12, H=768, A=12, 179M parameters
`bert-base-multilingual-uncased`	Largest Wikipedias (top 102 languages)	L=12, H=768, A=12, 168M parameters
`distilbert-base-german-cased`	16GB of German text (dbmdz)	L=6, H=768, A=12, 66M parameters
`distilbert-base-multilingual-cased`	Largest Wikipedias (top 104 languages)	L=6, H=768, A=12, 134M parameters

Table 5: Pre-trained models provided by huggingface `transformers` (version 4.0.1) suitable for German. For **all** available models, see: `https://huggingface.co/transformers/pretrained_models.html`.

(Hinton et al., 2015). The exact model specifications regarding number of layers (L), number of attention heads (A) and embedding size (H) for available German BERT models are depicted in the last column of Table 5. Both architectures were pre-trained on the Masked Language Modeling task as well as on the auxiliary Next Sentence Prediction task (only BERT) and can subsequently be fine-tuned on a task at hand.

We include three German (Distil)BERT models pre-trained by DBMDZ[3] and one by Deepset.ai[4]. The latter one is pre-trained using German `Wikipedia` (6GB raw text files), the `Open Legal Data` dump (2.4GB; Ostendorff et al., 2020) and news articles (3.6GB). DBMDZ combined `Wikipedia`, `EU Bookshop` (Skadiņš et al., 2014), `Open Subtitles` (Lison and Tiedemann, 2016), `CommonCrawl` (Ortiz Suárez et al., 2019), `ParaCrawl` (Esplà-Gomis et al., 2019) and `News Crawl` (Haddow, 2018) to a corpus with a total size of 16GB with $\sim$ 2,350M tokens. Besides this, we use the three multilingual (Distil)BERT models included in the `transformers` module. This amounts to five BERT and two DistilBERT models, two of which are "uncased" (i.e. every character is lower-cased) while the other five models are "cased" ones.

5 Results

For the re-evaluation, we used the latest data provided in `.xml`-format. Duplicates were not removed, in order to make our results as comparable as possible. We tokenized the documents and fixed single spelling mistakes in the labels[5]. For Subtask D, the BIO-tags were added based on the provided sequence positions, i.e. one entity corresponds to at least one token tag starting with `B-` for "Beginning" and continuing with `I-` for "Inner". If a token does not belong to any entity, the tag `O` for "Outer" is assigned. For instance, the sequence "fährt nicht" (engl. "does not run") consists of two tokens and would receive the entity `Zugfahrt:negative` and the token tags [`B-Zugfahrt:negative`, `I-Zugfahrt:negative`] if it refers to a DB train which is not running.

The models were fine-tuned on one Tesla V100 PCIe 16GB GPU using Python 3.8.7. Moreover, the `transformers` module (version 4.0.1) and `torch` (version 1.7.1) were used[6]. The considered values for the hyperparameters for fine-tuning follow the recommendations of Devlin et al. (2019):

- Batch size $\in \{16, 32\}$,
- Adam learning rate $\in \{5e,3e,2e\} - 5$,
- # epochs $\in \{2, 3, 4\}$.

After evaluating the model performance for combinations[7] of the different hyperparameters, all pre-trained architectures were fine-tuned with a learning rate of 5e-5 for four epochs, which turned out to be the most promising combination across the different models. The maximum sequence length was set to 256, which is sufficient since the evaluated data set consists of rather short texts from social media, and a batch size of 32 was chosen.

Other models Eight teams officially participated in the GermEval17 shared task, five of which analyzed Subtask A, all of them Subtask B and two repectively Subtask C and D. We furthermore consider the system by Ruppert et al. (2017) additionally to the participants' models from 2017, even

[3]MDZ Digital Library team at the Bavarian State Library. Visit https://www.digitale-sammlungen.de for details and https://github.com/dbmdz/berts for their repository on pre-trained BERT models.

[4]Visit https://deepset.ai/german-bert for details.

[5]"positve" in **train** set was replaced with "positive", " negative" in **test**$_{dia}$ set was replaced with "negative".

[6]Source code is available on GitHub: https://github.com/ac74/reevaluating_germeval2017. The results are fully reproducible for Subtasks A, B and C. For Subtask D, reproducibility could not be ensured. The micro F1 scores fluctuate across different runs between +/-0.01 around the reported values.

[7]Due to memory limitations, not every hyperparameter combination was applicable.

though they were the organizers and did not "officially" participate. They also tackled all four subtasks. Since 2017 several other authors analyzed (parts of) the GermEval17 subtasks using more advanced models, which we also consider for comparison here. Table 6 shows which authors employed which kinds of models to solve which task.

Subtask	A	B	C1	C2	D1	D2
Models from 2017 (Wojatzki et al., 2017; Ruppert et al., 2017)	X	X	X	X	X	X
Our BERT models	X	X	X	X	X	X
CNN (Attia et al., 2018)	–	X	–	–	–	–
CNN+FastText (Schmitt et al., 2018)	–	–	X	X	–	–
ELMo+GloVe+BCN (Biesialska et al., 2020)	–	X	–	–	–	–
ELMo+TSA (Biesialska et al., 2020)	–	X	–	–	–	–
FastText (Guhr et al., 2020)	–	X	–	–	–	–
`bert-base-german-cased` (Guhr et al., 2020)	–	X	–	–	–	–

Table 6: An overview on all the models discussed in this article, an "X" in a column indicates that the architecture was evaluated on the respective subtask.

Subtask A The Relevance Classification is a binary document classification task with classes `true` and `false`. Table 7 displays the micro F1 score obtained by each language model on each test set (best result per data set in bold).

Language model	$\textbf{test}_{syn}$	$\textbf{test}_{dia}$
Best model 2017 (Sayyed et al., 2017)	0.903	0.906
`bert-base-german-cased`	0.950	0.939
`bert-base-german-dbmdz-cased`	0.951	0.946
`bert-base-german-dbmdz-uncased`	**0.957**	**0.948**
`bert-base-multilingual-cased`	0.942	0.933
`bert-base-multilingual-uncased`	0.944	0.939
`distilbert-base-german-cased`	0.944	0.939
`distilbert-base-multilingual-cased`	0.941	0.932

Table 7: F1 scores for Subtask A on synchronic and diachronic test sets.

All the models outperform the best result achieved in 2017 for both test data sets. For the synchronic test set, the previous best result is surpassed by 3.8–5.4 percentage points. For the diachronic test set, the absolute difference to the best contender of 2017 varies between 2.6 and 4.2 percentage points. With a micro F1 score of 0.957 and 0.948, respectively, the best scoring pre-trained language model is the uncased German BERT-BASE variant by `dbmdz`, followed by its cased version. All the pre-trained models perform slightly better on the synchronic test data than on the diachronic data. Attia et al. (2018), Schmitt et al. (2018), Biesialska et al. (2020) and Guhr et al. (2020) did not evaluate their models on this task.

Subtask B Subtask B refers to the Document-level Polarity, which is a multi-class classification task with three classes. Table 8 demonstrates the performances on the two test sets:

Language model	$\textbf{test}_{syn}$	$\textbf{test}_{dia}$
Best models 2017 ($\textbf{test}_{syn}$: Ruppert et al., 2017) ($\textbf{test}_{dia}$: Sayyed et al., 2017)	0.767	0.750
`bert-base-german-cased`	0.798	0.793
`bert-base-german-dbmdz-cased`	0.799	0.785
`bert-base-german-dbmdz-uncased`	**0.807**	**0.800**
`bert-base-multilingual-cased`	0.790	0.780
`bert-base-multilingual-uncased`	0.784	0.766
`distilbert-base-german-cased`	0.798	0.776
`distilbert-base-multilingual-cased`	0.777	0.770
CNN (Attia et al., 2018)	0.755	–
ELMo+GloVe+BCN (Biesialska et al., 2020)	0.782	–
ELMo+TSA (Biesialska et al., 2020)	0.789	–
FastText (Guhr et al., 2020)	0.698†	–
`bert-base-german-cased` (Guhr et al., 2020)	0.789†	–

Table 8: Micro-averaged F1 scores for Subtask B on synchronic and diachronic test sets.
†*Guhr et al. (2020) created their own (balanced & unbalanced) data splits, which limits comparability. We compare to the performance on the unbalanced data since it more likely resembles the original data splits.*

All models outperform the best model from 2017 by 1.0–4.0 percentage points for the synchronic, and by 1.6–5.0 percentage points for the diachronic test set. On the synchronic test set, the uncased German BERT-BASE model by `dbmdz` performs best with a score of 0.807, followed by its cased variant with 0.799. For the diachronic test set, the uncased German BERT-BASE model exceeds the other models with a score of 0.800, followed by the cased German BERT-BASE model reaching a score of 0.793. The three multilingual models perform generally worse than the German models on this task. Besides this, all the models perform slightly better on the synchronic data set than on the diachronic one. The FastText-based model (Guhr et al., 2020) comes not even close to the baseline from 2017, while the ELMo-based models (Biesialska et al., 2020) are pretty competitive. Interestingly, two of the multilingual models are even outperformed by these ELMo-based models.

Subtask C Subtask C is split into *Aspect-only* (Subtask C1) and *Aspect+Sentiment* Classification (Subtask C2), each being a multi-label classification task[8]. As the organizers provide 20 aspect categories, Subtask C1 includes 20 labels, whereas Subtask C2 has 60 labels since each aspect category

[8]This leads to a change of activation functions in the final layer from softmax to sigmoid + binary cross entropy loss.

can be combined with each of the three sentiments. Consistent with Lee et al. (2017) and Mishra et al. (2017), we do not account for multiple mentions of the same label in one document. The results for Subtask C1 are shown in Table 9:

Language model	test_{syn}	test_{dia}
Best model 2017 (Ruppert et al., 2017)	0.537	0.556
`bert-base-german-cased`	0.756	0.762
`bert-base-german-dbmdz-cased`	0.756	0.781
`bert-base-german-dbmdz-uncased`	**0.761**	**0.791**
`bert-base-multilingual-cased`	0.706	0.734
`bert-base-multilingual-uncased`	0.723	0.752
`distilbert-base-german-cased`	0.738	0.768
`distilbert-base-multilingual-cased`	0.716	0.744
CNN+FastText (Schmitt et al., 2018)	0.523	0.557

Table 9: Micro-averaged F1 scores for Subtask C1 (*Aspect-only*) on synchronic and diachronic test sets. A detailed overview of *per-class* performances for error analysis can be found in Table 15 in Appendix A.

All pre-trained German BERTs clearly surpass the best performance from 2017 as well as the results reported by Schmitt et al. (2018), who are the only ones of the other authors to evaluate their models on this tasks. Regarding the synchronic test set, the absolute improvement ranges between 16.9 and 22.4 percentage points, while for the diachronic test data, the models outperform the previous results by 17.8–23.5 percentage points. The best model is again the uncased German BERT-BASE model by `dbmdz`, reaching scores of 0.761 and 0.791, respectively, followed by the two cased German BERT-BASE models. One more time, the multilingual models exhibit the poorest performances amongst the evaluated models. Next, Table 10 shows the results for Subtask C2:

Language model	test_{syn}	test_{dia}
Best model 2017 (Ruppert et al., 2017)	0.396	0.424
`bert-base-german-cased`	0.634	0.663
`bert-base-german-dbmdz-cased`	0.628	0.663
`bert-base-german-dbmdz-uncased`	**0.655**	**0.689**
`bert-base-multilingual-cased`	0.571	0.634
`bert-base-multilingual-uncased`	0.553	0.631
`distilbert-base-german-cased`	0.629	0.663
`distilbert-base-multilingual-cased`	0.589	0.642
CNN+FastText (Schmitt et al., 2018)	0.423	0.465

Table 10: Micro-averaged F1 scores for Subtask C2 (*Aspect+Sentiment*) on synchronic and diachronic test sets. A detailed overview of *per-class* performances for error analysis can be found in Table 16 in Appendix A.

Here, the pre-trained models surpass the best model from 2017 by 15.7–25.9 percentage points and 20.7–26.5 percentage points, respectively, for the synchronic and diachronic test sets. Again, the best model is the uncased German BERT-BASE `dbmdz` model reaching scores of 0.655 and 0.689, respectively. The CNN models (Schmitt et al., 2018) are also outperformed. For both, Subtask C1 and C2, all the displayed models perform better on the diachronic than on the synchronic test data.

Subtask D Subtask D refers to the Opinion Target Extraction (OTE) and is thus a token-level classification task. As this is a rather difficult task, Wojatzki et al. (2017) distinguish between exact (Subtask D1) and overlapping match (Subtask D2), tolerating a deviation of $+/-$ one token. Here, "entities" are identified by their BIO-tags. It is noteworthy that there are less entities here than for Subtask C since document-level aspects or sentiments could not always be assigned to a certain sequence in the document. As a result, there are less documents at disposal for this task, namely 9,193. The remaining data has 1.86 opinions per document on average. The majority class is now `Sonstige Unregelmäßigkeiten:negative` with around 15.4% of the true entities (16,650 in total), leading to more balanced data than in Subtask C.

	Language model	test_{syn}	test_{dia}
	Best model 2017 (Ruppert et al., 2017)	0.229	0.301
without CRF	`bert-base-german-cased`	0.460	0.455
	`bert-base-german-dbmdz-cased`	0.480	0.466
	`bert-base-german-dbmdz-uncased`	0.492	0.501
	`bert-base-multilingual-cased`	0.447	0.457
	`bert-base-multilingual-uncased`	0.429	0.404
	`distilbert-base-german-cased`	0.347	0.357
	`distilbert-base-multilingual-cased`	0.430	0.419
with CRF	`bert-base-german-cased`	0.446	0.443
	`bert-base-german-dbmdz-cased`	0.466	0.444
	`bert-base-german-dbmdz-uncased`	**0.515**	**0.518**
	`bert-base-multilingual-cased`	0.472	0.466
	`bert-base-multilingual-uncased`	0.477	0.452
	`distilbert-base-german-cased`	0.424	0.403
	`distilbert-base-multilingual-cased`	0.436	0.418

Table 11: Entity-level micro-averaged F1 scores for Subtask D1 (*exact match*) on synchronic and diachronic test sets. A detailed overview of *per-class* performances for error analysis can be found in Table 17 in Appendix B.

In Table 11, we compare the pre-trained models using an "ordinary" softmax layer to when using a CRF layer for Subtask D1.

The best performing model is the uncased German BERT-BASE model by `dbmdz` with CRF layer on both test sets, with a score of 0.515 and 0.518, respectively. Overall, the results from 2017 are outperformed by 11.8–28.6 percentage points

	Language model	test_{syn}	test_{dia}
	Best models 2017 (test_{syn}: Lee et al., 2017) (test_{dia}: Ruppert et al., 2017)	0.348	0.365
without CRF	bert-base-german-cased	0.471	0.474
	bert-base-german-dbmdz-cased	0.491	0.488
	bert-base-german-dbmdz-uncased	0.501	0.518
	bert-base-multilingual-cased	0.457	0.473
	bert-base-multilingual-uncased	0.435	0.417
	distilbert-base-german-cased	0.397	0.407
	distilbert-base-multilingual-cased	0.433	0.429
with CRF	bert-base-german-cased	0.455	0.457
	bert-base-german-dbmdz-cased	0.476	0.469
	bert-base-german-dbmdz-uncased	**0.523**	**0.533**
	bert-base-multilingual-cased	0.476	0.474
	bert-base-multilingual-uncased	0.484	0.464
	distilbert-base-german-cased	0.433	0.423
	distilbert-base-multilingual-cased	0.442	0.427

Table 12: Entity-level micro-averaged F1 scores for Subtask D2 (*overlapping match*) on synchronic and diachronic test sets. A detailed overview of *per-class* performances for error analysis can be found in Table 18 in Appendix B.

on the synchronic test set and 5.6–21.7 percentage points on the diachronic test set.

For the overlapping match (cf. Tab. 12), the best system from 2017 are outperformed by 4.9–17.5 percentage points on the synchronic and by 4.2–16.8 percentage points on the diachronic test set. Again, the uncased German BERT-BASE model by `dbmdz` with CRF layer performs best with an micro F1 score of 0.523 on the synchronic and 0.533 on the diachronic set. To our knowledge, there were no other models to compare our performance values with, besides the results from 2017.

Main Takeaways For the first two subtasks, which are rather simple binary and multi-class classification tasks, the pre-trained models are able to improve a little upon the already pretty decent performance values from 2017. Further, we do not see large differences between the different pre-trained models. Nevertheless, the small differences we can observe, already point in the same direction as what can be observed for the primary ABSA tasks of interest, C1 and C2:

- Uncased models have a tendency of outperforming their cased counterparts for the monolingual models, for multilingual models this cannot be clearly confirmed.
- Monolingual models outperform the multilingual ones.
- There are no large performance differences between the two cased BERT models by DBMDZ and Deepset.ai, which suggests only a minor influence of the different corpora, which the models were pre-trained on.
- The monolingual DistilBERT model is pretty competitive, it consistently outperforms its multilingual counterpart as well as the multilingual BERT models on the subtasks A – C and is at least competitive to the monolingual BERT models.

For D1 and D2 we observe a rather clear dominance of the uncased monolingual model which is not observable to this extent for the other tasks.

6 Discussion

After having observed a notable performance increase for German ABSA when employing pre-trained models, the next step is to compare these observations to what was reported for the English language. Therefore, we examine the temporal development of the SOTA performance on the most widely adopted data sets for English ABSA, originating from the SemEval Shared Tasks (Pontiki et al., 2014, 2015, 2016). When looking at public leaderboards, e.g. https://paperswithcode.com/, Subtask SB2 (*aspect term polarity*) from SemEval-2014 is the task which attracts most of the researchers. This task is related, but not perfectly similar, to Subtask C2, since in this case, the *aspect term* is always a word which has to present in the given review. For this task, a comparison of pre-BERT and BERT-based methods reveals no big "jump" in the performance values, but rather a steady increase over time (cf. Tab. 13).

	Language model	Laptops	Restaurants
pre-BERT	Best model SemEval-2014 (Pontiki et al., 2014)	0.7048	0.8095
	MemNet (Tang et al., 2016)	0.7221	0.8095
	HAPN (Li et al., 2018)	0.7727	0.8223
BERT-based	BERT-SPC (Song et al., 2019)	0.7899	0.8446
	BERT-ADA (Rietzler et al., 2020)	0.8023	0.8789
	LCF-ATEPC (Yang et al., 2019)	0.8229	0.9018

Table 13: Development of the SOTA Accuracy for the aspect term polarity task (SemEval-2014; Pontiki et al., 2014). Selected models were picked from https://paperswithcode.com/sota/aspect-based-sentiment-analysis-on-semeval.

Clearly more related, but unfortunately also less used, are the subtasks SB3 (*aspect category extraction*; comparable to Subtask C1) and SB4 (*aspect category polarity*; comparable to Subtask C2)

from SemEval-2014.[9] Limitations with respect to comparability arise from the different numbers of categories: Subtask SB4 only exhibits five aspect categories (as opposed to 20 categories for GermEval17) which leads to an easier classification problem and is reflected in the already pretty high scores of the 2014 baselines. Table 14 shows the performance of the best model from 2014 as well as performance of subsequent (pre-BERT and BERT-based) models for subtasks SB3 and SB4.

	Language model	Restaurants SB3	Restaurants SB4
pre-BERT	Best model SemEval-2014 (Pontiki et al., 2014)	0.8857	0.8292
	ATAE-LSTM (Wang et al., 2016)	—	0.840
BERT-based	BERT-pair (Sun et al., 2019)	0.9218	0.899
	CG-BERT (Wu and Ong, 2020)	0.9162†	0.901†
	QACG-BERT (Wu and Ong, 2020)	0.9264	0.904†

Table 14: Development of the SOTA F1 score (SB3) and Accuracy (SB4) for the aspect category extraction/polarity task (SemEval-2014; Pontiki et al., 2014). †*Additional auxiliary sentences were used.*

In contrast to what can be observed for SB2, in this case, the performance increase on SB4 caused by the introduction of BERT seems to be kind of striking. While the ATAE-LSTM (Wang et al., 2016) only slightly increased the performance compared to 2014, the BERT-based models led to a jump of more than 6 percentage points. So when taking into account the potential room for improvement (0.16 for SB4 vs. 0.60 for C2), the improvements *relative* to the potential (0.06/0.16 for SB4 vs. 0.23/0.60 for C2) are quite similar.

Another issue is that (partly) highly specialized (T)ABSA architectures were used for improving the SOTA on the SemEval-2014 tasks, while we "only" applied standard pre-trained German BERT models without any task-specific modifications or extensions. This leaves room for further improvements on this task on German data which should be an objective for future research.

[9] Since the data sets (*Restaurants* and *Laptops*) have been further developed for SemEval-2015 and SemEval-2016, subtasks SB3 and SB4 are revisited under the names Slot 1 and Slot 3 for the in-domain ABSA in SemEval-2015. Slot 2 from SemEval-2015 aims at OTE and thus corresponds to Subtask D from GermEval17. For SemEval-2016 the same task names as in 2015 were used, subdivided into Subtask 1 (*sentence-level ABSA*) and Subtask 2 (*text-level ABSA*).

7 Conclusion

As one would have hoped, all the state-of-the art pre-trained language models clearly outperform all the models from 2017, proving the power of transfer learning also for German ABSA. Throughout the presented analyses, the models always achieve similar results between the synchronic and the diachronic test sets, indicating temporal robustness for the models. Nonetheless, the diachronic data was collected *only* half a year after the main data. It would be interesting to see whether the trained models would return similar predictions on data collected a couple of years later.

The uncased German BERT-BASE model by `dbmdz` achieves the best results across all subtasks. Since Rönnqvist et al. (2019) showed that monolingual BERT models often outperform the multilingual models for a variety of tasks, one might have already suspected that a monolingual German BERT performs best across the performed tasks. It may not seem evident at first that an uncased language model ends up as the best performing model since, e.g. in Sentiment Analysis, capitalized letters might be an indicator for polarity. In addition, since nouns and beginnings of sentences always start with a capital letter in German, one might assume that lower-casing the whole text changes the meaning of some words and thus confuses the language model. Nevertheless, the GermEval17 documents are very noisy since they were retrieved from social media. That means that the data contains many misspellings, grammar and expression mistakes, dialect, and colloquial language. For this reason, already some participating teams in 2017 pursued an elaborate pre-processing on the text data in order to eliminate some noise (Hövelmann and Friedrich, 2017; Sayyed et al., 2017; Sidarenka, 2017). Among other things, Hövelmann and Friedrich (2017) transformed the text to lower-case and replaced, for example, "S-Bahn" and "S Bahn" with "sbahn". We suppose that in this case, lower-casing the texts improves the data quality by eliminating some of the noise and acts as a sort of regularization. As a result, the uncased models potentially generalize better than the cased models. The findings from Mayhew et al. (2019), who compare cased and uncased pre-trained models on social media data for NER, corroborate this hypothesis.

References

Mohammed Attia, Younes Samih, Ali Elkahky, and Laura Kallmeyer. 2018. Multilingual multi-class sentiment classification using convolutional neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Valentin Barriere and Alexandra Balahur. 2020. Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 266–271, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Salima Behdenna, Fatiha Barigou, and Ghalem Belalem. 2018. Document level sentiment analysis: A survey. *EAI Endorsed Transactions on Context-aware Systems and Applications*, 4:154339.

Katarzyna Biesialska, Magdalena Biesialska, and Henryk Rybinski. 2020. Sentiment analysis with contextual embeddings and self-attention. *arXiv preprint arXiv:2003.05574*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

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.

M. Esplà-Gomis, M. Forcada, Gema Ramírez-Sánchez, and Hieu T. Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *MT-Summit*.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans-Joachim Böhme. 2020. Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 1627–1632, Marseille, France.

Barry Haddow. 2018. News Crawl Corpus.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland. Linköping University Electronic Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Leonard Hövelmann and Christoph M. Friedrich. 2017. Fasttext and Gradient Boosted Trees at GermEval-2017 Tasks on Relevance Classification and Document-level Polarity. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020. Adversarial training for aspect-based sentiment analysis with bert.

Ji-Ung Lee, Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. UKP TU-DA at GermEval 2017: Deep Learning for Aspect Based Sentiment Detection. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany.

Lishuang Li, Yang Liu, and AnQiao Zhou. 2018. Hierarchical attention based position-aware network for aspect-level sentiment analysis. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 181–189, Brussels, Belgium. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processingand the*

9th International Joint Conference on Natural Language Processing, pages 6256–6261, Hong Kong, China. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, volume 30, pages 6294–6305. Curran Associates, Inc.

Pruthwik Mishra, Vandan Mujadia, and Soujanya Lanka. 2017. GermEval 2017: Sequence based Models for Customer Feedback Analysis. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an Open Platform for Legal Information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, pages 385—-388, New York, NY, USA. Association for Computing Machinery.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee de clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeny Kotelnikov, Nuria Bel, Salud María Zafra, and Gülşen Eryiğit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

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

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is Multilingual BERT Fluent in Language Generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Eugen Ruppert, Abhishek Kumar, and Chris Biemann. 2017. LT-ABSA: An Extensible Open-Source System for Document-Level and Aspect-Based Sentiment Analysis. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany.

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*.

Zeeshan Ali Sayyed, Daniel Dakota, and Sandra Kübler. 2017. IDS-IUCL: Investigating Feature Selection and Oversampling for GermEval 2017. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1114, Brussels, Belgium. Association for Computational Linguistics.

Uladzimir Sidarenka. 2017. PotTS at GermEval-2017 Task B: Document-Level Polarity Detection Using Hand-Crafted SVM and Deep Bidirectional LSTM Network. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, California, USA.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in neural information processing systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

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. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12, Berlin, Germany.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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, Online. Association for Computational Linguistics.

Zhengxuan Wu and Desmond C Ong. 2020. Context-guided bert for targeted aspect-based sentiment analysis. *arXiv preprint arXiv:2010.07523*.

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

Heng Yang, Biqing Zeng, JianHao Yang, Youwei Song, and Ruyang Xu. 2019. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *arXiv preprint arXiv:1912.07976*.

Appendix

A Detailed results (per category) for Subtask C

It may be interesting to have a more detailed look at the model performance for this subtask because of the high number of classes and their skewed distribution by investigating the performance on category-level. Table 15 shows the performance of the uncased German BERT-BASE model by `dbmdz` per test set for Subtask C1. The support indicates the number of appearances, which are also displayed in Table 4 in this case. Seven categories are summarized in *Rest* because they have an F1 score of 0 for both test sets, i.e. the model is not able to correctly identify any of these seven aspects appearing in the test data. The table is sorted by the score on the synchronic test set.

	test$_{syn}$		**test**$_{dia}$	
Aspect Category	**Score**	**Support**	**Score**	**Support**
Allgemein	0.854	1,398	0.877	1,024
Sonstige Unregelmäßigkeiten	0.782	224	0.785	164
Connectivity	0.750	36	0.838	73
Zugfahrt	0.678	241	0.687	184
Auslastung und Platzangebot	0.645	35	0.667	20
Sicherheit	0.602	84	0.639	42
Atmosphäre	0.600	148	0.532	53
Barrierefreiheit	0.500	9	0	2
Ticketkauf	0.481	95	0.506	48
Service und Kundenbetreuung	0.476	63	0.417	27
DB App und Website	0.455	28	0.563	18
Informationen	0.329	58	0.464	35
Komfort und Ausstattung	0.286	24	0	11
Rest	*0*	*24*	*0*	*20*

Table 15: Micro-averaged F1 scores and support by aspect category (Subtask C1). Seven categories are summarized in *Rest* and show each a score of 0.

The F1 scores for `Allgemein` (*General*), `Sonstige Unregelmäßigkeiten` (*Other ir-*

regularities) and `Connectivity` are the highest. 13 categories, mostly similar between the two test sets, show a positive F1 score on at least one of the two test sets. For the categories subsumed under *Rest*, the model was not able to learn how to correctly identify these categories.

Subtask C2 exhibits a similar distribution of the true labels, with the *Aspect+Sentiment* category `Allgemein:neutral` as majority class. Over 50% of the true labels belong to this class. Table 16 shows that only 12 out of 60 labels can be detected by the model (see Table 16).

	test_{syn}		test_{dia}	
Aspect+Sentiment Category	**Score**	**Support**	**Score**	**Support**
Allgemein:neutral	0.804	1,108	0.832	913
Sonstige Unregelmäßigkeiten:negative	0.782	221	0.793	159
Zugfahrt:negative	0.645	197	0.725	149
Sicherheit:negative	0.640	78	0.585	39
Allgemein:negative	0.582	258	0.333	80
Atmosphäre:negative	0.569	126	0.447	39
Connectivity:negative	0.400	20	0.291	46
Ticketkauf:negative	0.364	42	0.298	34
Auslastung und Platzangebot:negative	0.350	31	0.211	17
Allgemein:positive	0.214	41	0.690	33
Zugfahrt:positive	0.154	34	0	34
Service und Kundenbetreuung:negative	0.146	36	0.174	21
Rest	*0*	*343*	*0*	*180*

Table 16: Micro-averaged F1 scores and support by *Aspect+Sentiment* category (Subtask C2). 48 categories are summarized in *Rest* and show each a score of 0.

All the aspect categories displayed in Table 16 are also visible in Table 15 and most of them have negative sentiment. `Allgemein:neutral` and `Sonstige Unregelmäßigkeiten:negative` show the highest scores. Again, we assume that here, 48 categories could not be identified due to data sparsity. However, having this in mind, the model achieves a relatively high overall performance for both, Subtask C1 and C2 (cf. Tab. 9 and Tab. 10). This is mainly owed to the high score of the majority classes `Allgemein` and `Allgemein:neutral`, respectively, because the micro F1 score puts a lot of weight on majority classes. It might be interesting whether the classification of the rare categories can be improved by balancing the data. We experimented with removing general categories such as `Allgemein`, `Allgemein:neutral` or documents with sentiment `neutral` since these are usually less interesting for a company. We observe a large drop in the overall F1 score which is attributed to the absence of the strong majority class and the resulting data loss. Indeed, the classification for some single categories could be improved, but the rare categories could still not be identified by the model.

B Detailed results (per category) for Subtask D

Similar as for Subtask C, the results for the best model are investigated in more detail. Table 17 gives the detailed classification report for the uncased German BERT-BASE model with CRF layer on Subtask D1. Only entities that were correctly detected at least once are displayed. The table is sorted by the score on the synchronic test set. The classification report for Subtask D2 is displayed analogously in Table 18.

	test_{syn}		test_{dia}	
Category	**Score**	**Support**	**Score**	**Support**
Zugfahrt:negative	0.702	622	0.729	495
Sonstige Unregelmäßigkeiten:negative	0.681	693	0.581	484
Sicherheit:negative	0.604	337	0.457	122
Connectivity:negative	0.598	56	0.620	109
Barrierefreiheit:negative	0.595	14	0	3
Auslastung und Platzangebot:negative	0.579	66	0.447	31
Connectivity:positive	0.571	26	0.555	60
Allgemein:negative	0.545	807	0.343	139
Atmosphäre:negative	0.500	403	0.337	164
Ticketkauf:negative	0.383	96	0.583	74
Ticketkauf:positive	0.368	59	0	13
Komfort und Ausstattung:negative	0.357	24	0	16
Atmosphäre:neutral	0.348	40	0.111	14
Service und Kundenbetreuung:negative	0.323	74	0.286	31
Informationen:negative	0.301	68	0.505	46
Zugfahrt:positive	0.276	62	0.343	83
DB App und Website:negative	0.232	39	0.375	33
DB App und Website:neutral	0.188	23	0	11
Sonstige Unregelmäßigkeiten:neutral	0.179	13	0.222	2
Allgemein:positive	0.157	86	0.586	92
Service und Kundenbetreuung:positive	0.115	23	0	5
Atmosphäre:positive	0.105	26	0	15
Ticketkauf:neutral	0.040	144	0.222	25
Connectivity:neutral	0	11	0.211	15
Toiletten:negative	0	15	0.160	23
Rest	*0*	*355*	*0*	*115*

Table 17: Micro-averaged F1 scores and support by *Aspect+Sentiment* entity with exact match (Subtask D1). 35 categories are summarized in *Rest*, each of them exhibiting a score of 0.

For Subtask D1, the model returns a positive score on 25 entity categories on at least one of the two test sets. The category `Zugfahrt:negative` can be classified best on both test sets, followed by `Sonstige Unregelmäßigkeiten:negative` and `Sicherheit:negative` for the synchronic test set and by `Connectivity:negative` and `Allgemein:positive` for the diachronic set. Visibly, the scores between the two test sets differ more here than in the classification report of the previous task.

The report for the overlapping match (cf. Tab. 18) shows slightly better results on some categories

Category	test_{syn}		test_{dia}	
	Score	**Support**	**Score**	**Support**
Zugfahrt:negative	0.708	622	0.739	495
Sonstige Unregelmäßigkeiten:negative	0.697	693	0.617	484
Sicherheit:negative	0.607	337	0.475	122
Connectivity:negative	0.598	56	0.620	109
Barrierefreiheit:negative	0.595	14	0	3
Auslastung und Platzangebot:negative	0.579	66	0.447	31
Connectivity:positive	0.571	26	0.555	60
Allgemein:negative	0.561	807	0.363	139
Atmosphäre:negative	0.505	403	0.358	164
Ticketkauf:negative	0.383	96	0.583	74
Ticketkauf:positive	0.368	59	0	13
Komfort und Ausstattung:negative	0.357	24	0	16
Atmosphäre:neutral	0.348	40	0.111	14
Service und Kundenbetreuung:negative	0.323	74	0.286	31
Informationen:negative	0.301	68	0.505	46
Zugfahrt:positive	0.276	62	0.343	83
DB App und Website:negative	0.261	39	0.406	33
DB App und Website:neutral	0.188	23	0	11
Sonstige Unregelmäßigkeiten:neutral	0.179	13	0.222	2
Allgemein:positive	0.157	86	0.586	92
Service und Kundenbetreuung:positive	0.115	23	0	5
Atmosphäre:positive	0.105	26	0	15
Ticketkauf:neutral	0.040	144	0.222	25
Connectivity:neutral	0	11	0.211	15
Toiletten:negative	0	15	0.160	23
Rest	*0*	*355*	*0*	*112*

Table 18: Micro-averaged F1 scores and support by *Aspect+Sentiment* entity with overlapping match (Subtask D2). 35 categories are summarized in *Rest* and show each a score of 0.

than for the exact match. The third-best score on the diachronic test data is now `Sonstige Unregelmäßigkeiten:negative`. Besides this, the top three categories per test set remain the same.

Apart from the fact that this is a different kind of task than before, one can notice that even though the overall micro F1 scores are lower for Subtask D than for Subtask C, the model manages to successfully identify a larger variety of categories, i.e. it achieves a positive score for more categories. This is probably due to the more balanced data for Subtask D than for Subtask C2, resulting in a lower overall score and mostly higher scores per category.