Towards Aspect Extraction and Classification for Opinion Mining with Deep Sequence Networks

Joschka Kersting and Michaela Geierhos

Abstract This chapter concentrates on aspect-based sentiment analysis, a form of opinion mining where algorithms detect sentiments expressed about features of products, services, etc. We especially focus on novel approaches for aspect phrase extraction and classification trained on feature-rich datasets. Here, we present two new datasets, which we gathered from the linguistically rich domain of physician reviews, as other investigations have mainly concentrated on commercial reviews and social media reviews so far. To give readers a better understanding of the underlying datasets, we describe the annotation process and inter-annotator agreement in detail. In our research, we automatically assess implicit mentions or indications of specific aspects. To do this, we propose and utilize neural network models that perform the here-defined aspect phrase extraction and classification task, achieving F1-score values of about 80% and accuracy values of more than 90%. As we apply our models to a comparatively complex domain, we obtain promising results.

Keywords Sentiment analysis · Opinion mining · Text mining · Neural networks · Deep learning

1 Introduction

Researchers are becoming increasingly interested in sentiment analysis—also known as opinion mining—because of the steadily growing, mostly user-generated amount of textual data on the web. Pertinent examples are review websites and social media websites. Due to the nature of these websites, recent research in sentiment analysis

J. Kersting
Paderborn University, Warburger Str. 100, Paderborn, Germany
e-mail: joschka.kersting@upb.de

M. Geierhos (✉)
Bundeswehr University Munich, Research Institute CODE, Carl-Wery-Straße 22, Munich, Germany
e-mail: michaela.geierhos@unibw.de

R. Loukanova (ed.), *Natural Language Processing in Artificial Intelligence—NLPinAI 2020*, Studies in Computational Intelligence 939,
https://doi.org/10.1007/978-3-030-63787-3_6

has often been document-based, as its main goal is to identify an overall sentiment for whole documents or sentences. In contrast, aspect-based sentiment analysis (ABSA) aims at a fine-grained analysis of opinions expressed about the features of products, services, etc. This approach is necessary because most expressed opinions are related to single aspects of the overall entities. These aspects might contradict each other and numerous opinions might be voiced in one sentence. Up to now, research has mainly focused on domains with a common vocabulary represented by bag-of-words models, and it has neglected that aspects are sometimes represented implicitly by phrases rather than by direct remarks. The research has also overlooked including an understanding of why users rate in a certain manner, which is necessary when using review data that is available in substantial amounts [49]. ABSA is therefore a challenging and rewarding field of research to investigate.

Opinion mining (i.e., sentiment analysis) involves three approaches, focusing on the document, the sentences or the aspects. Document-level and sentence-level sentiment analysis suppose one opinion per document or sentence. In ABSA, this is different because it assumes that more than one opinion can be expressed. The following example demonstrates the necessity of ABSA rather well: "*Doctor Doe was **very nice*** but she ***did not shake my hand***." The expressions in bold demonstrate that the physician's friendliness is rated twice: once positively, and once negatively. Finding an expression such as the negative one shown above is simple for a human but difficult for an algorithm.

1.1 Contribution

Our research addresses two of the three subtasks involved in ABSA: namely, aspect term extraction and aspect class (i.e., category) classification, excluding aspect polarity classification [16]. With this research, we advance the field of ABSA by incorporating implicit, indicated aspect remarks in longer phrases, which are often long and complex in their form and also unique, due to differing wordings, insertions, etc. We decided to investigate German texts, a morphologically rich and complex language, and selected physician reviews for the dataset because of the high variety of vocabulary (e.g., professions and diseases in the medical sector versus health-related common vocabulary). By presenting our datasets, describing the annotation process and also delivering a rich set of examples, we create a new gold standard for user-generated content in analyzing physician reviews. These datasets are the base for training neural network architectures by applying a variety of word embedding technologies. In doing this, we evaluate our networks without separating the step of aspect phrase extraction and classification, in contrast to approaches used for some shared tasks [59].

This chapter is an extended version of [35], enhancing the previous study by adding a second dataset with a new set of aspect classes that are explained and presented in detail through several examples. This extended version also updates the state-of-the-art and adds more details regarding the annotation process. We also describe

the annotation process for the new dataset, which was evaluated by calculating the inter-annotator agreement. In comparison to our previous study, we added improved deep neural network architectures and word embedding configurations. For example, we use an architecture with an attention mechanism and also apply several newly trained word embedding algorithms. The combinations of these technologies were thoroughly evaluated. We also provide more training details and reasons for our design decisions.

1.2 Domain Focus

Our research focuses on physician reviews, where it is not possible to perform keyword spotting, due to the fact that several aspects are only implicitly mentioned or indicated by long phrases that also have off-topic insertions. This is especially the case for physician reviews because they deal with services that are usually personal, sensitive, and based on trust [10, 34]. Nevertheless, a great amount of ABSA studies proposes that nouns and noun phrases do represent aspects sufficiently or that aspects are explicitly mentioned in some form [8, 12, 32, 55, 59, 63]. This is especially the case for product and service reviews—so-called search goods. These search goods (e.g., smartphones) will always contain the same specific parts (e.g., a smartphone has a display, battery, etc.). But there are also experience goods, whose performance can only be evaluated after experiencing them because their nature is different and subjective [78]. Most ABSA studies focus on products [16] and services that use a rather limited vocabulary. For instance, the aspects of hotels or restaurants can be rated by using simple nouns: breakfast, cleanliness, etc. Such domains are more often characterized as experience domains than as services.

The term 'experience goods' characterizes physician reviews rather well. They characterize special services that involve intimate, personal, and private components. Each performed health service is unique, due to several involved factors such as the patient's personality, the symptoms, the current and any previous diseases, the personality of the health-care provider, and circumstances such as the rooms and the patient's current feelings. In addition, services performed by persons are generally reviewed on the basis of the staff's behavior, so keywords focus on the reliability, empathy, and the ambiance of the rooms [79]. These types of health reviews can be accessed on physician review websites (PRWs) such as RateMDs[1] in the USA, Jameda[2] in Germany and Pincetas[3] in Lithuania, to mention a smaller country. On PRWs, users can write qualitative review texts or assign quantitative grades (e.g., between one and five stars, where five is the best), which can usually be assigned to certain aspects such as the physician's friendliness or perceived competence. The review websites also offer functionalities such as blogging, appointment services, etc.

[1] http://ratemds.com, last visit was on 2020-05-19.

[2] http://jameda.de, last visit was on 2020-05-19.

[3] http://pincetas.lt, last visit was on 2020-05-19.

Yet, trust is still an important issue on PRWs. Some physicians do not prefer to be rated, so there are legal consequences, and others also feel unfairly rated. Moreover, users generally feel anonymous, even though they may be identified by the PRW or their physician through technical information or their review texts [2, 10, 34].

The outline of this chapter follows these steps: Sect. 2 deals with related research, current approaches and datasets in the domain of ABSA. Section 3 describes our dataset in general as well as the aspect classes and the annotation process in particular. Section 4 presents our neural network architectures and their functionality for identifying aspect phrases in German physician review texts. The discussion and evaluation can be found in Sect. 5. Section 6 concludes and proposes future work.

2 State of Research

Extracting aspect phrases and classifying them is the core task in ABSA [12]. This is different in comparison to classification tasks, where a whole document receives a target value. In ABSA, words and phrases must be found from sequential data. Consequent steps in extraction and classification are to identify sentiment-related words or phrases and then distinguish their polarity, meaning their positivity or negativity [12]. Besides this, sentiment analysis must also tackle other issues such as analyzing emotions or detecting sarcasm [80].

Since 2004, the subject of ABSA has been closely investigated by the research community. Especially, Hu and Liu [32] published a key work. Such studies typically deal with products like televisions, in which sentences like the following are common: "*The screen is sharp and nice, but I hate the voice quality.*" Since such reviews are published on the product page in an online shop, common aspects can be derived from other sources or by extracting the nouns and noun phrases [59, 61]. Another approach is to use topic modeling with a list of seed words [53, 81]. Furthermore, some scholars [59, 61] have written annotation guidelines which state 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 [61].

They used annotated data in more than one language for extracting aspect terms and polarity—but not in German, and their domains involved hotels and restaurants. The results for the separated steps, such as aspect term extraction and classification, were almost all below 50% [59].

Other approaches use dependency and constituency parsing in order to find relevant words. However, these words are mostly nouns, too [55]. The main issue with nouns is the assumption that aspects are mostly explicitly mentioned in texts through single words or short noun phrases. This does not comply with the complexity of human language texts. While there are explicit aspect remarks in the form of nouns and noun phrases, there are also implicit forms that come in forms such as adjectives, verbs, etc., and in complex constructions of these. A rather simple example is the adjective "*expensive*", which can refer to the price of a smartphone [44].

Section 3, which presents the data of this work, will demonstrate that the reviews contain implicit mentions of aspects in various forms (e.g., several word types). However, some scholars extract only the most frequent nouns and group them into synonym classes in order to find aspect phrases in texts [12].

In comparison to the previously mentioned studies, Wojatzki et al. [76] take a different path, by using customer reactions from Twitter and not review texts. In contrast to our study, their social media data evolve around the German railway company Deutsche Bahn. They annotated data with the goal of identifying aspects in texts by extracting the corresponding words or phrases, refraining from using data from customer reviews or from a sensitive health- and trust-related domain [10, 34]. Furthermore, the aspects in [76] relate to a large train corporation and are thus not as diverse as those related to all different kinds of physicians, medical professions, diseases, symptoms and the sensitive patient-physician relationship [34]. In the health-care domain, most aspects can be described with nouns, such as "*too loud atmosphere*" or the "*poor connectivity*", and specifically refer to things such as the seats, the noise level, etc. Hence, the general topic of this study has already been touched, yet the domain, approach, and dataset are different.

Another study to mention here builds an ABSA pipeline for data on retail, human resources, and banking in the Netherlands. It is the scientific contribution of [16] to ABSA in the sector of commercial and financial services. Here, it is interesting that their approach is based on earlier studies that suggest regarding aspect term extraction as a sequential labeling task. Inside, Outside, Beginning (IOB) tags [7] are used to determine inside, outside, and beginning tokens of aspect phrases. They used more than 20 classes per domain and annotated their data manually. For aspect term extraction, they achieved very high scores. But for the results of the aspect classification, they in part have values beneath 50%. According to the domain and presented samples, it seems that the most aspect terms are nouns.

Several datasets for ABSA have been published and used so far [16, 18, 20, 52, 59, 60, 62, 65–67, 76] and multiple surveys and summarizing studies have been published [32, 33, 44, 45, 50, 56, 74, 77, 80, 82]. However, some researchers regard nouns either as representations of aspects or as at least sufficient for extracting aspects from texts [8, 12, 55, 59]. Many of these datasets and thus the studies that use them focus on product reviews or commercial service reviews [16]. Among existing datasets, the investigated domains were products and services like restaurants [59], non-evaluative data like Tweets, and other social media data [20, 52] or newspaper articles and other texts [18].

Examples of other approaches are the semi-supervised or unsupervised methods for building ABSA systems [28, 53, 81]. For example, Garcia-Pablos et al. [28] start with a list of seed words to find aspects in large datasets. Yet, such topic models find clusters and topics that are usually not comprehensive and therefore do not comply with the goal of ABSA tasks [53]. Another issue is that such models cannot extract phrases. That is, they provide information on what is given in a text document but cannot sequentially identify the exact words that correspond to the information in the text. Thus, further linguistic analyses are excluded. There is also a variety of datasets that may be used for further analyses based on words in aspect phrases, such

as SentiWS [64] or SentiWordNet [3]. Identifying not only aspects but also the words that describe them in a document encourages further approaches for new analyses. This may enable researchers that use existing and tested data instead of annotating data anew.

In general, there are a multitude of studies that work with PRWs [5, 6, 10, 11, 21–24, 29, 30, 34, 35, 39, 41, 46, 48, 54, 68, 70, 75]. PRWs exist in at least 28 European countries [15]. The portals are quite similar, but show differences, too: RateMDs has four quantitative rating classes [27, 75] while Jameda has 25, Docfinder[4] has nine in Austria and Medicosearch[5] from Switzerland has six [15].

Zeithaml [78] already mentioned that medical diagnoses are among the most difficult things to evaluate. Other scholars found that physician reviews are very influential in choosing an adequate physician [22] and that most ratings are positive [24]. The importance of physician reviews is underlined by these studies. Finally, researchers must always remember that reviews from PRWs have a specific vocabulary and that trust is a basic requirement [34].

3 New Datasets for Aspect Phrases in Physician Reviews

Our dataset consists of physician reviews from several PRWs located in three German-speaking countries (Austria, Germany, and Switzerland) and in which the written language is German.

3.1 Data Collection and Overview

We gathered the ground data in mid-2018, from March to July. For data download, a distributed crawler framework was developed. We checked the sites manually and found an index that we crawled first. This enabled us to directly access all sites that correspond to a physician. Then, we saved all reviews including the ratings and other information that were publicly displayed. To keep the costs for all stakeholders as low as possible we respected good behavior and did not cause much site traffic. Hence, we collected the data over several weeks before saving them to a relational database [15]. Apart from reviews, including ratings and average ratings for a physician, we collected further information such as the address, opening hours, and introductory texts if written by the physicians themselves. This data might be useful in the future. To gain a broader view, we also collected data from a Spanish, an American, and a Lithuanian PRW in their corresponding language. This gave us the opportunity to make general observations or qualitative comparisons, e.g., regarding the use of

[4] http://docfinder.at, last visit was on 2020-05-19.

[5] http://medicosearch.ch, last visit was on 2020-05-19.

Table 1 Statistics for German-language PRWs [35, 36]

PRW	Jameda	Docfinder	Medicosearch
# Physicians	413,218	20,660	16,146
# Review texts	1,956,649	84,875	8,547
# Professions	293	51	139
Avg. rating	1.68	4.31	4.82
Rating system (best to worst)	1–6	5–1	5–1
Men/Women	53/47%	71/29%[a]	No Data
Length (Char.)	383	488	161

[a]Only few data were available

rating classes. Detailed observations on the researched data were taken from the three German-language PRWs: Jameda, Medicosearch, and Docfinder.

General statistics on this data can be found in Table 1, which demonstrates that most physicians are listed on Jameda in Germany. Fewer physicians are listed on the Austrian Docfinder and the Swiss Medicosearch, while Docfinder and Jameda are both visited more often than Medicosearch, presumably due to the higher number of reviews. On average, the ratings are very positive. The higher number of professions listed in Jameda and Medicosearch may be caused by allowing non-official professions to be listed. For our further steps, we excluded reviews that were in languages other than German.

However, the acquired data are not representative for all physicians in the named countries nor in any country, for that matter. The data neither represent all physicians or patients, nor do they provide a consistent rating for a country's health-care system. Still, they do provide many insights that would otherwise not be possible to date and the combination of textual reviews and quantitative ratings in the form of grades encourages a multitude of analyses. For example, as there are over 2 million sentences, there is a rich linguistic corpus of evaluative statements. The relation among physician and patients is sensitive [34] and, apart from protection, requires further evaluation due to data security questions. Moreover, since patients almost anonymously [10] reveal their findings and use subjective phrases, the data source is also rich in non-standard vocabulary. While these statements also contain many incorrectly spelled words and other typos, they are usually written honestly but may sometimes not reveal the true intention of the reviewer.

The analyses also involve how people rate services. While experience goods and services are more difficult to evaluate than products and medical diagnoses are the most difficult to evaluate [78], physician reviews are usually rated on the behavior of the staff [79]. This is revealed by statements such as: "*He did not look me in the eye.*" and "*Dr. Mueller did not shake my hand and answered my questions waaay too briefly.*" Further challenges are that PRWs can contain errors, they can decide which information (not) to publish [11], and rating schemes on PRWs are not scientifically

grounded, as far as we know. A scientific approach would be to employ scales such as the RAND questionnaire which uses 50 items to measure patient satisfaction [24].

3.2 Rating Classes

To find aspect classes that should be annotated, we used the available classes that can be assigned on Jameda, Docfinder, and Medicosearch. We further used qualitative methods to set fixed classes, using the base from all three PRWs. Examples of the classes that can be found on the PRWs are "*explanation*," "*friendliness*," etc. The rating classes for annotation were developed by discussing classes from the PRWs and then merging them semantically. A selection of classes that can be found on the websites is presented in Table 2 (translated from German). While Jameda has the most classes, it presents only a subset of them for each profession (Table 2).

To extract the aspect terms and classify them, we manually annotated a dataset in order to use it for supervised machine learning. For this study, we chose the aspect target "*physician*" and annotated over 26,000 sentences in two datasets. The first dataset ("fkza", an acronym of its German letters) has 11,237 sentences

Table 2 Rating classes on PRWs (selection; translated from German) [15, 35]

Source	Rating classes
Jameda	Treatment, counselling/care, commitment, discretion, explanation, relationship of trust, addressing concerns, time taken, friendliness, anxious patients, waiting time for an appointment, waiting time at the doctor's office, opening hours, consultation hours, entertainment in the waiting room [...]
Jameda—Profession "dentist[a]"	Treatment, explanation, mutual trust, time taken, friendliness, anxious patients, waiting time for an appointment, waiting time at the doctor's office, consultation hours, care, entertainment in the waiting room, alternative healing methods, child-friendliness, barrier-free access, equipment at the doctor's office, accessibility by telephone, parking facilities, public accessibility
Docfinder	Overall assessment, empathy of the physician, trust in the physician, peace of mind with treatment, range of services offered, equipment of the doctor's office/premises, care by medical assistants, satisfaction in waiting time for an appointment, satisfaction in waiting time in the waiting room
Medicosearch	Relationship of trust (the service provider has taken my problem seriously), relationship of trust (the service provider has taken time for me), information behavior (the service provider has given me comprehensive information), information behavior (the explanations of the service provider were understandable), recommendation (the service provider has fulfilled my expectations), recommendation (I recommend this service provider)

[a]Example extracted from Jameda (limited number of classes available).

containing the following classes: *friendliness*, *competence*, *time taken* and *explanation*.[6] The second dataset (the acronym "bavkbeg") contains 15,467 sentences with six classes: *treatment*, *alternative healing methods*, *relationship of trust*, *child-friendliness*, *care/commitment* and *overall/recommendation*,[7] where the final aspect class applies not just solely to the physician but to the physician, the team, and the offices in general. The second set of classes ("bavkbeg") has a larger dataset than the first because of the number of sentences that actually contain an aspect phrase. Both sets have roughly the same number of sentences that a machine learning model learns from, as experiments have shown.

Systems usually extract the target and the aspect together [80]. In our PRW data, we found three opinion targets: the physician, the team, and the doctor's office (e.g., "*parking situation*"). Furthermore, one particular target corresponds to a general evaluation that has just one aspect: *overall/recommendation*. An example here is: "*Totally satisfied.*" Before we describe the annotation process, the named aspect classes are clearly defined below.

3.2.1 First Rating Class Dataset (fkza)

- *Friendliness* deals with the question whether the physician treats his/her patients respectfully and kindly, whether he/she is nice or nasty when greeting them and whether he/she looks them in the eye? Generally, *friendliness* refers to the degree of devotion. Examples (translated from German) are:
  - "*She was nice to me and her assistants are quite efficient.*"
  - "*I don't understand it, he neither greeted me, nor did he listen.*"
- *Competence* describes the subjectively felt or demonstrated expertise of a physician. The question is whether a rater sensed that the physician knows what to do, why to do it, and how to do it. It neither asks about the general quality of treatment nor does it cover friendliness or empathy. This class also includes whether a physician is good in his/her profession in general: i.e., "*knows his job*" or "*knows how to reduce anxiety*" are seen as ratings of competence.
  - "*He is not competent but has a conscientious manner.*"
- *Time Taken* refers to the amount of time a physician takes during his/her appointments with patients. Time is a crucial aspect for the perceived quality of a treatment and the treatment itself. When a physician takes sufficient time or much time, patients see this as a positive signal and therefore express positive sentiment toward that physician. However, in the German language and especially in this class, other

[6]The aspect classes were translated from German: *Freundlichkeit*, *Kompetenz*, *Zeit genommen*, *Aufkärung*.

[7]The aspect classes were originally in German: *Behandlung*, *Alternativheilmethoden*, *Vertrauensverhältnis*, *Kinderfreundlichkeit*, *Betreuung/Engagement*, *Gesamt/Empfehlung*.

words are placed between words such as "*take*" and "*time*", comparable to these English phrases:

– "*She took a lot of time and [...]*"
– "*However, there is one remarkable drawback, because the practice is always overcrowded, so the personal consultation is rather short.*"

- *Explanation* deals with the clarifications a physician uses to explain symptoms, diseases, and (especially) treatments in an understandable manner, because patients naturally need to be informed well. This class is separated from the *time taken* class because a lengthy conversation may indicate the amount of time used, but the quality of the explanation depends on the details described during this conversation or the questions that the physician asks and answers.

  – "*I received a detailed clarification from Dr. Müller.*"
  – "*The consultation was great and I was very well informed about my disease pattern.*"

3.2.2 Second Rating Class Dataset (bavkbeg)

- *Alternative Healing Methods* describes whether a physician offers alternatives to his/her patients. That is, if the physician discusses possible treatments with patients and offers other ways. This does not involve the explanation of a treatment, but rather the general attitude towards alternatives, including alternatives that deviate from conventional medicine, which some patients seek.

  – "*She also offers alternative methods.*"
  – "*He is absolutely not a single bit open for homeopathy.*"
  – "*The doctor deliberately treats alternatively, which I disliked.*"

- *Treatment* deals with the way a physician treats his/her patients, in comparison to the *competence* class of the first dataset. As the datasets were annotated one after the other, certain decisions arose during annotation that needed to be made in order to classify certain cases. That is, a sentence such as "*The doctor is conscientious.*" belongs to *competence*, but the phrase "*treats conscientiously*" applies to *treatment*. In German, these two words (adjective and adverb) would look and sound the same. Therefore, this class is rather narrowly defined.

  – "*I was satisfied with the treatment.*"
  – "*She treats conscientiously.*"

- *Care/Commitment* includes all evaluations of whether a physician is further interested or involved in caring for and committing to the patient and the treatment. This can be seen in phone calls after a visit, in asking a patient further about his or her well-being, continuing to work even after the shift, etc.

  – "*After the surgery, the doctor came by to ask me how I am.*"
  – "*He also postponed his workday off for me. Very committed doctor!!!*"

- *Overall/Recommendation* indicates that the physician, the team, and the office are generally recommendable so that patients are satisfied and return. If this expressed satisfaction refers solely to the competence of the physician, it would be classified as *competence*.
  - "*I came here based on recommendations.*"
  - "*I'd love to go again. You're in good hands!*"
- *Child-Friendliness* describes how a physician takes care of minors. Children need special protection and when talking to a physician is involved, they may not be able to express their medical needs and current situation accordingly. Physicians are therefore especially required to be able to handle children, which this aspect class investigates: i.e., "*Is a physician treating pediatric patients well?*"
  - "*She does not talk to my child at all.*"
  - "*However, he listens to my son.*"
  - "*Listening to kids—not possible for Dr. Müller.*"
- *Relationship of Trust* describes the sensitive relationship between patient and physician. It concerns the question whether the patient has confidence in his/her medical service provider and whether this is expressed in a review. Patients often visit the same physician for years, which can be a way to express their trust. Other forms of trust can be seen in these examples:
  - "*I feel taken seriously.*"
  - "*He doesn't understand me.*"
  - "*I've been going to Dr. Doe for 10 years.*"

All of the aforementioned classes can be clearly distinguished. Just a noun or a single word usually do not clearly indicate the class, so multi-word phrases need to be annotated. The overall task was to review and discuss linguistic constructions that arose during the annotation. We created guidelines and documented cases that are on the edge. Yet the process is complicated because the typical German sentence structure provides much flexibility, especially since the word order can be changed quite freely.

3.3 Annotation Process

We started the annotation process by splitting reviews into sentences using the tokenizer of the spaCy library [25]. Extremely short sentences and rather long sentences were excluded; sentences that started in the heading and continued in the review were merged. We used sentences instead of full review texts as these contain aspects represented by complex phrases and (generally speaking) because sentence-level annotation is more efficient than at the document level. Overall, we had over 2 million review sentences. As many sentences do not contain relevant information, we annotated approx. 10,000 sentences on the basis of whether they contain an evaluative

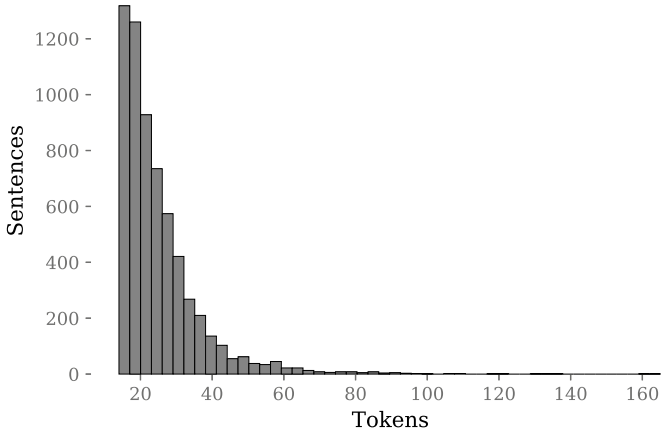

Fig. 1 Number of tokens per sentence with aspect phrases for the first dataset (fkza) [35, 36]

statement. Using this set, and with a high level of agreement among participating persons, we built a Convolutional Neural Network (CNN) [43] classifier,[8] which determined whether a sentence contained an evaluative statement. The annotation was conducted when the first set of rating classes, fkza, was already developed while the other classes were not set yet. This especially applies to fringe cases (i.e., phrases that could in theory be assigned to more than one class) that created the need to make decisions while annotating the corresponding dataset. Hence, this preselection may contain a certain bias. We tackled this issue by randomly saving all sentences with a minimal probability for an evaluative statement to a new file and then using this to annotate aspect phrases and their classes. For each dataset, we randomly set up a new input file. We regarded this approach as superior and less constrained in comparison to other approaches, such as using seed words [13] which would have caused a loss of information because our vocabulary is diverse and consists of longer phrases that indicate aspects. The dataset was annotated by one person while two other trained persons contributed to discussions, evaluations, and the inter-annotator agreement.

In the first rating class set (fkza), there are 11,237 sentences: 6,337 with one or more of the four classes, 4,900 without. In each sentence, it was possible to annotate several aspects and also the same aspect several times: i.e., a sentence could contain all four classes and also the *friendliness* class twice. We stored our annotations in a database where we also saved the tokenization. Figure 1 shows the sentence length for fkza measured in counted tokens. Most of the sentences are brief, but there are some instances of longer sentences.

The second rating class set (bavkbeg) includes 15,467 sentences: 6,600 contain evaluative statements and 8,867 did not. The higher number of sentences without evaluative statements derives from the fact that the six classes of the second set

[8]The neural network was inspired by [37].

did not appear in the data often, especially when compared to the first set. During annotation, this quota was even worse for some time. Consequently, we searched for approaches to increase the number of sentences for annotation that contain the desired classes. As we had viewed over 11,000 sentences, we were confident enough to train a neural network classifier that can predict whether a sentence contains one or more aspect classes in general (multi-output, multi-class classification). At that time, we had seen about 3,000 sentences with aspect phrases and about 8,000 without. The architecture we used is a CNN combined with a bidirectional Long Short-Term Memory [31, 37]. The accuracy was high for the test set (over 90%) and can be considered a good metric, because we also over- and undersampled the data, cutting the number of sentences without useful aspect phrases to reduce their overweight in the classification. This led to sufficient classification results and produced a new input file for the annotations. In each line there was a sentence that contained one of the six mentioned classes (on an alternating basis) and for each sentence more classes may have been predicted: For example, sentence one had the classes *child-friendliness* and also *treatment*. The sentences were chosen on the basis of whether they contained one of the classes and not on a high probability for a class. In the end, each sentence that was proposed for annotation had a high probability for at least one class, while each class appeared in every sixth sentence (with a high probability).[9] We did this again twice, while also increasing the value for what was regarded as a classification: i.e., going from a probability of 50%, which was required to expect a sentence to belong to a class, to a higher value such as 70%. Figure 2 presents the sentence length of the sentences that contained an aspect phrase. As can be seen, in comparison to Fig. 1, most sentences are rather short, but there are also many long sentences, such as the following example which was translated from German and based on the datasets:

> "*Is* **able to do what is necessary** [**competence**] *and I* **trust in him** [**relationship of trust**], *a good match: Dr. Meyer* **knows what he is doing** [**competence**] *and* **always welcomes me so perfectly**, [**friendliness**] *he* **takes a lot of time** [**time taken**] *for anyone who is visiting and* **answers every question** [**explanation**], *I* **totally recommend** [**overall/recommendation**] *him to anyone*."

This example is an illustration of a common review sentence. The aspect phrases are printed in bold, followed by their classes. The authors usually use colloquial language and often misspell names of diseases or treatments, in part referring to the same aspect more than once while using different phrases and implications, but often not just nouns. This is different than in other studies [59, 76]. Furthermore, our dataset is larger than [59] (e.g., English laptop ratings: 3,308 sentences; Dutch restaurant reviews: 2,286 sentences). What is more, they annotate just one of potentially several mentions of an aspect in a sentence. Wojatzki et al. [76] have a slightly larger dataset and not all sentences in their dataset contain aspect phrases. Although their dataset includes 2,000 sentences more than our first dataset (fkza), we have two datasets and thus more data.

[9]The scheme was as follows: sentence: 1, classes predicted: [1]/ sentence: 2, classes predicted: [2, 5]/ ...; sentence: 7, classes predicted: [1]/ etc.

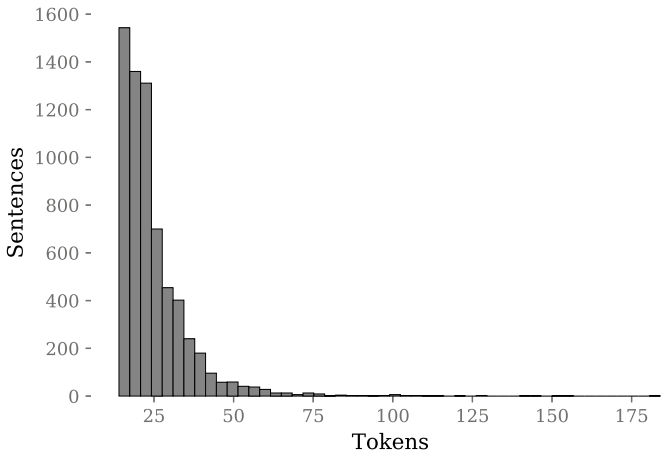

Fig. 2 Number of tokens per sentence with aspect phrases for the second dataset (bavkbeg)

The way how users describe their consultations in reviews makes the annotation task a difficult endeavor. The inter-annotator agreement was calculated on the basis of tagging: i.e., each word received a tag that marked it as part of an aspect phrase (including the corresponding class) or as part of a non-relevant phrase (*None-class*, see Sect. 4). For the first rating class set (fkza), the approach was conducted as follows.

From the data that were annotated by the main annotator, we selected roughly 3%: i.e., 337 sentences. The other two persons re-annotated the data from scratch. Looking into previously annotated datasets was allowed. Table 3 demonstrates our evaluation results. Cohen's Kappa [14] was calculated for each of the two annotators. We used the well-tested Scikit-learn software library for this [58], achieving a substantial agreement among annotators with scores of 0.722 up to 0.857. All values between 0.61 and 0.80 can be considered substantial agreement; values above this can be regarded as almost perfect [42]. In addition, we calculated Krippendorff's Alpha [40] by using the Natural Language Toolkit (NLTK) [7] because it allows one to consider all annotators at once, which we did. The score of 0.771 can be regarded as good; 1.0 would be the best. Alpha provides features such as calculating for many annotators at once (not only two) and can also handle missing data and any number of classes [40].

Table 3 Inter-annotator agreement for three annotators and two datasets [35, 36]

	First set (fkza)			Second set (bavkbeg)		
Annotators	**R & B**	**R & J**	**B & J**	**R & B**	**R & J**	**B & J**
Cohen's Kappa	**0.722**	**0.857**	**0.730**	**0.731**	**0.719**	**0.710**
Krippendorff's alpha (for all 3)	**0.771**			**0.720**		

The inter-annotator agreement for the second dataset was calculated in the same manner as for the first one. As we calculated it before having finished the annotation, 3% corresponds to 358 sentences of the data. If we sensed that the task was comprehended mostly incorrectly, the task was allowed to start over from the beginning (which happened once). As can be seen in Table 3, the scores were sufficient after starting again. The agreement can be considered substantial although there is no "almost perfect" agreement between two annotators. Krippendorff's Alpha has a value of 0.720, which is a good score. The Kappa scores all range slightly above 0.70.

4 Neural Networks for Aspect Phrase Extraction and Classification

In our solution for aspect phrase extraction and classification, we ultimately used our annotated dataset to perform supervised machine learning. Along the way, we tried several paths which all failed to work out, even though we searched the literature to find the best system. Liu [44], for example, presents four approaches to extract aspects: (1) using frequent noun (phrases), (2) utilizing opinion and target relations, (3) applying supervised learning and (4) making use of topic modeling [44]. We attempted all four approaches, none of which provided sufficient results, except for supervised learning. Furthermore, Sect. 2 supports our conclusion. In our tests, topic modeling found "topics" that were not sharply separated and were not understandable to a human: i.e., humans would have drawn a topic completely differently. Frequent nouns are not sufficient either, as mentioned above. The detection rate was extremely low and the extraction of relations led to no results. Some approaches, like the one of [57] from 2013, are outdated as they fail to use state-of-the art pretraining to enrich their algorithms with adequate text representations such as word embeddings [9, 19, 51]. For instance, to find candidate phrases, we used spaCy [25] for dependency parsing and the results of [38] for constituency parsing. Moreover, we constructed machine learning algorithms for IOB tagging, of which our final approach was the best. Evaluation results will be presented in Sect. 5.

Literature indicates an adequate use of IOB tagging [16]. For our domain, this does not seem appropriate: We have long phrases with differing start and end words, including punctuation, due to the user-generated nature. Hence, the aspect phrases are not as predictable as named entities, such as: "*Mrs. Jeanette Muller*" and "*Jeanette Muller*", in comparison to our data (in German) "*Dr. Meyer hat sich einige Zeit genommen.*" (translated: "*Dr. Meyer took a lot of time.*"). This can also be compared to cases such as: "*Dr. Meyer nimmt sich für seine Patienten viel Zeit.*" (translated: "*Dr. Meyer takes a lot of time for his patients.*"). In the German phrase, "*for his patients*" also has to be annotated because it is located in the middle of a phrase [35, 36]. IOB tagging seems to be a non-perfect fit for challenging cases like ours. However, "*I*" is sufficient when the start of a phrase receives less importance: i.e.,

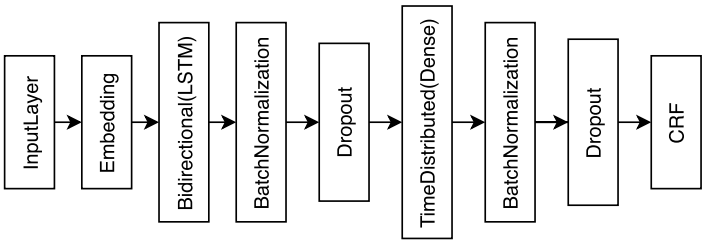

Fig. 3 BiLSTM-CRF model architecture [35, 36]

when the Beginning ("*B*") tag is left out. Hence, "*I*" and "*O*" are sufficient. In the end, every word is marked whether it belongs to an aspect phrase of a specific class or not. As experiments showed, using just "*I*" and "*O*" tags is sufficient and superior to using IOB tags.

For labeling sequential data such as text, studies like [71] propose an architecture composed of a Conditional Random Field (CRF) and a bidirectional Recurrent Neural Network (RNN) for feature extraction. We also tried to use additional features such as named entities, token lemmas, etc. But this approach proved to work not as well due to the use of user-generated content, which has various mistakes, while nouns are of limited importance to us. Furthermore, tests with Part-of-Speech (PoS) tags did not improve our results. Figure 3 shows the architecture of our system.

As Fig. 3 demonstrates, a bidirectional Long Short-Term Memory (biLSTM) [31] is used for feature extraction. This means that features are extracted in both directions from the text data and thus words before and after the current one are included in the calculations. A time-distributed dense layer then aligns the features and a CRF finds the optimal set of tags given for the words in a sentence. Using a biLSTM and CRF is state-of-the art [1]. In between, there are "BatchNormalization" layers in order to keep the activation values to a normalized score. We also use dropout[10] layers that regularize the data and thus hinder overfitting due to a limited dataset size (manual annotations require much human work). As input, we use sentences with vectorized tokens. For supervised learning, the tags resemble the following format: "*I-friendliness*" or "*O*" for a non-relevant word.

We trained our system to find aspect phrases and classes together, in the same step. We found that it is crucial to have pretrained vectors that embed common knowledge about words and subword units, so we trained vectors on all available reviews, not only the annotated ones. Further measures for embedding training focus on incorrectly split and spelled words by using only lowercase. Even though we have a comparatively limited amount of data, vectors with 300 dimensions worked the best. This dimensionality lead to no overfitting and increased the recall values,

[10] SpatialDropout was used only for the first dataset (fkza) and the biLSTM-CRF with FastText embeddings, while we relied on a regular dropout layer for the other cases. While SpatialDropout performed slightly better in this case, the overall effect was marginal and we thus trusted the normal Dropout, which is more suitable for sequence processing in general. The difference between Dropout and SpatialDropout is that the latter drops whole feature maps instead of single elements [72].

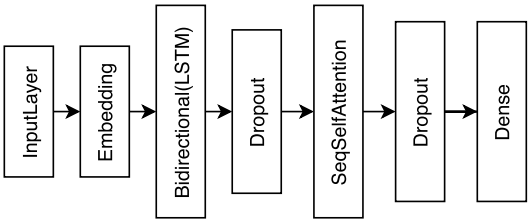

Fig. 4 BiLSTM-attention model architecture

which was especially the case when comparing them to a dimensionality of 25. We trained our own vectors using FastText [9]. When using other vectors apart from FastText, we spared the embedding layer and directly inserted the vectors to the model. However, as our user-generated data contained several mistakes, lowering the data and using the subword information, such as character n-Grams, helped achieve better representations. The Skipgram algorithm [9] learns vector representations for words that can predict their surrounding words. We allocated enough time for parameter tuning and testing a multitude of architecture configurations including CNNs, more RNN layers and other types of RNNs, with and without CRFs. The best results for the biLSTM-CRF model were reached with rather small values such as 0.3 for dropout, a unit size of around 30 for the biLSTM, and a small epoch and batch size. RMSprop also showed satisfactory results and we also tested current solutions such as BERT [19] (in its German, cased version by Deepset [17]). As BERT uses "wordpiece" tokens and not words, we use the weighted sum of the last hidden state [26]. The embedding layer in Fig. 3 contains all the vectors.

Another model architecture we tested performed well enough to be reported here (see Fig. 4). It consisted of a biLSTM and an attention layer, which are usually used together with recurrent neural networks [73] and calculates a weight (or attention) for data such as words [82]. Dependencies in sequence data can therefore be modeled without respect to the distance. We use multiplicative (dot-product) self-attention for sequence processing [69, 73]. However, using fewer layers and only a small portion of dropout worked here.

To achieve a better tested system, we also used different word embeddings. That is, we tested BERT vectors, as mentioned above. These were pretrained for German text, including capital letters (768 dimensions) [17]. As this model is well established, we fine-tuned it using physician reviews inspired by [4], thereby reaching a perplexity score of 3.78 and a loss of about 1.37. We used both language models to calculate the contextual input vectors (for results, see Table 4). We also used FLAIR embeddings [1] (also with uppercase letters) due to their contextual functionality, good performance results and the usage of characters, which makes them well suited for user-generated content. We also used BERT with uppercase letters because we wanted to benefit from the pretrained model available for the German language [17]. For FLAIR embeddings, we kept this configuration. We used a dimensionality of 768 to compare it to the BERT embeddings and trained RoBERTa embeddings [47] ourselves from scratch, using the physician reviews, but did not achieve usable results: i.e., the perplexity score of those embeddings remained high above 900.

5 Evaluation and Discussion

Tables 4, 5, 6, and 7 present our evaluation results such as precision, recall and F1 score per label as well as accuracy and an average per measure. We evaluated the two datasets each with a biLSTM-CRF and a biLSTM-Attention network using different word embeddings such as self-trained FastText, pretrained BERT, self-fine-tuned BERT and self-trained FLAIR embeddings. While our accuracy is always around 0.95 and thus high, we regard the F1 scores as more important. For example, in Table 4, we have an average F1 value of 0.80, which is not weighted. This score can be regarded as good.

This are better scores in comparison to [59] or [76]. Their domain uses a less complex wording and nouns usually describe most of the aspects. Furthermore, their models barely achieve a score of 0.50 and they perform aspect phrase extraction and classification in two steps instead of one, which propagates errors forward. We separated the classes with the "physician" aspect target into two datasets and thus trained models that recognize all the classes present in a dataset. However, there are also other approaches [71] that use trained models for each class. This may result in overlapping aspect phrase borders, with some words being labeled more than once for several classes. Our approach also achieves better results in bare numbers.

However, when taking a closer look at our results in Table 4, our precision scores are generally better than our recall scores. This is also consistent with Tables 5, 6, and 7. This may be caused by a limited amount of training data. As overfitting was an issue during model building and training, we therefore aimed at improving the recall results: Our model is relevant only in its practical application when it generalizes well to new data. Apart from our self-trained FastText vectors, we used BERT vectors, self-fine-tuned BERT vectors and self-trained FLAIR vectors. We expected generally and notably better results when using more sophisticated word

Table 4 Evaluation results of the biLSTM-CRF model for the first dataset (fkza) with different embeddings (except for BERT, the vectors were self-trained) [35, 36]

Vectors	FastText			BERT			own BERT			FLAIR		
Measures	P	R	F1	P	R	F1	P	R	F1	P	R	F1
I-explanation	0.81	0.71	0.76	0.73	0.67	0.70	0.75	0.65	0.69	0.82	0.44	0.57
I-friendliness	0.75	0.74	0.75	0.75	0.69	0.72	0.67	0.77	0.72	0.65	0.75	0.70
I-competence	0.68	0.67	0.67	0.69	0.65	0.67	0.84	0.53	0.65	0.94	0.38	0.54
I-time	0.85	0.80	0.82	0.87	0.77	0.82	0.85	0.80	0.82	0.82	0.78	0.80
O	0.97	0.98	0.97	0.97	0.98	0.97	0.96	0.98	0.97	0.95	0.99	0.97
Accuracy		**0.95**			**0.94**			**0.94**			**0.94**	
Average	**0.81**	**0.78**	**0.80**	**0.80**	**0.75**	**0.78**	**0.81**	**0.75**	**0.77**	**0.84**	**0.67**	**0.72**

Table 5 Evaluation results of the biLSTM-attention model for the first dataset (fkza) with different embeddings (except for BERT, the vectors were self-trained)

Vectors	FastText			BERT			own BERT			FLAIR		
Measures	P	R	F1	P	R	F1	P	R	F1	P	R	F1
I-explanation	0.80	0.60	0.69	0.71	0.64	0.67	0.73	0.63	0.68	0.74	0.61	0.67
I-friendliness	0.80	0.67	0.73	0.67	0.76	0.71	0.74	0.76	0.75	0.81	0.59	0.68
I-competence	0.77	0.58	0.66	0.70	0.63	0.66	0.77	0.61	0.68	0.73	0.55	0.62
I-time_taken	0.88	0.78	0.83	0.74	0.80	0.77	0.85	0.78	0.81	0.91	0.80	0.85
O	0.96	0.98	0.97	0.97	0.97	0.97	0.97	0.98	0.97	0.96	0.98	0.97
Accuracy		**0.95**			**0.94**			**0.95**			**0.94**	
Average	**0.84**	**0.72**	**0.78**	**0.76**	**0.76**	**0.76**	**0.81**	**0.75**	**0.78**	**0.83**	**0.71**	**0.76**

Table 6 Evaluation results of the biLSTM-CRF model for the second dataset (bavkbeg) with different vector representations (except for BERT, the vectors were self-trained)

Vectors	FastText			BERT			own BERT			FLAIR		
Measures	P	R	F1	P	R	F1	P	R	F1	P	R	F1
I-altern._healing_m.	0.78	0.71	0.75	0.65	0.71	0.68	0.80	0.60	0.68	0.85	0.49	0.62
I-treatment	0.71	0.71	0.71	0.70	0.58	0.63	0.84	0.63	0.72	0.75	0.52	0.61
I-care/commitment	0.73	0.62	0.67	0.77	0.50	0.61	0.82	0.65	0.73	0.79	0.32	0.45
I-overall/recom.	0.80	0.64	0.71	0.67	0.71	0.69	0.80	0.67	0.73	0.89	0.54	0.67
I-child-friendliness	0.76	0.81	0.79	0.77	0.81	0.79	0.75	0.81	0.78	0.64	0.75	0.69
I-relation._of_trust	0.83	0.79	0.81	0.83	0.80	0.81	0.83	0.80	0.81	0.95	0.59	0.73
O	0.96	0.97	0.97	0.96	0.97	0.96	0.96	0.98	0.97	0.94	0.99	0.97
Accuracy		**0.94**			**0.94**			**0.95**			**0.93**	
Average	**0.80**	**0.75**	**0.77**	**0.76**	**0.73**	**0.74**	**0.83**	**0.73**	**0.78**	**0.83**	**0.60**	**0.68**

vector calculation technologies such as BERT because these consider context, are state-of-the-art, and—in the case of BERT—were pretrained on larger quantities of data.

Still, our comparatively simple FastText embeddings achieved better recall and overall scores than in Table 4: The recall scores of 0.67 to 0.80 (and 0.98 for label "*O*") are favorable, especially when also reasoning about the F1 scores of 0.76, 0.75, 0.67, 0.82 and 0.97. In the domain and data that were used, these values are satisfying. We can explain the accuracy value of 0.95 with the high appearance of the label "*O*", as this boosts the accuracy score in general. We reduced the overweight of "*O*" labels by training our models only on sentences that contain an aspect phrase. It is crucial to have scores for precision and recall that are not too distinct. This is why when choosing the best model for the first dataset (fkza, Table 4 and 5), we regard the biLSTM-CRF model as superior to the biLSTM-Attention model due to the superior scores and because we regard FastText embeddings superior among the various word embedding configurations: The scores are higher yet not too distinct from each other.

Table 7 Evaluation results of the biLSTM-attention model for the second dataset (bavkbeg) with different vector representations (except for BERT, the vectors were self-trained)

Vectors	FastText			BERT			own BERT			FLAIR		
Measures	P	R	F1	P	R	F1	P	R	F1	P	R	F1
I-altern._healing_m.	0.78	0.69	0.73	0.70	0.67	0.68	0.73	0.61	0.66	0.93	0.54	0.68
I-treatment	0.71	0.69	0.70	0.84	0.63	0.72	0.82	0.65	0.72	0.85	0.57	0.68
I-care/commitment	0.63	0.69	0.66	0.71	0.54	0.62	0.79	0.56	0.65	0.66	0.58	0.61
I-overall/recom.	0.72	0.73	0.73	0.68	0.75	0.71	0.85	0.62	0.72	0.82	0.67	0.74
I-child-friendliness	0.77	0.75	0.76	0.79	0.77	0.78	0.80	0.84	0.82	0.79	0.72	0.76
I-relation._of_trust	0.77	0.80	0.79	0.80	0.79	0.80	0.86	0.76	0.81	0.85	0.79	0.81
O	0.97	0.97	0.97	0.96	0.97	0.97	0.96	0.98	0.97	0.96	0.98	0.97
Accuracy		**0.94**			**0.94**			**0.95**			**0.94**	
Average	**0.76**	**0.76**	**0.76**	**0.78**	**0.73**	**0.75**	**0.83**	**0.72**	**0.76**	**0.84**	**0.69**	**0.75**

For example, as Table 4 reveals, BERT embeddings [19] enable our model to achieve an F1 score of 0.67 for *competence*, the same as for our embeddings. While on average the precision scores are 0.80 compared to 0.81, the recall is lower with 0.75 to 0.78, so we prefer our model which we believe also reflects our user-generated data better. BERT embeddings are remarkable, though, because they are not trained on our domain and it is helpful to have embeddings calculated for every word in relation to its context words [19]. Yet, we must note that the fine-tuned embeddings actually perform slightly worse than embeddings that are not fine-tuned. We do not have an explanation for this, as we trained the embeddings for over 24 hours and saw a learning curve, evident in a decreasing loss.

Regarding the second dataset (bavkbeg) and the evaluation scores in Tables 6 and 7, we can also conclude that the biLSTM-CRF model using FastText vectors performs best. While obtaining an F1 score of 0.77, self-fine-tuned BERT embeddings achieve an F1 score of 0.78. However, FastText vectors achieve a smaller discrepancy of 0.05, compared to 0.10 in relation to precision and recall. The evaluation results in Table 7 are also almost as good. Here, the FastText vectors enable an F1 score of 0.76, which is balanced better because precision and recall also have a value of 0.76. Still, the biLSTM-CRF model performs slightly better. The attention model is simpler, though, as Figs. 3 and 4 demonstrate.

In this section, we have shown a number of approaches (along with evaluation scores) for automatically processing our annotated dataset and building a generalizing model that learns from the data to prove new ones. However, it is not possible to conduct a direct comparison to other models and datasets such as those of [59] or [76]. But a comparison to the presented values in studies dealing with shared tasks indicates the superiority of our approach. IO tagging with a biLSTM-CRF model improved our evaluation scores, but the self-trained word vectors were nevertheless important in achieving the final values shown in Tables 4, 5, 6, and 7. Numbers may not show the full picture, so we propose a manual evaluation as an additional step.

We therefore wrote sentences that we regard as fringe cases and cases that can be hard to tag. The results of our algorithm were positive.

In addition to these measures, we also annotated a dataset with higher inter-annotator agreement scores than Wojatzki et al. [76]. We have comparable Cohen's Kappa scores to [76], even though they do not clearly indicate scores for the aspect spans. We also used fewer human annotators. The inter-annotator agreement for aspects is 0.79–1.0. Pontiki et al. [59] utilize the F1 score for measuring the annotator agreement. However, this score is difficult to compare.

6 Conclusion

We presented two new datasets and several algorithms for ABSA by first introducing related literature and characterized numerous studies that tackle aspect extraction, classification and sentiment analysis in general. We then presented the data domain of physician reviews, where rating aspects are often implicitly mentioned or just indicated. Although we presented convincing algorithms and evaluation scores, there is still much work to be done before ABSA can be expanded from rather common review domains with a known and clearer vocabulary to fields that are more complicated and have a real-world use.

This study also provides a detailed presentation of our data, describing available rating classes on PRWs besides just providing general information. We then described the definition process for aspects and showed the classes to be annotated. We annotated four aspect classes for the first dataset (fkza): *friendliness*, *competence*, *time taken* and *explanation*, which are all equally related to physicians performing health-care services. The second dataset (bavkbeg) consists of six classes of which all but one apply to the physician: namely, *treatment*, *alternative healing methods*, *relationship of trust*, *child-friendliness*, *care/commitment* and *overall/recommendation* (which is a general class that does not involve the physician directly). The first dataset has 11,237 annotated sentences, the second 15,467. We plan to expand the process to other classes and opinion targets such as the doctor's office and the team until we have a holistic view over the entire domain of physician reviews and PRWs. To illustrate our findings, we provided example phrases and sentences, included comparisons to other datasets, and named details regarding the differentiation of classes. The inter-annotator agreement we calculated achieves favorable scores.

Our approach for extracting and classifying aspect phrases involves using a biLSTM-CRF model. We also tested a biLSTM-Attention model and various word embedding configurations, including transformers and subword information. To provide information for other scholars, we mentioned the difficulties involved when machine learning systems perform aspect extraction. In comparison to the work of other scholars, our approach conducts two steps in one: extracting and classifying the aspects phrases (their words). Even if the comparison is difficult to undertake, our datasets and models seem to outperform other approaches such as those of Pontiki et

al. [59]. This is impressive, because we consider our domain complex since it uses the morphologically complex German language.

For the future, we do not only plan to build more annotated datasets, but we also want to include opinion extraction: A possible method would be to build on existing knowledge such as integrating a second stack of neural network layers that learn from other annotated datasets. We could also annotate more sentences in our dataset in relation to their general sentiment. During the initial trial runs, it became evident that most sentences can be considered either positive or negative. A more fine-grained scale is not possible, because the wording does not exhibit a fine-grained gradation. For instance, the phrase "*He was **very/really/quite** friendly.*" will express the same information no matter which of the three adverbs is chosen, whether the adverbs are left out or whether two adverbs are applied. In theory, it would be possible to find guidelines that make a distinction possible, but this would be very subjective and not easy for human annotators to agree on. Moreover, such rules would be difficult to remember and reaching an agreement is challenging because of phrases such as: "*He was very friendly and competent.*" The moderating word "*very*" cannot be annotated twice with the guidelines and tool we use. So if this were the decisive factor between a positive and very positive phrase, we would have to rely on the fact that a neural network can consider the context accordingly. In this case, the word "*competent*" would be labeled as very positive. But in other cases where this word occurs again, but alone, it would be labeled as positive, not as very positive.

Acknowledgements This study is an invited, extended work based on [35]. Another related study is [36], which was written and submitted during the same period as [35]. This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Centre On-The-Fly Computing (SFB 901). We thank Rieke Roxanne Mülfarth, Frederik Simon Bäumer and Marvin Cordes for their support with the data collection.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638–1649. ACL, Santa Fe, NM, USA (2018). https://www.aclweb.org/anthology/C18-1139
2. Apotheke-Adhoc: Von Jameda zur Konkurrenz geschickt. [sent by Jameda to the competitors]. https://www.apotheke-adhoc.de/nachrichten/detail/apothekenpraxis/von-jameda-zur-konkurrenz-geschickt-bewertungsportale/ (2018). Accessed 28 Oct 2019
3. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th LREC, vol. 10, pp. 2200–2204. ELRA (2010)
4. Beltagy, I., Lo, K., Cohan, A.: SCIBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3615–3620. ACL (2019)
5. Bidmon, S., Elshiewy, O., Terlutter, R., Boztug, Y.: What patients value in physicians: Analyzing drivers of patient satisfaction using physician-rating website data. J. Med. Internet Res. **22**(2), e13830 (2020). https://doi.org/10.2196/13830

6. Bidmon, S., Elshiewy, O., Terlutter, R., Boztug Y.: What patients really value in physicians and what they take for granted: an analysis of large-scale data from a physician-rating website. J. Med. Internet Res. **22**(2), e13830 (2019). https://doi.org/10.2196/13830
7. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python, 1st edn. O'Reilly Media, Sebastopol (2009)
8. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G.A., Reynar, J.: Building a sentiment summarizer for local service reviews. In: Proceedings of the WWW Workshop on NLP Challenges in the Information Explosion Era, vol. 14, pp. 339–348. ACM (2008)
9. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. ACL **5**, 135–146 (2017)
10. Bäumer, F.S., Grote, N., Kersting, J., Geierhos, M.: Privacy matters: detecting nocuous patient data exposure in online physician reviews. In: Proceedings of the 23rd International Conference on Information and Software Technologies, vol. 756, pp. 77–89. Springer (2017). https://doi.org/10.1007/978-3-319-67642-5_7
11. Bäumer, F.S., Kersting, J., Kuršelis, V., Geierhos, M.: Rate your physician: findings from a Lithuanian physician rating website. In: Proceedings of the 24th International Conference on Information and Software Technologies, Communications in Computer and Information Science, vol. 920, pp. 43–58. Springer (2018). https://doi.org/10.1007/978-3-319-99972-2_4
12. 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). https://doi.org/10.1109/icosc.2015.7050774
13. Cieliebak, M., Deriu, J.M., Egger, D., Uzdilli, F.: A Twitter corpus and benchmark resources for German sentiment analysis. In: Proceedings of the 5th International Workshop on Natural Language Processing for Social Media, pp. 45–51. ACL (2017). https://doi.org/10.18653/v1/W17-1106
14. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
15. Cordes, M.: Wie bewerten die anderen? Eine übergreifende Analyse von Arztbewertungsportalen in Europa. [What do the others think? An overarching analysis of doctor rating portals in Europe]. Master's thesis, Paderborn University (2018)
16. 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). https://doi.org/10.18653/v1/w17-5218
17. deepset: deepset – open sourcing German BERT (2019). https://deepset.ai/german-bert. Accessed 28 Nov 2019
18. Deng, L., Wiebe, J.: MPQA 3.0: an entity/event-level sentiment corpus. In: Proceedings of the 2015 Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 1323–1328. ACL (2015)
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint (2018)
20. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K.: Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the ACL, pp. 49–54. ACL (2014)
21. Ellimoottil, C., Leichtle, S.W., Wright, C.J., Fakhro, A., Arrington, A.K., Chirichella, T.J., Ward, W.H.: Online physician reviews: the good, the bad and the ugly. Bull. Am. Coll. Surg. **98**(9), 34–39 (2013)
22. Emmert, M., Meier, F., Pisch, F., Sander, U.: Physician choice making and characteristics associated with using physician-rating websites: cross-sectional study. J. Med. Internet Res. **15**(8), e187 (2013)
23. Emmert, M., Sander, U., Esslinger, A.S., Maryschok, M., Schöffski, O.: Public reporting in Germany: the content of physician rating websites. Methods Inf. Med. **51**(2), 112–120 (2012)
24. Emmert, M., Sander, U., Pisch, F.: Eight questions about physician-rating websites: a systematic review. J. Med. Internet Res. **15**(2), e24 (2013). https://doi.org/10.2196/jmir.2360

25. ExplosionAI: Spacy (2019). https://spacy.io/. Accessed 06 Nov 2019
26. ExplosionAI: GitHub - explosion/spacy-transformers/ – spaCy pipelines for pre-trained BERT, XLNet and GPT-2 (2020). https://github.com/explosion/spacy-transformers. Accessed 20 May 2020
27. Gao, G.G., McCullough, J.S., Agarwal, R., Jha, A.K.: A changing landscape of physician quality reporting: Analysis of patients' online ratings of their physicians over a 5-year period. J. Med. Internet Res. **14**(1), e38 (2012). https://doi.org/10.2196/jmir.2003
28. Garcia-Pablos, A., Cuadros, M., Rigau, G.: W2VLDA: almost unsupervised system for aspect based sentiment analysis. Expert Syst. Appl. **91**, 127–137 (2018). https://doi.org/10.1016/j.eswa.2017.08.049
29. Geierhos, M., Bäumer, F., Schulze, S., Stuß, V.: "I grade what I get but write what I think." inconsistency analysis in patients' reviews. In: ECIS 2015 Completed Research Papers. AIS (2015). https://doi.org/10.18151/7217324
30. Hao, H., Zhang, K.: The voice of Chinese health consumers: A text mining approach to web-based physician reviews. J. Med. Internet Res. **18**(5), e108 (2016). https://doi.org/10.2196/jmir.4430
31. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
32. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
33. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of the 19th National Conference on Artificial Intelligence, pp. 755–760. AAAI (2004)
34. Kersting, J., Bäumer, F., Geierhos, M.: 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, pp. 147–155. SCITEPRESS (2019). https://doi.org/10.5220/0007745401470155
35. 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)
36. Kersting, J., Geierhos, M.: Neural learning for aspect phrase extraction and classification in sentiment analysis. In: Proceedings of the 33rd International Florida Artificial Intelligence Research Symposium (FLAIRS) Conference. AAAI (2020)
37. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1746–1751. ACL (2014)
38. Kitaev, N., Klein, D.: Constituency parsing with a self-attentive encoder. In: Proceedings of the 56th Annual Meeting of the ACL, vol. 1, pp. 2676–2686. ACL (2018)
39. Kordzadeh, N.: Investigating bias in the online physician reviews published on healthcare organizations' websites. Decis. Support Syst. **118**, 70–82 (2019). https://doi.org/10.1016/j.dss.2018.12.007
40. Krippendorff, K.: Computing Krippendorff's alpha-reliability. Technical report 1-25-2011, University of Pennsylvania (2011). https://repository.upenn.edu/asc_papers/43
41. Lagu, T., Norton, C.M., Russo, L.M., Priya, A., Goff, S.L., Lindenauer, P.K.: Reporting of patient experience data on health systems' websites and commercial physician-rating websites: mixed-methods analysis. J. Med. Internet Res. **21**(3), e12007 (2019). https://doi.org/10.2196/12007
42. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977). https://doi.org/10.2307/2529310
43. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
44. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)

45. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C.C., Zhai, C.X. (eds.) Mining Text Data, pp. 415–463. Springer, Berlin (2012)
46. Liu, J., Hou, S., Evans, R., Xia, C., Xia, W., Ma, J.: What do patients complain about online: a systematic review and taxonomy framework based on patient centeredness. JMIR **21**(8), e14634 (2019). https://doi.org/10.2196/14634
47. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR p, o. S. (2019)
48. Löpez, A., Detz, A., Ratanawongsa, N., Sarkar, U.: What patients say about their doctors online: a qualitative content analysis. J. Gen. Intern. Med. **27**(6), 685–692 (2012). https://doi.org/10.1007/s11606-011-1958-4
49. McAuley, J., Leskovec, J., Jurafsky, D.: Learning attitudes and attributes from multi-aspect reviews. In: Proceedings of the 12th IEEE International Conference on Data Mining, pp. 1020–1025. IEEE (2012). http://arxiv.org/pdf/1210.3926v2
50. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014). https://doi.org/10.1016/j.asej.2014.04.011
51. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR pp. 1–12 (2013)
52. Mitchell, M., Aguilar, J., Wilson, T., Van Durme, B.: Open domain targeted sentiment. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1643–1654. ACL (2013)
53. Mukherjee, A., Liu, B.: Aspect extraction through semi-supervised modeling. In: Proceedings of the 50th Annual Meeting of the ACL, vol. 1, pp. 339–348. ACL (2012)
54. Murphy, G.P., Radadia, K.D., Breyer, B.N.: Online physician reviews: is there a place for them. Risk Manag. Healthc. Policy **12**, 85–89 (2020)
55. 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)
56. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008). https://doi.org/10.1561/1500000001
57. Paul, M.J., Wallace, B.C., Dredze, M.: What affects patient (dis) satisfaction? analyzing online doctor ratings with a joint topic-sentiment model. In: Proceedings of the Workshops at the 27th AAAI Conference on Artificial Intelligence. AAAI (2013)
58. 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., Duchesnay, E.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
59. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryiğit, G.: SemEval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation, pp. 19–30. ACL (2016). http://www.aclweb.org/anthology/S16-1002
60. 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). http://aclweb.org/anthology/S/S15/S15-2082.pdf
61. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval 2016 task 5: aspect based sentiment analysis (ABSA-16) annotation guidelines (2016)
62. 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)
63. Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. Comput. Linguist. **37**(1), 9–27 (2011). https://doi.org/10.1162/coli_a_00034

64. Remus, R., Quasthoff, U., Heyer, G.: SentiWS - A publicly available German-language resource for sentiment analysis. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the International Conference on Language Resources and Evaluation, pp. 1168–1171. ELRA (2010). http://www.lrec-conf.org/proceedings/lrec2010/summaries/490.html
65. Ruppenhofer, J., Klinger, R., Struß, J.M., Sonntag, J., Wiegand, M.: IGGSA shared tasks on German sentiment analysis (GESTALT). In: Proceedings of the 12th KONVENS, pp. 164–173 (2014). http://nbn-resolving.de/urn:nbn:de:gbv:hil2-opus-3196
66. Ruppenhofer, J., Struß, J.M., Wiegand, M.: Overview of the IGGSA 2016 shared task on source and target extraction from political speeches. In: Proceedings of the IGGSA 2016 Shared Task on Source and Target Extraction from Political Speeches, pp. 1–9. Ruhr Universität Bochum, Bochumer Linguistische Arbeitsberichte (2016)
67. Saeidi, M., Bouchard, G., Liakata, M., Riedel, S.: SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In: Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1546–1556. COLING/ACL (2016)
68. Sharma, R.D., Tripathi, S., Sahu, S.K., Mittal, S., Anand, A.: Predicting online doctor ratings from user reviews using convolutional neural networks. Int. J. Mach. Learn. Comput. **6**(2), 149–154 (2016). https://doi.org/10.18178/ijmlc.2016.6.2.590
69. Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C.: Disan: Directional self-attention network for RNN/CNN-free language understanding. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. AAAI (2018)
70. Terlutter, R., Bidmon, S., Röttl, J.: Who uses physician-rating websites? Differences in sociodemographic variables, psychographic variables, and health status of users and nonusers of physician-rating websites. J. Med. Internet Res. **16**(3), e97 (2014). https://doi.org/10.2196/jmir.3145
71. Toh, Z., Su, J.: Nlangp at SemEval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In: Proceedings of the 10th International Workshop on Semantic Evaluation, pp. 282–288. ACL (2016). https://doi.org/10.18653/v1/s16-1045
72. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656. IEEE (2015). https://doi.org/10.1109/cvpr.2015.7298664
73. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems, pp. 5998–6008. Curran Associates (2017)
74. Vinodhini, G., Chandrasekaran, R.: Sentiment analysis and opinion mining: a survey. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **2**(6), 282–292 (2012)
75. Wallace, B.C., Paul, M.J., Sarkar, U., Trikalinos, T.A., Dredze, M.: A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J. Am. Med. Inform. Assoc. **21**(6), 1098–1103 (2014). https://doi.org/10.1136/amiajnl-2014-002711
76. 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. Springer (2017)
77. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. **13**(3), 55–75 (2018)
78. Zeithaml, V.: How consumer evaluation processes differ between goods and services. Mark. Serv. **9**(1), 186–190 (1981)
79. Zeithaml, V.A., Parasuraman, A., Berry, L.L., Berry, L.L.: Delivering Quality Service: Balancing Customer Perceptions and Expectations. Free Press (1990). https://books.google.de/books?id=RWPMYP7-sN8C
80. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. **8**(4), 1–25 (2018). https://doi.org/10.1002/widm.1253
81. Zhao, W.X., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a maxent-lda hybrid. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 56–65. ACL (2010)

82. 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). https://doi.org/10.1109/access.2019.2920075