Human Language Comprehension in Aspect Phrase Extraction with Importance Weighting

Joschka Kersting[1](✉) and Michaela Geierhos[2]

[1] Paderborn University, Warburger Str. 100, Paderborn, Germany
joschka.kersting@uni-paderborn.de
[2] Bundeswehr University Munich, Research Institute CODE,
Carl-Wery-Straße 22, Munich, Germany
michaela.geierhos@unibw.de

Abstract. In this study, we describe a text processing pipeline that transforms user-generated text into structured data. To do this, we train neural and transformer-based models for aspect-based sentiment analysis. As most research deals with explicit aspects from product or service data, we extract and classify implicit and explicit aspect phrases from German-language physician review texts. Patients often rate on the basis of perceived friendliness or competence. The vocabulary is difficult, the topic sensitive, and the data user-generated. The aspect phrases come with various wordings using insertions and are not noun-based, which makes the presented case equally relevant and reality-based. To find complex, indirect aspect phrases, up-to-date deep learning approaches must be combined with supervised training data. We describe three aspect phrase datasets, one of them new, as well as a newly annotated aspect polarity dataset. Alongside this, we build an algorithm to rate the aspect phrase importance. All in all, we train eight transformers on the new raw data domain, compare 54 neural aspect extraction models and, based on this, create eight aspect polarity models for our pipeline. These models are evaluated by using Precision, Recall, and F-Score measures. Finally, we evaluate our aspect phrase importance measure algorithm.

Keywords: Aspect-based Sentiment Analysis · Aspect Polarity Model

1 Introduction

Sentiment Analysis (SA) is the process of automatically identifying and categorizing opinions expressed in a text, especially to determine whether the author's attitude towards a particular topic, product, etc. is positive, negative, or neutral. There are different approaches: Aspect-based Sentiment Analysis (ABSA) aims to identify expressed opinions about aspects of services or products. SA at the document or sentence-level does not address conflicting feelings, feelings expressed towards different aspects, and the granularity of human language in general. ABSA is therefore an alternative method that allows fine-grained analysis, automatically extracting individual aspects and their scores. The development of ABSA has led to various studies and shared tasks [10,15,17].

Previous approaches have often failed to pursue a human-centered method by considering implicit or indirect mentions of aspects and ratings, as the studies focused on domains with common vocabulary in which nouns often explicitly indicate an aspect. These approaches treat nouns and noun phrases as the representation of aspects, or they consider them as sufficient [2,16,17], due to the commonly used review domains: Most reviews are written for products [6,17] or services [17]. Despite the available domains and their particularities, it is necessary to understand how users rate and why they do so in order to use the reviews available on the Internet. Hence, ABSA is a promising research topic.

However, to find complex indirect aspect phrases, current deep learning approaches need to be combined with supervised training data. Due to implicit mentions and the use of longer phrases, keyword spotting is not an option.

Example 1 (Sentence from Physician Review). "*Dr. Stallmann has* ***never once looked me in the eye****, but he* ***accurately described*** *the options and he also seemed to* ***know, and this is important to me, what he is doing****.*"

In this example, some ratings are given for the aspects "*friendliness*", "*explanation*", and "*competence*" (printed in bold). As shown, these aspect phrases are rather complex, using insertions and different wording. They are not covered by previous machine learning models targeting ABSA, partly because they often appear in a different form and expression. For example, they do not directly mention that a physician has a "*good friendliness*" because this is a rather uncommon style in written physician reviews or everyday conversations.

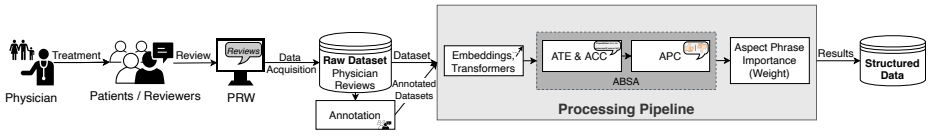

Fig. 1. Processing pipeline to structure and analyze unstructured text data.

Physician reviews can be found in various languages on physician review websites (PRWs) such as Ratemds[1] in English, or Jameda[2] in German. For example, users can rate a physician by assigning scores for rating classes and by writing a textual evaluation. Quantitative scores can be assigned to classes, such as the "*competence of the physician*". Assessed health services are strongly associated with trust; they are sensitive and personal.

As shown in Figure 1, we build a fully functional text processing pipeline that takes raw text as input, vectorizes it, then extracts aspect phrases to finally add polarity scores. That is, we classify the extracted aspect phrases to determine whether the author evaluates a characteristic of the doctor negatively or

[1] `https://ratemds.com`, accessed: 2020-12-17.
[2] `https://jameda.de`, accessed: 2020-12-17.

positively. Then, our pipeline determines which of the phrase(s) has an increased importance weight. Overall, this implements a complete cycle from unstructured user-generated text to structured data.

Compared to related literature and our previous work [8,9,10], we here present a new aspect phrase dataset dealing with a physician practice team, an aspect polarity dataset and supervised learning algorithms, and a method for measuring aspect phrase's weight of importance. Furthermore, we train and test a number of machine learning approaches, including numerous domain-specific transformer models, and build a processing pipeline that converts unstructured physician reviews into structured data (cf. Figure 1).

2 State of Research

Physician reviews are not like the standard data used for ABSA research. There is no standard service in the healthcare sector, as treatments from physicians and other healthcare providers heavily depend on the practitioner and the patient.

There are three core tasks in ABSA research: ATE, ACC, and APC (Aspect Term Extraction, Category Classification, and Polarity Classification) [2]. ATE means finding aspects in texts. This is important for performing subsequent steps, but as we discussed in a related study [8], much previous work relies on nouns, seed words, etc. For example, Pontiki et al. [18] write that "[a]n opinion target expression [...] is an explicit reference (mention) to the reviewed entity [...]. This reference can be a named entity, a common noun or a multi-word term". In their annotated datasets, they used common product or service domains (e.g., hotels) and achieved evaluation scores for ATE and ACC of about 50%. However, most studies use the data of the shared task by Pontiki et al. [17] or its predecessors [19,20], as survey studies show [24,25].

Previous ABSA approaches have neglected human-like language understanding without artificial constraints, thus limiting their methods and data domains, as we have previously described [8,9,10]. Therfore, most study designs cannot be applied to physician review data.

Recent approaches to ABSA use neural networks and deep learning methods, as surveys show [15,24,25]. They differ not only in the applied data (mostly from shared tasks [17,20,19]), but also in neural network architectures and do not perform ATE. Thus, they rather perform SA at the sentence or document level. However, it is clear that transformers such as BERT [7] have improved vector representations for use in other algorithms, while they can also be fine-tuned for downstream tasks such as tagging words and classifying texts. For example, our previous work [10] successfully applied transformer models to PRW data, but more traditional methods for language modeling such as FastText [1] are still competitive for physician reviews, as we have shown [8]. All in all, previous research has not explored and made the contributions described in Section 1, although researchers such as DeClercq et al. [6] built an ABSA pipeline for Dutch social media data on retail, banking, and human resources. Nevertheless, the domain, approach, and data are entirely different. Based on our previous

remarks and current studies, there is no alternative to supervised deep learning for ABSA with human-like aspect understanding [15,24,25].

Several datasets [6,14,17,20,23] have been created for ABSA so far. Here it can be seen that the polarity scales are usually threefold or twofold, i.e., they use either the positive, negative and neutral classes or only the first two. The importance weight of an aspect phrase is difficult to determine because aspect phrases are not very heterogeneous. There is no uniform vocabulary; so it is not sufficient to use rule-based or list-based approaches that determine importance with the infrequent preference of a word from a predefined list. In German, longer off-topic insertions are also common ("*He took a lot, and I want to add this after I clarify how I encountered my friend in the office, of time ...*") and such cases are numerous, making it difficult to adapt ideas from the literature. Moreover, it is not known which rating scale should be applied here. However, from the ABSA datasets and their polarity scales, it can be inferred that a rather simple scale with two or three values is applicable. However, it is obvious that users assign different weights to aspect classes [13].

One of the many various approaches is to calculate the semantic information value, e.g., using the entropy or by measuring the cosine similarity between the embedding vectors of a phrase and the corresponding annotations for the class. However, this misses the point, because we have neural vectors with embedded semantics but do not see information scores or vector similarity as measures for importance. Another approach might be sentiment-intensity ranking: A study [21] uses words with the same meaning and ranks polar words by intensity, e.g., "*pleased, exhilarated*". Such approaches do not fit because we do not have a traditional separation into sentiment and aspect words, and lexicon-based approaches are not flexible enough. Our phrases mostly cover both at once, e.g. in just one word like "*friendly*", which indicates both friendliness as an aspect and a positive evaluation. The same applies to longer phrases (cf. Example 1). Therefore, a promising approach is to calculate the normalized frequency of aspect classes in the respective dataset. This provides a unique measure that also allows a comparison of the classes. A second possibility is to analyze linguistic structures which indicate a higher importance. Since adjectives are common in our data, intensified adjectives or additional adverbs could be a solid way to identifying important aspect phrases from physician reviews.

3 Data and Annotation Process

In our data and annotation process (aspect and polarity data), some of the data are based on our previous works which contain additional information, especially for the `fkza` and `bavkbeg` datasets [8,9,10].

Raw data were collected from three German-language PRWs[3] between March and July 2018 by using a spider to crawl all review and physician pages to reach a total of 400,000 physicians and over 2,000,000 review texts. The scales are

[3] Jameda: `https://jameda.de`; Docfinder: `https://docfinder.at`; Medicosearch: `https://medicosearch.ch`; accessed 2021-01-11.

based on the German/Austrian school grade system as well as star ratings. The number of quantitative rating classes varies greatly among the PRWs [5,8,9,10]. To train algorithms that extract and classify aspect phrases, we needed to find classes that could be annotated.

We considered all available quantitative rating classes from the three crawled websites and qualitatively merged classes, e.g., those related to the team's "*competence*" or to the "*waiting time for [an] appointment*", etc. The semantic merging of quantitative classes resulted in a larger set of rating classes. For ATE and ACC, we use three datasets in this study. The first two, `fkza` and `bavkbeg`, were taken from our previous studies [8,9,10], which present the dataset in detail and provide a tableau of examples. In short, `fkza` is an acronym of the German names of the classes translated into English as "*friendliness*", "*competence*", "*time taken*", and "*explanation*". The `bavkbeg` dataset covers the classes of "*treatment*", "*alternative healing methods*", "*relationship of trust*", "*child-friendliness*", "*care/commitment*", and "*overall/recommendation*". `Fkza` and `bavkbeg` apply to the physician as an aspect target. `Bavkbeg` has an overall rating class that applies equally to the physician, the practice, and the team. These three are the available aspect targets in the data. Like many systems, we perform ATE and ACC together [24], which is due to their mutual influence.

The third and newly annotated dataset is called `bfkt`, which aims at the physician's team as an aspect target. Since the target is different, some of the classes are similar to those in the `fkza` package. However, for human annotators identifying the aspect target clearly on the basis of the text is not an issue. To avoid annotation conflicts, certain rules can be established. The classes of `bfkt` are these: "*care/commitment*", "*friendliness*", "*competence*", and "*accessibility by telephone*"[4].

- "*Care/commitment*" refers to whether the practice team is (further) involved or interested in the patient's care and treatment: "***Such a demotivated** assistant!*"
- "*Friendliness*" deals with the friendliness, as in the package `fkza`, but aims at the team: "*Due to their **very nice manner**, there was no doubt about the team at any time.*"
- "*Competence*" describes the patient's perception of the team's expertise: "*The staff at the reception makes an **overstrained impression**.*"
- "*Accessibility by telephone*" indicates how easy it is to reach the team: "*You have to **try several times before you get someone on the phone**.*"

Since the PRWs focus on reviews of "doctors", this may explain why there are far fewer aspect phrases for `bfkt`. The annotation process began with one person annotating the package, while we held ongoing discussions and reviews among a team of (computational) linguists. Active learning was performed once for all packages before annotations began, consistent with previous work [10].

[4] Translated from German, with the team as the aspect target: "*Betreuung/Engagement*", "*Freundlichkeit*", "*Kompetenz*", and "*Telefonerreichbarkeit*".

Here, the goal was to find sentences that generally contain an evaluative statement. For this purpose, we used a neural network classifier. We then annotated several thousand sentences for `bfkt`. Since most of the sentences did not contain relevant statements, we again trained a sentence-level classifier using the existing annotations, ordering the sentences in the resulting file so that they contained at least one predicted class per line. This multi-label, multi-class classification problem at the sentence level helped us save time, which is consistent with what was done for `bavkbeg` [10]. Of more than 15,000 sentences in the `bfkt` dataset, about half contain an evaluative statement, and it was possible to annotate more than one mentioned aspect in a sentence. Most sentences tend to be short, and users generally write as they speak, indicating rating aspects in longer phrases like: "*It doesn't matter how many times you try, you will* ***never catch any of them over the wire!***" During annotation, we also formulated rules and examples as guidelines for the annotators to follow, such as that phrases should be as short as possible but contain all important information, preferably without punctuation.

The annotation task was rather difficult due to the data and the direct and indirect long phrases it contained. We computed an IAA based on the tagged words, assigning a tag to each word indicating its class. All non-annotated words were tagged "no class". We used the annotations of the first annotator and randomly selected about 330–360 sentences (about 3% of `fkza` [8], `bavkbeg` [10]). The second annotator and another person then performed new annotations for the agreement. The values of all IAAs are shown in Table 1. All Cohen's Kappa [3] values can be considered as "substantial" agreement (0.61–0.80). One pair of annotators, "B&J", achieved an "almost perfect" agreement [12]. Krippendorff's Alpha [11] can be considered good as it leans to 1.0. However, the values are worse than for `fkza` and `bavkbeg` [10] with 0.654 (R&B) to 0.722 (R&B) for `bfkt` and `fkza`.

Table 1. Inter-annotator agreements for all used datasets (fkza & bavkbeg: [8,10]).[5]

	Dataset fkza			**Dataset bavkbeg**			**Dataset bfkt**			**Polarity Dataset**		
	R&B	R&J	B&J	R&B	R&J	B&J	R&B	R&J	B&J	R&M	R&J	M&J
CK	0.722	0.857	0.730	0.731	0.719	0.710	0.654	0.673	0.806	0.917	0.923	0.918
KA		0.771			0.720			0.711			0.919	

The sentiment polarity annotations were conducted differently. As mentioned above, a distinction between aspect phrases and sentiment words is not possible. Since the aspect phrase and class annotations were difficult and several tens of thousands of sentences had to be annotated, the steps were separated and the polarity step was conducted later. For the polarity annotation, we randomly selected sentences containing aspect phrases from the datasets and deleted erroneous annotations from the file. We also included two newly annotated aspect

[5] CK = Cohen's Kappa; KA = Krippendorff's Alpha.

datasets that were not yet complete, so we had the annotator also check the aspect phrases for errors. For each phrase, we needed to assign a positive or negative sentiment polarity. At first, we tried finer scales by using a neutral value. After testing and discussions we discovered that neither a finer granularity nor a neutral label are appropriate for our data, as the phrases do not have patterns that reveal finer nuances, and neutral evaluative statements are almost nonexistent in physician reviews. As for nuances such as a "highly positve", "midly positive" or "normal positive" polarity, it is difficult to distinguish between the phrases such as: "*very friendly*", "*expressively friendly*", "*always very friendly*" or "*always friendly as every time except once*", "*indeed he was friendly today*". These phrases show that nuances are hard to systematize, so adequate and consistent annotations for scales with increments are not possible.

After deleting the sentences that contained mistakes and the ones in which we experimentally annotated potentially neutral values, we have over 9,300 sentences with polarity annotations in general. For quality reasons, we computed the IAA shown in Table 1. As shown, the results are quite good. Since the Cohen's Kappa values are all above 0.90, the agreement is almost perfect [12]. However, this is not surprising for a human annotation of a binary phrase-sentiment polarity classification task. Krippendorff's Alpha can be considered as very good, with a value of 0.919, which is quite close to 1.0.

4 Method and Results

As our previous work has shown, supervised neural learning is the most promising path for ABSA in a serious data domain such as ours [8,10]. However, it was also shown that transformers perform well in ATE and ACC, especially when pre-trained on raw PRW data. Nevertheless, more traditional solutions such as FastText provided the best results, while a domain-trained BERT [7] performed slightly better or almost as well [10]. Due to this information, we want to further investigate using transformer models for our case, so we searched Huggingface for pre-trained transformer models for German.

For our experimental setup, we used IO tags (Inside, Outside) for ATE and ACC [8], e.g., "*I-friendliness_T*". This step is critical because it is the most challenging and it starts the pipeline, so the other steps depend on the results (cf. Figure 1). Therefore, we tested a large number of transformers and show these results in Table 2. First, we domain-trained the existing transformer models for German as well as the multilingual XLM-RoBERTa [4]. The domain-trained models are marked with a "+". As tests have shown, we do not have enough PRW data to train a transformer from scratch (no useful results), so we tested pre-trained transformers and domain-trained these further. In addition to fine-tuning, we built our own neural networks that used the word vectors generated by the transformer as input. We used XLM-RoBERTa for this purpose because the loss in domain training was extremely small. The loss was about 0.37 after 4 epochs compared to about 1.1–1.3 after 10 epochs for most German language models such as BERT (bert-base cased). This was different for Electra (a loss

Table 2. Results[6] for the extraction and classification of aspect phrases (ATE, ACC) using broadly pre-trained and domain-trained ("+") transformers.

	bfkt			**bavkbeg**			**fkza**		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
xlm-roberta-base+	0.81	0.70	0.75	0.83	0.82	**0.82**	0.86	0.80	**0.83**
⌞ biLSTM-CRF+	0.78	0.76	**0.77**	0.81	0.81	0.81	0.86	0.79	**0.83**
⌞ biLSTM-Attention+	0.82	0.70	0.75	0.78	0.81	0.79	0.83	0.80	0.82
xlm-roberta-base	0.80	0.70	0.74	0.83	0.81	0.81	0.85	0.80	0.82
MedBERT+	0.81	0.70	0.75	0.84	0.82	**0.82**	0.86	0.80	**0.83**
MedBERT	0.80	0.68	0.73	0.83	0.79	0.80	0.86	0.78	0.82
electra-base uncased+	0.16	0.20	0.18	0.10	0.14	0.12	0.15	0.20	0.17
electra-base uncased	0.79	0.70	0.74	0.82	0.81	0.81	0.85	0.80	0.82
distilbert-base cased+	0.80	0.69	0.74	0.83	0.80	0.80	0.85	0.80	0.82
distilbert-base cased	0.78	0.67	0.72	0.81	0.78	0.79	0.84	0.78	0.81
dbmdz bert-base uncased+	0.81	0.72	0.76	0.82	0.82	0.81	0.86	0.80	**0.83**
dbmdz bert-base uncased	0.80	0.70	0.74	0.83	0.80	0.81	0.86	0.80	0.82
dbmdz bert-base cased+	0.80	0.70	0.74	0.83	0.81	0.81	0.86	0.81	**0.83**
dbmdz bert-base cased	0.79	0.70	0.74	0.83	0.79	0.80	0.85	0.80	0.82
bert-base cased+	0.81	0.71	0.75	0.83	0.81	**0.82**	0.86	0.81	**0.83**
bert-base cased	0.79	0.68	0.73	0.82	0.80	0.80	0.86	0.78	0.82
FastText biLSTM-CRF+	0.77	0.70	0.73	0.80	0.76	0.78	0.83	0.79	0.81
FastText biLSTM-Attention+	0.74	0.71	0.72	0.81	0.74	0.77	0.82	0.77	0.79

over 6.7). The parameters were tuned before the final runs. We used a train-test split of 90%/10% of the sentences extracted from the raw data (cf. Section 3).

XLM-RoBERTa achieves the best scores for the datasets `bavkbeg` (F1: 0.82) and `fkza` (F1: 0.82) with transformer fine-tuning and for `bfkt` (F1: 0.77) with a biLSTM-CRF model [10]. The train-test split was 80%/20% for transformers (epochs: 10) in most cases, and 90%/10% for the other neural networks (epochs: 6) after tuning the parameters. FastText was trained uncased, as we are using error-prone user-generated text data with medical terms. A general advantage cannot be seen (in contrast to previous work [8]), since cased transformers also perform well. This may be because the transformer approach computes embeddings ad-hoc, based on context, while FastText computes a fixed table in which each string is given a vector.

The other models in Table 2 that are not explicitly marked as (un-)cased are cased. While XLM-RoBERTa is well documented, which is another reason for its use, other German transformers were not. For MedBERT, a related paper was published after we had used it [22]. At least it was obvious that MedBERT was trained on data related to the medical domain. We use Precision, Recall and F1 as scores because accuracy is prone to error considering our high class imbalance. Most words in a sentence are not tagged as a specific class, but as "O" such as

[6] P = Precision, R = Recall, F1 = F1-score; all pre-trained transformer models are in German and can be found by their names on `https://huggingface.co/models`, accessed 2020-12-28. BiLSTM-CRF and Attention models are based on [10].

outside a phrase. Therefore, the accuracy values were all the same but did not reveal differences in the models. To reduce the imbalance and because the results were better, we used only the sentences containing aspect phrases.

The second step in the pipeline (cf. Figure 1) involves sentiment polarity classification. We used the transformer architectures that performed best in the previous step, so only domain-tuned transformers and FastText embeddings were used. The results are shown in Table 3. Again, our own neural network performs best with an F1-score of 0.96, strengthened by XLM-RoBERTa embeddings. Again, we obtained the best results with a train-test split of 80%/20% for the transformer fine-tuning (epochs: 4) and 90%/10% for the other neural networks (epochs: 6). The task was performed as a binary text classification. We used two input layers that received the corresponding aspect phrase and its context. Our goal was to classify the aspect phrase; the context was represented by the sentence from which the phrase was extracted. The multilingual XLM-RoBERTa outperformed the transformers trained specifically for German.

Table 3. Sentiment polarity classification results.

Model	**P**	**R**	**F1**
xlm-roberta-base+	0.93	0.95	0.94
⌞ biLSTM+	0.97	0.95	**0.96**
⌞ CNN+	0.92	0.97	0.94
MedBERT+	0.90	0.95	0.92
dbmdz bert-base uncased+	0.92	0.93	0.92
bert-base cased+	0.92	0.91	0.91
FastText biLSTM+	0.92	0.95	0.93
FastText CNN+	0.91	0.93	0.92

The third step of the pipeline deals with measuring the importance of aspect phrases. After studying the available methods in Section 2, we concluded that three approaches are promising: First, importance can be derived from a normalized frequency of each aspect class. On this basis, the most frequent aspect classes are ranked as most important. Second, as suggested in Section 2, we set up a linguistic approach that uses part-of-speech (POS) tagging to identify adverbs and adjective superlatives. We suggested that the presence of an adverb increases importance, which is often the case: "*They were **very compassionate**.*", instead of just "*compassionate*", "*friendly*", etc. This also applies to longer phrases. The use of superlatives also shows a high importance: "*The woman at the front desk is the **worst listener** I have ever seen!*" We also included German indefinite pronouns: "*They had **many friendly** words.*" Third, we combined the two approaches and suggested that whenever either one of both suggests a high importance, this should be respected as an outcome. The exploration of possible methods also led to an investigation of which scale is appropriate. Consistent with our observations regarding the polarity scale (cf.

Section 3), only a binary classification into higher and lower (normal) importance is possible. Finer gradations are not possible.

Table 4. Accuracy-agreement of humans with the importance weighting method.

Person	Statistic	Linguistic	Combined
J	0.51	**0.82**	0.62
R	0.54	**0.87**	0.65

To test our aspect-phrase-importance weighting, we had two human annotators label approximately 340 random phrases with higher or lower relative importance. Both knew the domain and were introduced to the task. During the initial annotations, they were allowed to see the results of the automatic approach. The evaluation results in Table 4 show the accuracy of the annotations with the automatic methods. As can be seen, the linguistic approach has the highest agreement: 0.82 for annotator J and 0.87 for annotator R. The high scores indicate the quality of the approach. The disadvantages of this method are that POS tagging sometimes fails, especially when distinguishing between adverbs and adjectives. Furthermore, POS tagging may fail for longer phrases and due to insertions that may contain superlatives that are not relevant to the corresponding aspect. The evaluation results may have a limited value because annotators may be biased on their linguistic knowledge or knowledge of the used methods.

5 Conclusion

We showed three datasets for ATE and ACC, one is new and deals with the team of a physician's office, and two deal with the physician as the aspect target. We also presented a new dataset for APC and calculated IAAs for all datasets, achieving good scores (cf. Table 1). To build a pipeline that converts user-generated, unstructured physician reviews into structured data, we trained a set of deep-learning models and developed a method for measuring the importance of aspect phrases. All of these were evaluated in detail in Tables 2–4 and obtained good results. We tested 54 models for ATE and ACC, and another eight for APC. XLM-RoBERTa in its basic version emerged as the best model among all those tested. It is a multilingual model that also outperformed German-only models, which we consider a major finding, especially as we applied the models to long, complex, and user-generated phrases. Furthermore, due to resource constraints, we trained the base version of this pre-trained transformer instead of the large version. This large version is a promising tool for future experiments.

In all training steps, we applied human language comprehension to extract information in a human-like manner, conducting broad research by using and comparing a wide variety of neural models. In the future, we can build on these experiments to extract other aspect classes from data such as the accessibility

and the opening hours. We see potential applications in domains with implicit aspect phrases. Since XLM-RoBERTa is capable of working with multiple languages, we plan to test our fine-tuned models on English physician reviews. The binary scales discussed and used to measure sentiment polarity and aspect phrase importance emerged as the only feasible solutions based on the data. Annotating the data based on context allows us and our models to treat irony accordingly. Parts of the pipeline methods presented here are in further development for a related study dealing with possible analyses based on it.

Acknowledgments. This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center On-The-Fly Computing (SFB 901). We thank F. S. Bäumer, M. Cordes, and R. R. Mülfarth for their assistance with the data collection.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the ACL **5**, 135–146 (2017)
2. Chinsha, T.C., Shibily, J.: A Syntactic Approach for Aspect Based Opinion Mining. In: Proceedings of the 9th IEEE International Conference on Semantic Computing. pp. 24–31. IEEE (2015)
3. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the ACL. pp. 8440–8451. ACL, Online (2020)
5. Cordes, M.: Wie bewerten die anderen? Eine übergreifende Analyse von Arztbewertungsportalen in Europa. Master's thesis, Paderborn University (2018)
6. De Clercq, O., Lefever, E., Jacobs, G., Carpels, T., Hoste, V.: Towards an Integrated Pipeline for Aspect-based Sentiment Analysis in Various Domains. In: Proceedings of the 8th ACL Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 136–142. ACL (2017)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. ACL, Minneapolis, Minnesota, USA (2019)
8. Kersting, J., Geierhos, M.: Aspect Phrase Extraction in Sentiment Analysis with Deep Learning. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence: Special Session on Natural Language Processing in Artificial Intelligence. pp. 391–400. SCITEPRESS (2020)
9. Kersting, J., Geierhos, M.: Neural Learning for Aspect Phrase Extraction and Classification in Sentiment Analysis. In: Proceedings of the 33rd International FLAIRS. pp. 282–285. AAAI (2020)

10. Kersting, J., Geierhos, M.: Towards Aspect Extraction and Classification for Opinion Mining with Deep Sequence Networks. In: Loukanova, R. (ed.) Natural Language Processing in Artificial Intelligence – NLPinAI 2020, Studies in Computational Intelligence (SCI), vol. 939, pp. 163–189. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-63787-3_6
11. Krippendorff, K.: Computing Krippendorff's Alpha-Reliability. Tech. Rep. 1-25-2011, University of Pennsylvania (2011)
12. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics **33**(1), 159–174 (1977)
13. Liu, Y., Bi, J.W., Fan, Z.P.: Ranking Products through Online Reviews: A Method Based on Sentiment Analysis Technique and Intuitionistic Fuzzy Set Theory. Information Fusion **36**, 149–161 (2017)
14. López, A., Detz, A., Ratanawongsa, N., Sarkar, U.: What Patients Say About Their Doctors Online: A Qualitative Content Analysis. Journal of General Internal Medicine **27**(6), 685–692 (2012)
15. Nazir, A., Rao, Y., Wu, L., Sun, L.: Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey. IEEE Transactions on Affective Computing pp. 1–1 (2020). https://doi.org/10.1109/TAFFC.2020.2970399
16. Nguyen, T.H., Shirai, K.: PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2509–2514. ACL (2015)
17. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation. pp. 19–30. ACL (2016)
18. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2016 Task 5: Aspect Based Sentiment Analysis (ABSA-16) Annotation Guidelines (2016)
19. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation. pp. 27–35. ACL (2014)
20. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation. pp. 486–495. ACL (2015)
21. Sharma, R., Somani, A., Kumar, L., Bhattacharyya, P.: Sentiment Intensity Ranking among Adjectives Using Sentiment Bearing Word Embeddings. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 547–552. ACL (2017)
22. Shrestha, M.: Development of a Language Model for Medical Domain. Master's thesis, Rhine-Waal University of Applied Sciences (2021)
23. Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., Biemann, C.: 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. pp. 1–12. Springer (2017)
24. Zhang, L., Wang, S., Liu, B.: Deep Learning for Sentiment Analysis: A Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), 1–25 (2018)
25. Zhou, J., Huang, J.X., Chen, Q., Hu, Q.V., Wang, T., He, L.: Deep Learning for Aspect-Level Sentiment Classification: Survey, Vision, and Challenges. IEEE Access **7**, 78454–78483 (2019)