Aspect Phrase Extraction in Sentiment Analysis with Deep Learning

Joschka Kersting and Michaela Geierhos
Semantic Information Processing Group, Paderborn University, Warburger Str. 100, Paderborn, Germany
{jkers, geierhos}@mail.upb.de

Keywords: Deep Learning, Natural Language Processing, Aspect-based Sentiment Analysis.

Abstract: This paper deals with aspect phrase extraction and classification in sentiment analysis. We summarize current approaches and datasets from the domain of aspect-based sentiment analysis. This domain detects sentiments expressed for individual aspects in unstructured text data. So far, mainly commercial user reviews for products or services such as restaurants were investigated. We here present our dataset consisting of German physician reviews, a sensitive and linguistically complex field. Furthermore, we describe the annotation process of a dataset for supervised learning with neural networks. Moreover, we introduce our model for extracting and classifying aspect phrases in one step, which obtains an F1-score of 80%. By applying it to a more complex domain, our approach and results outperform previous approaches.

1 INTRODUCTION

Aspect-based Sentiment Analysis (ABSA) aims at finding expressed opinions towards attributes of products or services. Sentiment analysis (or opinion mining) experienced a large interest during the last years due to a growing amount of user-generated content and missing methods to make use of them. However, the focus so far was on identifying an overall sentiment of full documents or sentences. This is, however, not sufficient when it comes to conflicting sentiments on different aspects of a product or service. Consequently, ABSA was invented in order to identify aspects in natural language texts and calculate the corresponding sentiment expressed by users. This has led to numerous studies (Liu and Zhang 2012; Sun, Huang, and Qiu 2019; Tang et al. 2016) and shared tasks (Wojatzki et al. 2017; Pontiki et al. 2015; Toh and Su 2016; Danda et al. 2017). Yet the research so far, has mostly neglected or omitted the existence of aspects represented implicitly by phrases rather than directly and only by nouns. What is more, in order to make use of the widely available reviews, an understanding of why users rate how is required (McAuley, Leskovec, and Jurafsky 2012). Hence, ABSA is an important field to be investigated.

There are three main approaches in sentiment analysis research. These are document, sentence-level as well as aspect-based sentiment analysis. While the first only covers an overall sentiment for a whole document, the latter supposes that there is just one sentiment expressed per sentence. Hence, cases are neglected in which even opposing sentiments for the same aspect or opinions regarding different aspects in a sentence are expressed. The following sentence serves as an example: "The doctor was **very friendly** but he **did not offer to shake my hand**." In this sentence, the friendliness of a physician is rated two times by the bold printed expressions. Finding an expression such as the latter may be easy for a human but hard for a machine.

1.1 Domain of Research

This paper deals with the domain of physician reviews. ABSA in this domain cannot be performed by keyword spotting, because of implicit mentions and the use of phrases for expressing opinions about the usually personal, trustful and sensitive services (Bäumer et al., 2017; Kersting, Bäumer, and Geierhos, 2019). However, most of ABSA research proposes that nouns are representative for aspects or at least settles with nouns (and noun phrases) explicitly mentioned as aspect indicators (Pontiki et al., 2016b; Nguyen and Shirai, 2015; Qiu et al., 2011; Hu and Liu, 2004; Blair-Goldensohn et al., 2008; Chinsha and Shibily, 2015). This has reasons: Many reviews are written about products or services. There exist search goods, i.e., products such as smartphones or keyboards (e.g., smartphone: battery, memory), which can be interchanged and will roughly be the

same every time. And there are experience goods, whose performance can only be evaluated after experiencing them due to their subjective, every time different nature (Zeithaml, 1981). However, most of ABSA research was done focusing on products (De Clercq et al., 2017) or such services using rather simple vocabulary. Even experience goods such as hotels receive reviews based on nouns describing the breakfast, bed, etc. These can be characterized as experience domains rather than services.

The area of physician reviews is characterized by experience goods, but its nature is that of very special services with private and personal components. Each treatment performed by a health care provider is unique. Personally provided services are usually reviewed on the basis of the behavior of the staff. Components of this are empathy, reliability, but also the ambiance of the rooms (Zeithaml et al. 1990). The reviews can be found on Physician Review Websites (PRW). Examples for PRWs are Ratemds[1] in English, Jameda[2] in German or Pincetas in Lithuanian[3]. On a PRW, users can rate their physician quantitatively using stars or grades and qualitatively by writing a review text. Typically, the grades can be assigned to different aspects such as the competence of the physician. Besides grading, the PRWs also offer blogging functionality, appointment services and more. However, while many physicians feel unfairly treated and to not want to be rated, legal repercussions are the consequences. Users feel anonymous even though they can be identified by the PRW or by the physician on the basis of the review texts. Trust is an important issue regarding PRWs (Kersting, Bäumer, and Geierhos 2019; Bäumer et al. 2017; Apotheke-Adhoc 2018).

1.2 Contributions

In general, ABSA has three subtasks of which this paper addresses two: Aspect term extraction and aspect category classification, but not aspect polarity classification (De Clercq et al., 2017). What is more, we contribute by advancing the field of ABSA with phrases implicitly indicating rating aspects. These phrases are often complex and in their form of appearance not frequent. Additionally, we use German as a complex and morphologically rich language. We use the service domain of physician reviews due to its broad variety of fields (professions and diseases) and sensitive, health-related nature. Here, we contribute by presenting our dataset in general and the annotated texts. Moreover, we present a neural network for aspect phrase extraction and evaluate it in one step without separating the step of identifying phrases and classifying them in contrast to the procedures used at shared tasks such as (Pontiki et al., 2016b).

The outline is as follows: The second section presents related literature and the third section our dataset. Here, we present the whole dataset as well as the aspect categories and our annotated data. The fourth section deals with our method and the implementation of a neural network aimed at identifying aspect phrases in German physician review texts. The fifth deals with the evaluation and discussion of our proposed dataset, domain and system. The last section concludes the paper and draws implications for future work.

2 STATE-OF-THE-ART

The identification and extraction of aspects[4] from texts is the core task in ABSA (Chinsha and Shibily, 2015). This distinguishes it from tasks such as classification where grades for a text document are predicted. Further steps are the identification of opinion words related to aspects and their polarity (positive or negative sentiment) (Chinsha and Shibily, 2015). Besides, sentiment analysis has to tackle other issues such as the detection of sarcasm or analyzing emotions (Zhang, Wang, and Liu 2018).

Hu and Liu (2004) published one of the earlier works on ABSA. Typical examples of such works deal with products such as smartphones: “The screen is perfect, but the voice quality sucks.” When this review is published on the online shop page of a smartphone model, recognizing the aspects is trivial, e.g., by focusing on noun phrases (Pontiki et al., 2016b; 2016a) or using topic modeling based on a list of seed words as it was performed by earlier studies (Mukherjee and Liu, 2012; Zhao et al., 2010). Some researchers stated in their annotation guidelines 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” (Pontiki et al., 2016a). They used annotated datasets in several languages (not German) in order to extract aspect terms, polarity, etc. for domains such as hotels or restaurants. For the extraction of aspect phrases and their classification, the results vary and are almost all below 50% (Pontiki

[1] Available at http://ratemds.com
[2] Jameda can be accessed at http://jameda.de
[3] Pincetas can be found at http://pincetas.lt
[4] Also referred to as feature, topic or target extraction or identification (Chinsha and Shibily, 2015).

et al., 2016b). Other researchers used constituency and dependency parsing in order to identify relevant words which are mostly nouns, too (Nguyen and Shirai, 2015). However, the problem with nouns is the assumption that aspects are mostly directly mentioned by words. There are explicit mentions of aspects that come in the form of nouns and noun phrases, and there are implicit mentions coming with every other possible construction such as adjectives, verbs, etc. For instance, "expensive" would refer to the price of a smartphone (Liu, 2012). The data of the current paper, which will be presented in the next section, contains implicit aspect mentions by all kinds of word types and forms. Some scholars go as far as extracting only the most frequent nouns and grouping them into synonym classes for the aspect phrase extraction (Chinsha and Shibily, 2015).

There are other works like Wojatzki et al., (2017), who use customer reactions from different channels such as twitter instead of reviews. They annotated a dataset in order to identify aspects in texts. And try to find the corresponding words or phrases. Still, they neither use customer reviews nor data from a sensitive domain such as the health care sector. What is more, all aspects in Wojatzki et al., (2017) are related to the main German railway company and are thus not as diverse as those related to all physicians, their professions, possible diseases and the sensitive patient-physician relationship (Kersting, Bäumer, and Geierhos, 2019). Many aspects can be described with nouns (e.g., atmosphere, train ride, connectivity). Hence, this approach touches the field of the present paper, but has still a different domain, approach and dataset. However, De Clercq et al. (2017) build an ABSA pipeline for Dutch retail, banking and human resources data in order to contribute to ABSA by using data form service domains. They rely on earlier studies that recommend seeing aspect term extraction as a sequential labelling task using Inside, Outside, Beginning (IOB) tags (Bird, Klein, and Loper 2009) for marking the beginning, inside and outside tokens of aspect phrases. They use more than 20 classes per domain. Their data are annotated manually. While they achieve very high scores for aspect term extraction, their category classification results are in part below 50%. According to their chosen domains and examples, the aspect terms seem to be nouns.

Other scholars follow an unsupervised path for building an ABSA system (Garcia-Pablos, Cuadros, and Rigau 2018; Mukherjee and Liu 2012; Zhao et al. 2010). Garcia-Pablos, Cuadros, and Rigau (2018) use a list of seed words in order to find aspects in large data quantities. A problem with such approaches is that topic models find clusters and topics that are not comprehensive for humans and thus miss the point for ABSA (Mukherjee and Liu, 2012).

When it comes to physician reviews and PRWs, there are numerous studies dealing with them in general and with their sensitive data (Emmert, Sander, and Pisch, 2013; Emmert et al., 2012; Bäumer et al., 2018; 2017; Kersting, Bäumer, and Geierhos, 2019). Medical diagnoses are among the hardest things to evaluate (Zeithaml, 1981) and physician reviews are utterly important for the choice of the right physician (Emmert et al., 2013), while most ratings are positive (Emmert, Sander, and Pisch, 2013). These studies underline the importance of physician reviews. What is more, physician reviews have a specific vocabulary and PRWs require trust (Kersting, Bäumer, and Geierhos, 2019).

3 DATA

The dataset consists of German-language physician reviews from several PRWs located in three German speaking countries (Austria, Germany, Switzerland).

3.1 Data Collection and Overview

Data collection took place in mid-2018, from March to July. For downloading the data, a distributed crawler framework was developed. At first, the websites were manually checked for an index. The index sites were collected in order to directly access the physician's (sub-)sites and save the corresponding reviews and ratings. We followed the rule of not causing too much traffic in order to keep the costs for the community as low as possible. Thus, data collection took several weeks (Cordes, 2018). All results were saved in a relational database.

We collected reviews, ratings and additional information concerning the physician and the office: e.g., opening hours, address and further training. We regard this information as useful for the future. Additionally, we collected data from an English, a Lithuanian and a Spanish PRW which enables qualitative comparisons, e.g., regarding the use of rating classes. The German-language PRWs are Jameda, Medicosearch [5] and Docfinder [6]. General statistics can be found in Table 1.

[5] Medicosearch can be reached at http://medicosearch.ch
[6] Docfinder can be found at http://docfinder.at

Table 1: Statistics for German-language PRWs.

PRW	Jameda	Docfinder	Medico-search
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%[7]	No Data
Length (Char.)	383	488	161

As can be seen, most physicians are listed on Jameda, while fewer are available on Docfinder and Medicosearch. However, the Austrian Docfinder and German Jameda are visited much more than Medicosearch, according to reviews. The ratings are, as stated before, very good on average. Interestingly, Jameda and Medicosearch have a higher number of professions, which may come from listing non-official professions and specializations. Non-German reviews were excluded for further steps.

3.2 Rating Categories

The aspect categories to be annotated were identified on the basis of qualitative methods. Here, we used the available categories that can be assigned on Jameda, Docfinder and Medicosearch, e.g., "friendliness", "explanation", etc. A set of rating categories were set up by discussing them in the team and semantically merging them. In Table 2, we present a selection of the categories given on the websites, which built the basis for our choice. Jameda has the most categories, but it presents only a selection for certain professions on its website (Table 2).

For the aspect term extraction and aspect category classification, we built a manually annotated dataset in order to use it for supervised machine learning. For this paper, a first set of categories was chosen which consequently was annotated in 11,237 sentences: "friendliness", "competence", "time taken", "explanation"[8]. All categories apply to the physician as the aspect target (/entity target/opinion target).

In most cases, systems combine aspect target extraction and aspect extraction (Zhang, Wang, and Liu 2018). In general, we identified three opinion targets in our data: the physician, the team and the doctor's office (e.g., "parking situation"). Additionally, we have the target of a general evaluation with only one aspect; an example sentence is the following: "Satisfied all round."

Table 2: Rating classes on PRWs (selection; translated from German) (Cordes 2018).

PRW	Rating Classes
Jameda	Treatment, counselling/care, commitment, discretion, explanation, relationship of trust, entering into concerns, time taken, friendliness, anxiety patients, waiting time for appointment, waiting time in the practice, opening hours, consultation hours, entertainment in the waiting room […]
Example: profession "dentist" on Jameda (limited classes available)	Treatment, explanation, mutual trust, time taken, friendliness, anxiety patients, waiting time for an appointment, waiting time in the practice, consultation hours, care, entertainment in the waiting room, alternative healing methods, child-friendliness, barrier-free access, practice equipment, 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 practice/premises, care by medical assistants, satisfaction in waiting time for 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 informed me comprehensively), information behavior (the declarations of the service provider were understandable for me), recommendation (the service provider has fulfilled my expectations), recommendation (I recommend this service provider)

The four named classes are distinguished clearly before we describe the annotation process in detail:

- "Friendliness" refers to the degree of devotion of the physician. That is, does the physician treat his/her patients respectfully and kindly, is he/she

[7] Only few data were available.

[8] All translated from German: "Freundlichkeit", "Kompetenz", "Zeit genommen", "Aufklärung."

nice or nasty when greeting them, does he/she look them in the eye? Examples (aspect phrases in bold, translated from German):

- "He was very friendly and his assistants are very efficient."
- "She did not even greet and did not listen."
- "Competence" deals with the (subjective) expertise of the physician. It asks whether the raters felt that the doctor knows what to do and how to do it. It does not ask for the treatment quality in general and it does not ask for friendliness, empathy, etc. The general competence is also included by this category. This means that whether a physician is good in his/her profession, "knows his job" or knows how to reduce anxiety is regarded as competence.
- "He was very competent and has a conscientious manner."
- "Time taken" refers to the time a physician uses in his appointments. Time is crucial for a treatment and the felt quality of a treatment. Patients may see it as positive, negative or even as a necessity that a health care provider takes enough time to treat them. In German, especially when it comes to this category, there are words which are between words such as "take" and "time", comparable to the English phrase:
- "She took a lot of time and […]."
- "The only drawback is that the practice is always overcrowded, so the personal consultation is rather short."
- "Explanation" describes the way the physician explains diseases and treatments to his patients. Here, patients want to be informed well. This class should be differed from the class "time taken" accordingly, as a long conversation indicates the amount of time used but the details mentioned or the questions the physicians asks deal with the explanation.
- "After a detailed clarification, the treatment was started immediately."
- "I was very well informed about my disease pattern, the consultation was excellent."

The classes can be distinguished clearly. In most cases, muli-word phrases need to be annotated, only nouns or single words do not indicate a category. However, there are cases in which a distinction between classes is not natural even for human beings and thus we provide the brief explanations above. However, for annotators, we created guidelines and documented cases that are on the edge. The annotation process is still very complicated because the phrases and the complexity of German sentence structure make it more difficult (e.g., word order may be changed quite freely).

3.3 Annotation Process

The process started by splitting the language reviews into sentences using the spaCy library (ExplosionAI, 2019). Annotation at the sentence-level instead of the document level is more efficient, especially when having aspects represented by complex phrases. What is more, Pontiki et al., (2016b) also annotated at the sentence level. However, we have roughly over 2 million sentences of reviews.

We then annotated 10,000 sentences to indicate whether they contain an evaluative statement. Based on this, after having achieved a high agreement among annotators, we built a Convolutional Neural Network (CNN) classifier in order to calculate a probability for every sentence to determine whether it contains an evaluation or not. We saved all sentences with a probability of over 50% randomized to a file and used them for the annotation of aspect phrases and their categories. We regarded this as a better approach compared to the use of seed words as performed by other scholars (Cieliebak et al., 2017). Our vocabulary, when it comes to the description of aspect classes, is complex and oftentimes consists of longer phrases. Thus, we expect a narrowed selection of sentences when using seed words for our dataset.

The dataset was annotated mainly by one person while two other people contributed. The annotators are all specialists who discussed and reviewed the annotations. Of 11,237 sentences, 6,337 contained at least one of the four classes, 4,900 did not. It was possible to annotate several aspects in one sentence. Our annotations are stored to a database and were exported into text files for further processing. Tokenization is saved here, too. The average sentence length in tokens can be found in Figure 1. Most sentences are short, while there is a certain number of longer ones as well.

The following sentence is a good example based on sentences from the dataset (translated from German): "**Competence [competence]** and **connectedness [friendliness]**, a good match: Dr. Meyer **knows what he is doing [competence]** and is **cordial [friendliness]** and **takes time [time taken]** for the patient, his **explanations are great [explanation]**." The aspect phrases are printed in bold, categories bold and in brackets. Here, the example delivers an idea of common phrases. Mostly, users write the way they speak. They rate the same thing with longer phrases or short words – often not

nouns – even several times in the same sentence. This is different in comparison to Pontiki et al. (2016b) and Wojatzki et al. (2017). However, compared to Pontiki et al. (2016b), our dataset is larger (e.g., for English in the laptop domain: 3,308 sentences; Dutch in the restaurant domain: 2,286 sentences). Furthermore, Pontiki et al. (2016a) includes only one of potentially several mentions of the same opinion target phrase or aspect entities. The dataset of Wojatzki et al. (2017), however, seemed larger at first to us, but reducing it to sentences with annotated aspect phrases reveals that it is only slightly larger than ours (roughly 2,000 sentences more).

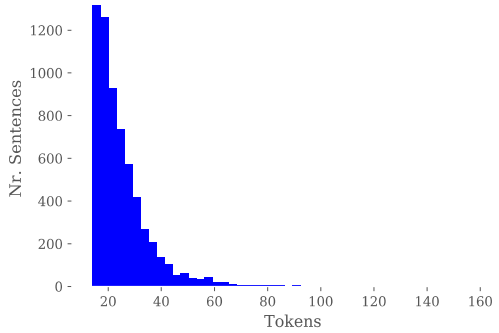

Figure 1: Number of tokens per sentence.

The annotation task was serious due to the nature of our data and we calculated the inter-annotator agreement on the basis of tagging, that is, every word received a tag with its class and every word not being annotated received a tag with a "None-class" (see Section 4). We randomly selected 337 (3%) of the data that were annotated by the main annotator before. Then, the other two, re-annotated them from scratch. We achieved sufficient agreement scores as can be seen in Table 3. We calculated Cohen's Kappa (Cohen 1960) for each two of three annotators using Scikit-learn (Pedregosa et al., 2011). The agreement among annotators is substantial with a minimum of 0.722 and a maximum between R and J of 0.857. According to Landis and Koch (1977), all values between 0.61 and 0.80 can be considered as substantial agreement, values above 0.81 as almost perfect. We regard those values as good for our dataset. What is more, we calculated Krippendorf's Alpha (Krippendorff, 2011) using NLTK (Bird et al., 2009) for all three annotators at once. Here, we achieve a score of 0.771 which can be seen as a good value, where 1.0 would be the best. Alpha provides several advantages such as calculating it for many annotators at once (not only two). Missing data and any number of categories can also be used (Krippendorff, 2011).

Table 3: Inter-annotator agreement between annotators R, B and J.

Annotators	**R & B**	**R & J**	**B & J**
Cohen's Kappa	**0.722**	**0.857**	**0.730**
Krippendorf's Alpha (**for all 3**)	**0.771**		

4 METHOD

In this section, we briefly describe our approach to perform aspect phrase extraction and the classification of aspect phrases on the basis of our annotated dataset. At first, we describe paths we followed in search for a working system. We scanned the literature for building the ideal extraction system. For example, Liu (2012) proposes four approaches to extract aspects: Extraction (1) using frequent noun (phrases), (2) by making use of opinion and target relations (3) by supervised learning (4) based on topic modeling (Liu, 2012). We tried them and had to conclude that only supervised approaches are promising. This came from test results as well as from the related literature section. For instance, topic modeling did not find clearly separated topics as humans would define them. Frequent nouns lead to an extremely low detection rate for aspects and the extraction of relations produced no usable results. We used spaCy (ExplosionAI, 2019) for dependency parsing and results of Kitaev and Klein (2018) for constituency parsing to find candidate phrases. Moreover, we built several machine learning architectures for IOB tagging of which our final approach was the best.

Literature indicated a superiority of IOB tagging (De Clercq et al., 2017). This seems unsuitable for our case, as we have long phrases with differing start words that may not be as predictable as in named entities. Examples are "Mr John Doe" and "John Doe" in comparison to our data (in German): "Dr. Müller hat sich viel Zeit genommen" (translated: "Dr. Müller took a lot of time") in comparison to "Dr. Müller nimmt sich für seine Patienten viel Zeit." (translated: "Dr. Müller takes a lot of time for his patients."; In German, "for his patients" must be annotated along as it stands in the middle of the phrase.) However, while IOB tagging does not fit, the idea is sufficient when leaving out the Beginning (B) tag in favor of only I and O. We tried both and the binary IO tagging proved to be the best solution. When it comes to sequential labeling tasks, studies suggest using a Conditional Random Field (CRF) in combination with a bidirectional Recurrent Neural Network (RNN) (Toh and Su, 2016) that extracts

features, which we did. We did not use additional features as mentioned by other scholars, e.g., named entity information or token lemmas, as we rely on user-generated content that has too many mistakes and nouns are not dominant for us. Nevertheless, tests with Part-of-Speech tags and other common features did not improve our results. The architecture of our system can be found in Figure 1.

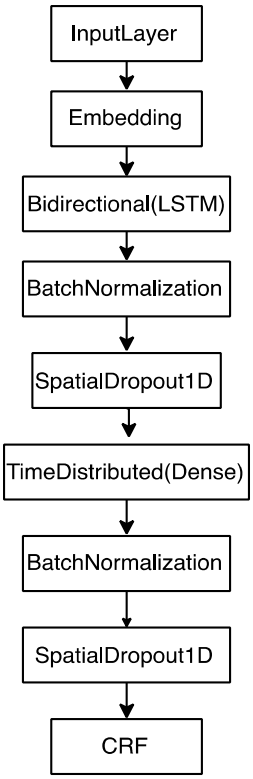

Figure 2: Model architecture.

Our architecture builds mainly on a bidirectional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) for extracting features from the sequential text data in both directions, using words before and after the current one. A time-distributed dense layer aligns all those features, before we hand them over to a CRF that considers the whole sentence in order to assign tags. The "BatchNormalization" layers are meant to keep the activation smaller, i.e., normalized. The dropout layers are used to prevent overfitting as our manually annotated dataset is relatively small. The input consists of sentences whose tokens were vectorized. At first, we used our tokens together with their tags in the form of "I-friendliness" or "O" for a non-relevant word.

That is, we directly trained the system for detecting aspect phrases together with their category. Secondly, it was crucial to have pretrained vectors. We trained our vectors on all of our sentences with further measures for avoiding incorrectly split words and using only lowercase. Interestingly, vectors with 300 dimensions turned out to work best for the model. This dimensionality helps to avoid overfitting and increasing recall, especially in comparison to a smaller dimensionality such as 25. What is more, current solutions such as BERT (Devlin et al., 2018) (in its German, cased version by Deepset (2019)) were outperformed by our vectors, which may be caused by the user generated nature of our data, its specific vocabulary and the better knowledge embedded in our vectors. The embedding layer in Figure 1 does contain all vectors. We trained our own vectors using FastText (Bojanowski et al., 2017). As user-generated content contains a lot of mistakes, we lowercased it to exclude such errors. Our embeddings are enriched with subword information (character n-Grams), which is helpful when dealing with user-generated content to cover mistakes. We used the skipgram algorithm proposed by Bojanowki et al., (2017). It learns vector representations of words that can predict words appearing in the context. We spent time for parameter tuning and testing other model architectures using CNNs, more RNN layers, other types of RNNs, models without CRF layers, etc. Our parameters showed best results with values such as a dropout of 0.3, a small unit size of 30 in the LSTM layer, RMSprop as optimizer, a small epoch size due to a small dataset and a batch size of about 10.

5 EVALUATION AND DISCUSSION

Table 2 presents our evaluation results such as precision, recall, F1-score per label as well as accuracy and an average per measure. While our accuracy of 0.95 is high, we regard our F1-score as more important. The F1-value of 0.80 is unweighted and can be regarded as good, especially in comparison to results in Pontiki et al., (2016b) or Wojatzki et al., (2017) who barely reach values of 0.50 in a domain with less complex wording and language while they separate extraction of phrases and classification of them which leads to forward propagation of errors. Furthermore, authors such as Toh and Su (2016) trained separate models for each category which may leads to better results, but we trained a unified model that obtains superior scores. We also did this in order to do not get overlapping aspect phrases for different categories.

Table 4: Evaluation results[9] of our model (self-trained and BERT embeddings.

Measures	P	R	F1	P (B)	R (B)	F1 (B)
I-explanation	0.81	0.71	0.76	0.73	0.67	0.70
I-friendliness	0.75	0.74	0.75	0.75	0.69	0.72
I-competence	0.68	0.67	0.67	0.69	0.65	0.67
I-time taken	0.85	0.80	0.82	0.87	0.77	0.82
O	0.97	0.98	0.97	0.97	0.98	0.97
Accuracy			**0.95**			**0.94**
Average	**0.81**	**0.78**	**0.80**	**0.80**	**0.75**	**0.78**

However, our precision scores are generally better than our recall scores. We think that this comes from a rather small amount of annotated training data. During training, overfitting was an issue and thus it was a goal to improve the recall: We want our model to be applicable to new data and therewith contain a relevance in application. Next to our self-trained embeddings, we tried BERT embeddings. Still, our own embeddings achieve better recall and overall scores: The current recall values of 0.67 to 0.80 (and 0.98 for label "O") are regarded as favorable, especially when taking into consideration that F1-scores of 0.76, 0.75, 0.67, 0.82 and 0.97 are more than satisfying when considering the domain and data. The accuracy shows a very high value of 0.95 which can be explained by the fact that generally, the "O" label appears mostly and thus boosts the accuracy score, thus we relate on F1. Furthermore, we regard it as crucial that precision and recall are not too far apart. This is the reason why we prefer our model with embedding layer and self-trained word vectors over BERT vectors. As Table 4 reveals, BERT embeddings (Devlin et al., 2018) 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. This is why we prefer our model which, as we think, also reflects our user-generated data better. However, BERT embeddings are remarkable while being not trained on our domain and it is helpful to have embeddings that are calculated for every word depending on its context words (Devlin et al., 2018).

To discuss our evaluation scores, it can be said that a direct comparison to other models and studies is not possible. This comes from the dataset we built and presented earlier in the third section. However, a comparison as indicated to the commonly presented values in studies dealing with shared tasks and their numerous results achieved in them indicates the superiority of our approach. While IO tagging combined with an LSTM-CRF model lead to success, self-trained word vectors finally enabled the evaluation scores in Table 2. Numerical scores can be misleading. Hence, we regard a manual evaluation as crucial. We wrote several sentences that we regard as edge-cases and cases that can be hard to classify in general. However, the aspect extraction and classification performed by our model is more than satisfying.

Additionally, we annotated a dataset with high inter-annotator agreement scores. Having used fewer human resources, we achieved comparable Cohen's Kappa scores to Wojatzki et al., (2017), even though they do not clearly indicate scores for the aspect spans. Their inter-annotator agreement for aspects lies between 0.79 and 1.0. Pontiki et al., (2016b) use the F1-score for the annotator agreement. We consider this score to be difficult to compare.

6 CONCLUSION

At first, we introduced the topic of ABSA. Here we indicated issues that are currently not sufficiently addressed. What is more, we enhanced this understanding in the literature section by presenting current approaches and general ideas from the area of ABSA. However, there is still much work to be done in order to expand research from common reviews for products and services to linguistically cover more complex review areas and languages, before ABSA can serve for further domains. Then, we present our data. Here, we collected a large number of review texts of which we use the German-language texts to train word embeddings and extract a number of sentences that were annotated. We annotate four classes related to a physician performing a health care service: "friendliness", "competence", "time taken", "explanation." Our annotated dataset currently consists of 11,237 sentences. We plan to further expand the process to other categories and opinion targets such as the doctor's office and the team. We also provided examples and comparisons to other datasets, named details of the category differentiation and calculated an inter-annotator agreement that achieves good scores.

We then described our method for extracting and classifying aspect phrases. This includes the model building process as well as the process of parameter tuning and details regarding word embedding training. However, we mention what has not led to success as well as difficulties for aspect extraction performed by a machine learning system.

[9] P = Precision, R = Recall, F1 = F1-score, B = BERT

Lastly, we evaluate and discuss our performance scores. Here, we compare our model with self-trained word embeddings to BERT embeddings. In opposite to other scholars, our approach performs two steps in one: aspect phrase extraction and classification. Nevertheless, we outperform other approaches such as Pontiki et al., (2016b) though we use another domain and dataset. However, we regard this domain as a complex one using the morphologically rich German language.

In the future, we not only plan to build more annotated datasets, but want to include the opinion extraction part, too.

ACKNOWLEDGEMENTS

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Centre On-The-Fly Computing (SFB 901).

REFERENCES

Apotheke-Adhoc. 2018. *Von Jameda zur Konkurrenz geschickt. [Sent by Jameda to the competitors]* URL: https://www.apotheke-adhoc.de/nachrichten/detail/apothekenpraxis/von-jameda-zur-konkurrenz-geschickt-bewertungsportale/ (visited on 10/28/2019).

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media, 1. Edition.

Blair-Goldensohn, S.; Hannan, K.; McDonald, R.; Neylon, T.; Reis, G. A.; and Reynar, J. 2008. Building A Sentiment Summarizer For Local Service Reviews. In *Proceedings of the WWW Workshop NLP Challenges in the Information Explosion Era*, 14, 339–348. ACM.

Bäumer, F. S.; Grote, N.; Kersting, J.; and Geierhos, M. 2017. Privacy Matters: Detecting Nocuous Patient Data Exposure In Online Physician Reviews. In *Proceedings of the 23rd International Conference on Information and Software Technologies,*, 756, 77–89. Druskininkai, Lithuania: Springer.

Bäumer, F. S.; Kersting, J.; Kuršelis, V.; and Geierhos, M. 2018. 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*, 920, 43–58. Vilnius, Lithuania: Springer.

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

Cieliebak, M., Deriu, J. M., Egger, D. & Uzdilli, F. (2017). A Twitter Corpus and Benchmark Resources for German Sentiment Analysis. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*. 45–51. Valencia, Spain: ACL.

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*. 24–31. Anaheim, CA, USA: IEEE.

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

Cordes, M. 2018. *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 Thesis. Universität Paderborn.

Danda, P.; Mishra, P.; Kanneganti, S.; and Lanka, S. 2017. IIIT-H at IJCNLP-2017 Task 4: Customer Feedback Analysis Using Machine Learning and Neural Network Approaches. In *Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks,* 155–160. Taipei, Taiwan: AFNLP.

Deepset. 2019. *Deepset – Open Sourcing German BERT*. URL: https://deepset.ai/german-bert. / (visited on 11/28/2019).

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*. 136–142, Copenhagen, Denmark: ACL.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*.

Emmert, M.; Sander, U.; Esslinger, A. S.; Maryschok, M.; and Schöffski, O. 2012. Public Reporting In Germany: The Content of Physician Rating Websites. *Methods of Information in Medicine.* 51(2):112–120.

Emmert, M.; Meier, F.; Pisch, F.; and Sander, U. 2013. Physician Choice Making and Characteristics Associated With Using Physician-Rating Websites: Cross-Sectional Study. *Journal of Medical Internet Research* 15(8):e187.

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.

ExplosionAI. 2019. *Release de_core_news_sm-2.1.0 explosion/spacy-models GitHub*. URL: https://github.com/explosion/spacy-models/releases//tag/de_core_news-sm-2.1.0 (visited on 11/13/2019).

Garcia-Pablos, A.; Cuadros, M.; and Rigau, G. 2018. W2VLDA: Almost Unsupervised System for Aspect-based Sentiment Analysis. *Expert Systems with Applications* 91:127–137.

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

Hu, M., and Liu, B. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, 168–177. Seattle, WA, USA: ACM.

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*, 147–155. Heraklion, Greece: SCITEPRESS.

Kitaev, N., and Klein, D. 2018. Constituency Parsing with a Self-attentive Encoder. In *Proceedings of the 56th Annual Meeting of the ACL*, 2676–2686. Melbourne, Australia: ACL.

Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. University of Pennsylvania. Retrieved from https://repository.upenn.edu/asc_ papers/43

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

Liu, B., and Zhang, L. 2012. A Survey of Opinion Mining and Sentiment Analysis. In Aggarwal, C. C., and Zhai, C. X., eds., *Mining Text Data*. 415–463. Springer.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5(1):1–167.

McAuley, J.; Leskovec, J.; and Jurafsky, D. 2012. Learning Attitudes and Attributes From Multi-Aspect Reviews. In *Proceedings of the 12th IEEE International Conference on Data Mining*, 1020–1025. Brussels, Belgium: IEEE.

Mukherjee, A., and Liu, B. 2012. Aspect Extraction Through Semi-Supervised Modeling. In *Proceedings of the 50th Annual Meeting of the ACL*, 1, 339–348. Jeju, South Korea: ACL.

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*, 2509–2514. Lisbon, Portugal: ACL.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … 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*, 486–495. Denver, CO, USA: ACL.

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

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.; and Eryiğit, G. 2016b. Semeval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, 19–30. Denver, CO, USA: ACL.

Qiu, G.; Liu, B.; Bu, J.; and Chen, C. 2011. Opinion Word Expansion and Target Extraction Through Double Propagation. *Computational Linguistics* 37(1):9–27.

Sun, C.; Huang, L.; and Qiu, X. 2019. Utilizing BERT for Aspect-based Sentiment Analysis Via Constructing Auxiliary Sentence. *arXiv preprint*.

Tang, D.; Qin, B.; Feng, X.; and Liu, T. 2016. Effective LSTMs for Target-Dependent Sentiment Classification. In *Proceedings of the 26th COLING*. Osaka, Japan: ICCL.

Toh, Z., and Su, J. 2016. 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*. 282–288. San Diego, CA, USA: ACL.

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*, 1–12. Berlin, Germany: Springer.

Zeithaml, V. A.; Parasuraman, A.; Berry, L. L.; and Berry, L. L. 1990. *Delivering Quality Service: Balancing Customer Perceptions and Expectations*. Free Press.

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.

Zhao, W. X.; Jiang, J.; Yan, H.; and Li, X. 2010. Jointly Modeling Aspects and Opinions With A Maxent-LDA Hybrid. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 56–65. Cambridge, MA, USA: ACL.