Well-Being in Plastic Surgery: Deep Learning Reveals Patients' Evaluations

Joschka Kersting[1] and Michaela Geierhos[2]
[1]*Paderborn University, Warburger Str. 100, Paderborn, Germany*
[2]*Bundeswehr University Munich, Research Institute CODE, Carl-Wery-Straße 22, Munich, Germany*

Keywords: Aspect-based Sentiment Analysis, Information Extraction, Deep Learning, Transformer Applications.

Abstract: This study deals with aspect-based sentiment analysis, the correlation of extracted aspects and their sentiment polarities with metadata. There are millions of review texts on the Internet that cannot be analyzed and thus people cannot benefit from the contained information. While most research so far has focused on explicit aspects from product or service data (e.g., hotels), we extract and classify implicit and explicit aspect phrases from German-language physician review texts. We annotated aspect phrases that indicate ratings about the doctor's practice, such as waiting time or general perceived well-being conveyed by all staff members of a practice. We also apply a sentiment polarity classifier. While we compare several traditional and transformer networks, we apply the best model, the multilingual XLM-RoBERTa, to a dedicated German-language dataset dealing with plastic surgeons. We choose plastic surgery as sample domain because it is especially sensitive with its relation to a person's self-image and felt acceptance. In addition to standard evaluation measures such as Precision, Recall, and F1-Score, we correlate our results with metadata from physician review websites, such as a physician's gender. We figure out several correlations and present methods for analyzing unstructured review texts to enable service improvements in healthcare.

1 INTRODUCTION

Handling unstructured data such as text has made significant progress so far. Among others, transformers (Vaswani et al., 2017) have enabled great improvements when it comes to word representations, information extraction, or text classification (Devlin et al., 2019). One field of study that has also benefited from recent developments in natural language processing is called Aspect-based Sentiment Analysis (ABSA), which aims to identify fine-grained evaluations in texts about products and services. ABSA extracts words or phrases from texts and classifies them according to the rated aspects, their targets, and their polarities. Sentiment polarity describes whether something is talked about positively or negatively (Pontiki et al., 2016a; He et al., 2019). Due to the nature of natural language, the ABSA task is a challenging one that has led to numerous studies (Li et al., 2019; Kersting and Geierhos, 2020a; Kersting and Geierhos, 2020b; Kersting and Geierhos, 2021b), surveys (Do et al., 2019; Zhou et al., 2019; Nazir et al., 2020), and shared tasks (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016a; Wojatzki et al., 2017). In addition to ABSA, there is also sentiment analysis at the document and sentence level. Both assume one opinion per text or sentence and therefore do not deal with ambivalence, sentiments on different aspects and the granularity of human language in general.

Our goal is to investigate patient reviews for plastic surgeons on German-language Physician Review Websites (PRWs). The reasons for the use of physician reviews lie not only in the challenging domain but more especially in the wording used, which mostly consists of longer phrases and insertions. Unlike other studies, we deal with implicit and indirect aspect mentions. Scholars mostly used exclusively commercial reviews (Zhou et al., 2019; Nazir et al., 2020) for ABSA research, in which nouns often explicitly mention an aspect by its name or a synonym. Many studies either take nouns and noun phrases for aspect representation for granted or at least as sufficient (Chinsha and Shibily, 2015; Nguyen and Shirai, 2015; Pontiki et al., 2016a). The reason for this may lie in the review domains commonly used in research. As mentioned, these are product and service reviews (Pontiki et al., 2016a; De Clercq et al., 2017). For PRW data, for example, the phrase "*he didn't look*

Kersting, J. and Geierhos, M.
Well-Being in Plastic Surgery: Deep Learning Reveals Patients' Evaluations.
DOI: 10.5220/0010576002750284
In *Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA 2021)*, pages 275-284
ISBN: 978-989-758-521-0

me in the eye" is more common than a noun-based construction. Commercial reviews, typically used for ABSA research, usually feature adjective-noun combinations such as "*bright display*".

In addition to performing ABSA, we also want to present our research design and subsequent data analyses. Thus, we correlate our results with meta-information from the websites. Therefore, we chose a specific physician specialty, plastic surgery. We consider our correlation results to be more valuable when the extracted aspects, their polarity, etc., are brought together with metadata such as assigned grades for a particular medical specialty. This allows us to draw conclusions if the equipment of a plastic surgeon's practice is of particular importance for patients.

Since our data contain indirect, complex, and long aspect phrases, we train and compare a number of recent deep learning approaches to find the best performing solution. To find mentioned aspects in texts, ABSA research has developed three subtasks: Aspect Term Extraction (ATE), Aspect Category Classification (ACC), and Aspect Polarity Classification (APC) (Chinsha and Shibily, 2015). ATE means to find and extract aspect phrases in texts. ACC focuses on classifying them into the correct aspect category and APC classifies the extracted phrases into the appropriate sentiment polarity class, e.g., positive or negative sentiment. ATE and ACC are usually performed together, which is consistent with our previous work (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2021b). For our deep learning approaches, we need to use supervised deep learning, because implicit mentions and longer phrases cannot be captured by keyword spotting. The following sentence exemplifies it:

Example 1 (typical physician review)**.** "*I want to emphasize how well Dr. Myers conducted my plastic surgery.* ***Appointments are usually made to accommodate the wishes of the patients*** *and* ***waiting in the practice never (except once) takes long****. However, I think that the* ***newest technology is available****, they also own a* ***new apparatus for breast examinations****. The* ***rooms are full of light****, while privacy is maintained.*"

Example 1 shows common aspect phrases. Previously, we examined aspect classes dealing with the doctor and his or her team (Kersting and Geierhos, 2021b; Kersting and Geierhos, 2021a). The aspect classes we examine in this study deal with the practice of a physician: "*Waiting Time for an Appointment*", "*Waiting Time*" (in the practice), "*Equipment/Facilities*", and "*Well-Being*"[1]. As the example shows, aspect phrases are quite complex and use insertions. In contrast to common product domains, it is not mentioned that "*the battery is good and the display is bright enough*". Our examples are longer, more complex, and have many insertions and a different word order because of the use of German. It is a rather unusual style to say briefly that the "*waiting time for an appointment is good*", it is rather said that "*you MUST wait VERY LONG to get just a tiny, useless 3-minute appointment*".

Contributions. First, we present our annotated dataset `wtwawo`, which consists of physician review sentences that deal with the entire practice as the aspect target instead of focusing directly on human employees (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2020b; Kersting and Geierhos, 2021a; Kersting and Geierhos, 2021b). Therefore, and second, we contribute by training and evaluating a number of recent deep learning architectures such as transformers for ABSA. Third, we collected a dataset from a German-language PRW that deals exclusively with plastic surgeons. Fourth, we apply the previously described models as well as additional models on the gathered reviews. Fifth, we correlate and associate our results with themselves and with metadata obtained from the PRW. To get good correlation data, we also use domain-trained transformers that have been additionally trained on a large set of raw physician reviews to achieve better results with deep learning models. We also present our solution for polarity classification, for which we have also annotated data. Furthermore, we apply our algorithm to determine the relative importance of aspect phrases (Kersting and Geierhos, 2021a).

This paper is organized as follows: Section 1 introduces the topic. After that, Section 2 outlines the relevant state-of-the-art literature, while Section 3 then presents the datasets of this study. Section 4 shows examples of typical physician reviews and describes the annotation task along with its results. Section 5 summarizes the applied methods and their results before Section 6 concludes.

2 RELATED LITERATURE

The following areas of interest relate to the state-of-the-art for this study: (physician) reviews, ABSA, deep learning for ABSA, and correlation and association techniques.

2.1 Physician Reviews

There are several studies that have looked at physi-

[1]Translated from these German terms: "*Wartezeit auf einen Termin*", "*Wartezeit in der Praxis*", "*Ausstattung*", and "*Wohlfühlen*". The acronym `wtwawo` sums them up.

cian reviews (Emmert et al., 2013; Kersting et al., 2019), while most research has dealt with commercial reviews (Mayzlin et al., 2014; Pontiki et al., 2016a). Perhaps the most important finding regarding physician reviews is that trust is extremely important (Kersting et al., 2019), while ratings are mostly positive and the PRWs have some shortcomings (Emmert et al., 2013). However, medical treatments are the most difficult domain to rate (Zeithaml, 1981).

2.2 Aspect-based Sentiment Analysis

While most researchers have not looked at ABSA and physician reviews or comparable data domains, we have recently attempted to bridge the gap and use sophisticated PRW data to perform real-world analyses (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2020b; Kersting and Geierhos, 2021b). That is, in previous work we have investigated implicitly mentioned aspect phrases, focusing on a human-like language understanding. Many other works rely on nouns, seed words, etc. As an example, scholars 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 expression" (Pontiki et al., 2016b). Given the complexity and diversity of human language, such approaches do not go far enough.

Most previous studies use this understanding based on commercial review data (Pontiki et al., 2016a; Pontiki et al., 2015; Pontiki et al., 2014) and also often do not perform ATE (Zhou et al., 2019). It is important that most studies use the same data (Zhou et al., 2019), as the data will shape subsequent processing methods and machine learning approaches, and thus subsequent research. That is, most studies will take similar approaches for their deep learning and other machine learning systems for ABSA. Furthermore, as most studies have neglected human-like language comprehension, they are limited in terms of methods and data (domains) (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2020b; Kersting and Geierhos, 2021b) and thus cannot be applied to physician reviews. We focus here on implicit aspect mentions, insertions, and how we have applied human understanding without limiting the aspects that may be annotated.

2.3 Deep Learning for ABSA

In this study, we want to compare a number of deep learning approaches for ABSA. In our previous work (Kersting and Geierhos, 2021b), we compared neural networks using (bidirectional) Long-Short Term Memories ((bi)LSTMs) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) with Conditional Random Fields (CRFs) (Lafferty et al., 2001) or Attention (Vaswani et al., 2017). We combined them with word embedding techniques that compute vector representations for words and n-grams, i.e., parts of words. We successfully used FastText (Bojanowski et al., 2017) that equates word representations with representations for partial word units. However, this is a static method of embedding computation, newer approaches compute such embeddings ad-hoc based on context or they can be fine-tuned directly to downstream tasks. Such recent approaches are transformers (Vaswani et al., 2017) such as XLM-RoBERTa (Conneau et al., 2020) or BERT (Devlin et al., 2019). While these outperform previous research, methods such as FastText can keep up to some extent (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2021b). Nevertheless, transformers have shifted the performance of natural language processing and emerged a number of different model types. It is not yet sure which one performs best for German data or for physician reviews. Therefore, we compare a number of pre-trained transformers for German and multiple languages, and also domain-train them on a large corpus of raw physician reviews to improve their performance (cf. Section 5).

Here we do not only perform ABSA but also aspect phrase weight computation as well as metadata correlation and association. In a related study, we discuss algorithms for calculating importance based on the weight of an aspect phrase (Kersting and Geierhos, 2021a). Briefly, we can say that those aspect phrases with additional adverbs that serve as modifiers, for example, are more important than those without. An example is the phrase: "*very pleasant atmosphere*" compared to "*pleasant atmosphere*". The same applies to comparative adjectives such as "*longer*" or the superlative "*longest*". Increased weight, i.e., importance, means that reviewers or patients have put more emphasis on what they have written and have thus elaborated it a bit more.

2.4 Correlation and Association

There are several methods we introduce to correlate and associate our data. For correlating numbers, we used the Spearman's correlation coefficient (Spearman, 1904), which is a standard measure that shows how two columns of data correlate with each other on a scale from -1 to $+1$ (perfect monotonic relationship), where 0 means no association. Correlations are also called associations and do not indicate a causal relationship. This means, for example, that a negative

correlation indicates that a higher value of x is associated with a lower value of y (Schober et al., 2018). We use this measure instead of Pearson's correlation coefficient because we do not expect to have normally distributed data (i.e., a bell-shaped curve) (Rodgers and Nicewander, 1988). For instance, quantitative scores from PRWs can be used, but our observations show that they are mostly positive.

For data that are not numerical or cannot be represented by numbers, we used a (different) measure of association, Cramér's V (Acock and Stavig, 1979), which measures the association between two categorical variables. That is, Cramér's V describes the strength, not the direction, of an association (Benning, 2021).

3 DATA

In this section, we first briefly describe the raw data and then the data on plastic surgeons.

3.1 Raw Data from PRWs

All available raw data were crawled from three German-language PRWs, mostly between March and July 2018. We aim at a platform-independent approach and thus collected data from more than one platform. We started with a crawler that collected all physician profile links to cause as few website hits as possible. In total, there are over 400,000 physician profiles on German-language PRWs and over 2,000,000 review texts in the raw dataset. Among other things, physician reviews with associated ratings, the cumulative ratings per physician, and participation in continuing medical education were crawled. The quantitative rating systems use German and Austrian school grades as well as stars. The German PRW is by far the largest in terms of available data (physician profiles, reviews). Therefore, we use it as the basis for our investigation. The German school grading system uses grades $1-6$, starting with the best. It should be noted that the German grade 5 means inadequate ("*mangelhaft*"), which is a devastating grade to receive, while a 6 means insufficient and a total failure ("*ungenügend*"). The number of quantitative rating classes varies widely among PRWs (Cordes, 2018; Kersting and Geierhos, 2020a; Kersting and Geierhos, 2020b; Kersting and Geierhos, 2021b).

3.2 Plastic Surgery Data

In our dataset, we have over 1,600 physicians that were labeled as male or female plastic surgeons. Over half of them are men, less than 20% are women and the rest have no gender assigned. However, only over 800 reviews have cumulative ratings and not all grades were provided. This is quite normal for platforms with user-generated data that are social networks. We applied lower boundaries by selecting those reviews that had both a text and a title. This resulted in over 35,000 reviews. Counting these textless reviews listed on a physician's profile page, there are over 40 ratings per physician. The cumulative grade is 1.27, where 1.0 would be the best possible grade. This means that German plastic surgeons are rated well in quantitative terms. The number of reviews with the best grade is overwhelming.

We split all collected reviews into sentences before any annotation step was performed (Kersting and Geierhos, 2020a). However, due to quality reasons, sentences were excluded, such as extremely short and long ones. These quality limits included a minimum length of the entire review text (280 characters) or the requirement that multiple quantitative scores were assigned to the corresponding reviews (e.g., grade for the physician's competence). These limits were previously established to encourage broader research with the data (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2020b; Kersting and Geierhos, 2021b). Since we have a total of over 2,000,000 review texts, we were confident that exclusion of some would not hinder the overall annotation. However, for our plastic surgeon reviews, we adhere to these quality constraints to maintain consistency, and remain with less than 60,000 sentences out of 16,000 reviews.

4 ANNOTATIONS

Here, we explain how we annotated the relevant data.

4.1 Aspect Phrase Annotation

To find relevant aspect classes, we evaluated all available rating categories from the PRWs (Kersting and Geierhos, 2021b). However, we decided to quantitatively combine the available classes into a set where possible. As described, in this case we aimed at finding classes that evaluate entire medical practices. Examples of the existing classes include team competence. However, our approach is consistent with the literature and involves our previous work that focuses on the physician or his/her team as the aspect target (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2020b; Kersting and Geierhos, 2021b). In the physician reviews, there are three aspect targets at all. We intend to perform ATE and ACC together, as

many scholars have done before (Zhang et al., 2018). The reason for this is the mutual influence between ATE and ACC.

Our newly annotated dataset is called `wtwawo`, which is an acronym for the German names of the aspect classes it contains. As described in Section 1, these are: "*Waiting Time for an Appointment*", "*Waiting Time*" (in the practice), "*Equipment/Facilities*", and "*Well-Being*". However, patients are extremely concerned about how long they must wait to get an appointment. Furthermore, they do not want to spend a lot of time waiting when they have arrived at a practice (cf. Section 5). The following list explains the used annotation labels and their topics:

- "*Waiting Time for an Appointment*" deals with the perceived duration of time a patient has to accept to get an appointment: "*Got an **appointment the next day**!*"
- "*Waiting Time*" (in the practice) also describes the subjectively perceived time a patient has to wait in the physician's practice: "*It **took more than one hour** for my name to be called.*
- Patients use "*Equipment/Facilities*" to describe a physician's technological equipment and tools: "*Her practice has the **latest X-ray machine**.*"
- "*Well-Being*" describes the subjective well-being felt in the practice and how comfortable visitors feel in its rooms: "*The treatment took place in a **very pleasant, relaxed atmosphere**.*" "*The practice is also **very nice and spacious**, there is **water and tea** in the waiting room.*"

PRWs focus on doctors and reviews about them. This may be the reason why most sentences do not contain aspect phrases that deal with the practice. As can be seen in the list above, the classes are quite broad, especially the last two. However, there were classes related to similar topics, which encouraged us to merge them into one global "*Well-Being*" class. We also felt that the well-being factor of a practice should be presented after reading review texts.

Active learning was performed once for all packages before annotations began, consistent with previous work (Kersting and Geierhos, 2021b). Our goal was to find sentences that generally contain an evaluative statement. Therefore, a neural network classifier was used. Then, the annotations for `wtwawo` started.

The process was carried out mostly by one person and followed the general approach of our previous work (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2021b). Because we looked for external help, two external persons were involved in the annotation. Three internal staff members took turns reviewing the annotations of the two external persons. All were trained German linguists and language specialists. The first external person was only available for a short period of time. Since we were unsure about the quality of his annotations, we kept only the sentences that contained aspect phrases (several hundred) and reviewed them. In the following, we excluded all sentences without aspect phrases from this first phase. This, and the fact that we applied active learning, caused the number of sentences without aspect phrases in them to shrink dramatically. Consistent with out earlier studies, we applied an active learning approach several times, in agreement with our previous studies (Kersting and Geierhos, 2021b) (multi-label, multi-class classification) to pre-select sentences with a higher probability of containing one of the four aspect classes. The reason is that, in part, we had to annotate several dozen sentences until a sentence with aspect phrases appeared. In the end, we have over 8,000 sentences, most of which, i.e., over 7,000, contain aspect phrases. This contrasts our earlier work (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2021b), where we were dealing with a higher number of sentences.

The annotations were consistently monitored during the annotation process by discussing them in the annotation team. In the end, we scanned almost all sentences with aspect phrases for errors and deleted them generously from the dataset (< 400), resulting in higher data quality. Examples of errors include missing attention to context, incorrectly assigned labels or annotated phrases that are either too short or too long. During annotation, it was possible to label multiple aspect phrases in a sentence and most sentences tended to be colloquial and thus unstructured. Users tend to write as they would speak. Annotation at the sentence level was sufficient for this type of data (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2021b). Many also contain insertions making the annotation task more difficult: "*I love visiting his practice because the **atmosphere, and this is something interesting (after last Thursday), is lovely and charming**, because of the beautiful pictures that hang everywhere."*

Table 1: Inter-Annotator Agreement (IAA) for our annotated `wtwawo` dataset.

Annotators	Cohen's Kappa	Krippendorff's Alpha
R & Y	**0.710**	
R & J	**0.871**	**0.768**
Y & J	**0.727**	

The nature of the review texts made the annotation task highly difficult. Inter-annotator agreement (IAA) was calculated based on the tagged words, so we as-

signed a tag to each word in a sentence indicating a class. All words that were not annotated with one of the `wtwawo` classes were assigned a "no class" tag. We compared the annotations of the second external annotator, who performed the majority of annotations, and randomly selected about 340 sentences (Kersting and Geierhos, 2020a). The scores of all IAAs can be found in Table 1. The Cohen's Kappa values (Cohen, 1960) show a "substantial" agreement (0.61–0.80) for two of three pairs of annotators. "R & J" achieved an "almost perfect" agreement (0.871) (Landis and Koch, 1977). Krippendorff's Alpha (Krippendorff, 2011) is considered to be good, being close to 1.0. All in all, the agreements are more than satisfactory and reward the challenging annotation task and the amount of manual effort involved.

4.2 Sentiment Polarity Annotation

We performed the sentiment polarity annotations in a different way. For instance, in most cases the distinction between aspect phrases and sentiment words is not possible for physician reviews, which the examples demonstrate: "*A **friendly** doctor.*" Another example could be that "*water and tea*", as shown in the list above, are an implicit indicator of the class "*Well-Being*". These words convey the feeling that it would not be possible to identify a polarity word and a word indicating the aspect class in the phrase. Therefore, we extracted aspect phrases from the `wtwawo` dataset[2] and annotated them with respect to polarity. As it turned out, it is only possible to assign positive or negative polarity, another scale is not adequate[3].

5 METHODS AND RESULTS

In the following, we present our extraction and polarity classification approaches for aspect phrases before discussing the results obtained on plastic surgery data.

5.1 Aspect Phrase Extraction

We trained supervised learning algorithms on the basis of previous research and investigated neural network approaches. In addition to our biLSTM algorithms (Kersting and Geierhos, 2021b), we also used numerous transformers. As tests revealed, we do not have enough PRW data to train a transformer from scratch (no useful results). Here, we turned to those available for German data from Huggingface[4] and a multilingual model (XLM-RoBERTa). These were further domain-trained on raw physician review texts (Kersting and Geierhos, 2021a). That is, as Table 2 shows, we have fine-tuned both: pre-trained transformers and further domain-adapted transformers. The domain-trained models are marked with a "+". FastText can keep up with transformers to some extent, as described, which is why we included it here. We want to further explore the use of transformer models for our case.

Our experimental setup refers to IO (Inside, Outside) tags for ATE and ACC (Kersting and Geierhos, 2020a), e.g., "*I-waiting_time_T*". During domain-training transformers, the loss for XLM-RoBERTa (base) was about 0.37 after 4 epochs, while for most German-language models such as BERT (bert-base cased) it was about 1.1–1.3 after 10 epochs. This is different for Electra (loss above 6.7), which also performed poorly in the aspect extraction task (cf. Table 2). 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 3.1).

As shown in Table 2, XLM-RoBERTa achieves the best scores (F1: 0.86). Interestingly, it also significantly outperforms models with self-trained FastText embeddings (F1: 0.79), which in some cases performed better than transformers in our previous work (Kersting and Geierhos, 2020a; Kersting and Geierhos, 2021b). To obtain the best results, we performed parameter tuning and tested different train-test splits. Since our goal was also to maintain comparability, we mostly used a train-test split of 80%/20% for transformers (epochs: 10) and 90%/10% for the other neural networks (epochs: 6). Because our previous work suggested an advantage of training uncased embeddings (Kersting and Geierhos, 2020a), we trained FastText vectors uncased. The reasons for this lie in the nature of our data: Physician reviews are error-prone and contain user-generated text with medical terms. However, we cannot see any advantage from this. The other models in Table 2 that are not explicitly labeled as (un-)cased are cased. However, some of them are not well-documented, e.g., for MedBERT a reference was published long after the model was used for this study, unlike XLM-RoBERTa (Conneau et al., 2020). Precision, Recall, and the F1-Score are preferred over Accuracy, because Accuracy is likely to show fuzzy results: Most words are not in an aspect phrase and thus labeled as "O", so our classes are very unbalanced. Computing Accuracy would lead to

[2]We also extracted aspect phrases from related datasets dealing with a physician as the aspect target (Kersting and Geierhos, 2021a; Kersting and Geierhos, 2021b).

[3]The generation of this dataset and the classification methods are prior work (Kersting and Geierhos, 2021a).

[4]Besides Huggingface, we thank several developers (Pedregosa et al., 2011; Biewald, 2020; Hugging Face, 2020).

Table 2: Results for the extraction and classification of aspect phrases (ATE, ACC) using broadly pre-trained and domain-trained transformers for the `wtwawo` dataset[5].

Model	P	R	F1
xlm-roberta-base+	0.81	0.91	**0.86**
⌞ biLSTM-CRF+	0.83	0.79	0.81
⌞ biLSTM-Attention+	0.84	0.77	0.80
xlm-roberta-base	0.79	0.89	0.83
MedBERT+	0.80	0.90	0.84
MedBERT	0.80	0.89	0.84
electra-base uncased+	0.15	0.20	0.17
electra-base uncased	0.79	0.90	0.84
distilbert-base cased+	0.81	0.89	0.85
distilbert-base cased	0.80	0.88	0.84
dbmdz bert-base uncased+	0.81	0.90	0.85
dbmdz bert-base uncased	0.80	0.90	0.85
dbmdz bert-base cased+	0.82	0.90	0.85
dbmdz bert-base cased	0.80	0.90	0.85
bert-base cased+	0.81	0.90	0.85
bert-base cased	0.80	0.88	0.84
FastText biLSTM-CRF+	0.84	0.75	0.79
FastText biLSTM-Attention+	0.83	0.77	0.79

very high values for most models; Precision, Recall, and F1-Score averaged across aspect classes show a clearer picture here. To reduce the imbalance and achieve better results, we used only the sentences containing aspect phrases for training the models. As the tests have shown, this is not an obstacle for use with unseen data. The reason could be that the "O" tag is disproportionately represented and hence the models can easily identify irrelevant words.

5.2 Aspect Phrase Polarity Detection

The polarity classification is not described in detail here. We performed it as a binary classification. Qualitative testing has shown that there are almost no neutral aspect phrases and that there are not enough nuances (e.g., highly positive, moderately positive, etc.) to assign polarity scores. The models were given the aspect phrase and the sentence as the context for training. We annotated several thousand sentences for this task and achieved an IAA of three times over 0.91 for Cohen's Kappa and over 0.91 for Krippendorff's Alpha. This is considered as almost perfect. For the application, we chose XLM-RoBERTa, which achieved an F1-Score of 0.94 (Precision: 0.93, Recall: 0.95). XLM-RoBERTa outperformed other transformers trained for German.

[5]P = Precision, R = Recall, F1 = F1-Score. All pre-trained transformer models are in German and are accessible with their names on https://huggingface.co/models, accessed 2020-12-28. BiLSTM-CRF and Attention models are based on previous work (Kersting and Geierhos, 2021b).

5.3 Application to Plastic Surgery Data

Our aspect phrase importance calculation works by identifying phrases that may indicate increased relative weight, e.g., by superlative adjectives, comparatives, and the use of adverbs. We also tested statistical approaches that did not succeed (Kersting and Geierhos, 2021a). Overall, we applied the XLM-RoBERTa models for ATE, ACC, and APC and the aspect phrase importance algorithm to all extracted sentences for plastic surgeons. The results are as follows: Out of about 60,000 sentences, almost 14,000 had aspect phrases of the `wtwawo` dataset in them. This affects over 8,000 out of 16,000 reviews (cf. Table 3).

The number of labels that appeared in plastic surgery reviews (cf. Table 3) demonstrates the importance of each label. Most notably, "*Well-Being*" is featured by far the most often, followed by "*Waiting Time for an Appointment*". Phrases that may be mentioned more often are more important to reviewers. However, mentioning things may not automatically make them important, as mentioning the "*Well-Being*" can be a standard behavior of patients

Table 3: Number of appearances of each aspect class of the `wtwawo` dataset after application to plastic surgery reviews.

Labels	#
Waiting Time for an Appointment	3,720
Waiting Time	1,727
Well-Being	10,914
Equipment/Facilities	967

Table 4: Number of appearances of each aspect class related to its importance and polarity.

Labels	Importance	#	Polarity	#
Waiting Time for an Appointment	high	1,315	positive	3,168
	low	2,405	negative	552
Waiting Time	high	755	positive	1,089
	low	972	negative	638
Well-Being	high	4,327	positive	10,035
	low	6,587	negative	879
Equipment/ Facilities	high	131	positive	902
	low	836	negative	65

reviewing plastic surgeons. Hence, we applied our aspect phrase importance algorithm that determines the true importance on the basis of linguistic features such as superlatives (Kersting and Geierhos, 2021a). The results can be found in Table 4. Apart from importance, it is notable whether aspect phrases are rather positive or negative, i.e. how patients rate plastic surgeons in their evaluative texts, apart from assigned grades. This is equally shown in Table 4.

As Table 4 shows, there are constantly more aspect phrases of *low* or *normal* importance than phrases with *high* importance. The same applies to positive aspect phrases, compared to negative ones. This is not surprising, as reviews are generally very positive. Moreover, it can be seen that the "*Equipment/Facilities*" are generally less important, as the ratio of high and low importance reveals. This is different for the "*Waiting Time for an Appointment*", "*Waiting Time*", and "*Well-Being*".

5.4 Association and Correlation Results

We applied a Spearman's correlation (Spearman, 1904) in order to investigate whether labels have an association with polarity scores. They do have a slight positive correlation of 0.18, when transforming polarity scores to $0 = negative$ and $1 = positive$. A stronger association can be observed when applying Cramér's V (Acock and Stavig, 1979) (0.26). There is, also an association between polarity and importance scores (0.13). But this is not the case for a Spearman's correlation of polarity and importance. We see correlation scores of ≥ 0.10 as notable (Xiao et al., 2016), though larger values indicate a stronger relationship.

However, there is a negative correlation when bringing together each occurring aspect phrase label with the corresponding accumulated grades of physicians (-0.10). Equally, when taking the average polarity per sentence and the accumulated grades of physicians, they are negatively correlated (-0.11). That is, a higher (more positive) polarity of observed aspect phrases in the reviews comes with a better grade. Cramér's V shows that polarity and accumulated grades are quite notably associated (0.29).

We made other findings when organizing the data per review. Unsurprisingly, the overall grade of a review is strongly positively correlated with the physician's accumulated grade (Spearman: 0.26; Pearson: 0.50). This is indeed a finding, because not all reviews received a text and we only considered those containing one of our aspect classes. There seems to be a discrepancy between reviews with texts and `wtwawo` phrases compared to those without texts. Furthermore, we found that negative sentiment polarities among the aspect phrases in a review are slightly associated with a poorer overall grade for each review, as Cramér's V reveals (0.14). This is less or almost not the case for the physician's accumulated grade (0.08).

When correlating the accumulated grade per physician with the polarity of each aspect phrase in the surgery reviews, we find that "*Well-Being*" and "*Waiting Time for an Appointment*" (-0.15, -0.12) are less strongly correlated than "*Equipment/Facilities*" or "*Waiting Time*" ($-0.24, -0.29$). Separating the data by important aspect phrases does not reveal new findings.

Moreover, we selected each label and the gender of the physician. Then we correlated each aspect phrase's polarity with the cumulative grade of a physician. For "*Equipment/Facilities*", we found a strong difference because for males the correlation was -0.23 and for females -0.46. For the "*Well-Being*" class this was different ($-0.27, -0.35$). There is a smaller number of female plastic surgeons, but in general, the observed scores for "*Equipment/Facilities*" are applicable ($-0.21, -0.45$).

6 CONCLUSION

In our study, we annotated a dataset of physician reviews that have the physician's practice as the aspect target. We calculated the IAA and achieved good results, which encouraged us to train and compare a number of transformers and other neural networks to extract implicit aspect phrases from texts, categorize them, and assess their polarity. These ABSA steps (ATE, ACC, and APC) were applied together with an

importance weighting on a dataset of plastic surgery reviews. We found indicators that patients care about a physician's practice, so they rate it quite frequently, especially perceived well-being. In addition, gender appears to have a strong association with the polarity of reviews, according to our data. What is new is that we were able to compare and correlate the polarity of ratings with assigned scores on rating portals. Here we found that assigned quantitative grades and polarity in review texts are associated with each other.

With the methods and analyses presented here, we are able to analyze patient reviews on a large scale and provide physicians, their practices and hospitals with deep insights and relevant information that can help to improve their medical services. This can lead to an improved curative process. Here, our goal was to present general capabilities for further analysis that can help medical providers understand and improve their services in the future. Together with related studies that address aspects targeting the physician and his or her team, we can fully cover the topics related to healthcare providers in review texts. We can also apply our findings to other domains in the future, since implicit expressions indicating aspects are also present in other challenging domains, such as everything related to interpersonal communication.

ACKNOWLEDGEMENTS

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 data collection.

REFERENCES

Acock, A. C. and Stavig, G. R. (1979). A Measure of Association for Nonparametric Statistics. *Social Forces*, 57(4):1381–1386.

Benning, V. (2021). Cramer's V Verstehen, Berechnen und Interpretieren. https://www.scribbr.de/statistik/cramers-v/. Accessed 2021-04-20.

Biewald, L. (2020). Experiment Tracking with Weights & Biases. Technical report. Software available from wandb.com.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the ACL*, 5:135–146.

Chinsha, T. C. and Shibily, J. (2015). A Syntactic Approach for Aspect Based Opinion Mining. In *Proceedings of the 9th IEEE International Conference on Semantic Computing*, pages 24–31. IEEE.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 8440–8451, Online. ACL.

Cordes, M. (2018). Wie bewerten die anderen? Eine übergreifende Analyse von Arztbewertungsportalen in Europa. Master's thesis, Paderborn University.

De Clercq, O., Lefever, E., Jacobs, G., Carpels, T., and Hoste, V. (2017). 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*, pages 136–142. ACL.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186. ACL.

Do, H. H., Prasad, P. W. C., Maag, A., and Alsadoon, A. (2019). Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications*, 118:272–299.

Emmert, M., Sander, U., and Pisch, F. (2013). Eight Questions About Physician-Rating Websites: A Systematic Review. *Journal of Medical Internet Research*, 15(2):e24.

He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2019). An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 504–515. ACL.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hugging Face (2020). Hugging Face – On a Mission to Solve NLP, One Commit at a Time. [Pretrained Models]. https://huggingface.co/models. Accessed 2020-05-14.

Kersting, J., Bäumer, F., and Geierhos, M. (2019). In Reviews We Trust: But Should We? Experiences with Physician Review Websites. In *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security*, pages 147–155. SCITEPRESS.

Kersting, J. and Geierhos, M. (2020a). 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*, pages 391–400. SCITEPRESS.

Kersting, J. and Geierhos, M. (2020b). Neural Learning for Aspect Phrase Extraction and Classification in Sentiment Analysis. In *Proceedings of the 33rd International FLAIRS*, pages 282–285. AAAI.

Kersting, J. and Geierhos, M. (2021a). Human Language Comprehension in Aspect Phrase Extraction with Importance Weighting. In Kapetanios, E., Horacek, H., Métais, E., and Meziane, F., editors, *Natural Language Processing and Information Systems*, vol. 12801 of *LNCS*. Springer. In Press.

Kersting, J. and Geierhos, M. (2021b). Towards Aspect Extraction and Classification for Opinion Mining with Deep Sequence Networks. In Loukanova, R., editor, *Natural Language Processing in Artificial Intelligence – NLPinAI 2020*, volume 939 of *SCI*, pages 163–189. Springer.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Technical Report 1-25-2011, University of Pennsylvania.

Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conf. on Machine Learning*, pages 282–289. ACM.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Li, X., Bing, L., Zhang, W., and Lam, W. (2019). Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. In *Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*, pages 34–41. ACL.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *The American Economic Review*, 104(8):2421–2455.

Nazir, A., Rao, Y., Wu, L., and Sun, L. (2020). Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey. *IEEE Transactions on Affective Computing*, pages 1–1.

Nguyen, T. H. and Shirai, K. (2015). PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514. ACL.

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

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 486–495. ACL.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2016a). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 19–30. ACL.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2016b). SemEval-2016 Task 5: Aspect Based Sentiment Analysis (ABSA-16) Annotation Guidelines.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 27–35. ACL.

Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008. Curran Associates.

Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., and Biemann, C. (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. Springer.

Xiao, C., Ye, J., Esteves, R. M., and Rong, C. (2016). Using Spearman's Correlation Coefficients for Exploratory Data Analysis on Big Dataset. *Concurrency and Computation: Practice and Experience*, 28(14):3866–3878.

Zeithaml, V. (1981). How Consumer Evaluation Processes Differ between Goods and Services. *Marketing of Services*, 9(1):186–190.

Zhang, L., Wang, S., and Liu, B. (2018). Deep Learning for Sentiment Analysis: A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):1–25.

Zhou, J., Huang, J. X., Chen, Q., Hu, Q. V., Wang, T., and He, L. (2019). Deep Learning for Aspect-Level Sentiment Classification: Survey, Vision, and Challenges. *IEEE Access*, 7:78454–78483.