The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Neural Learning for Aspect Phrase Extraction and Classification in Sentiment Analysis

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

Abstract

In this study, we present an approach and a dataset for aspect-based sentiment analysis, showing how we extract and classify aspect phrases. The research field of aspect-based sentiment analysis aims at finding opinions expressed for individual characteristics of products or services in natural language texts. In the literature, reviews for common products or services such as smartphones or restaurants were mostly investigated. We describe our newly annotated dataset of German physician reviews, which presents a sensitive and linguistically complex domain, taking care to describe the annotation process and the functionality of our neural network approach. Finally, we introduce a model that can extract and classify aspect phrases in one step while obtaining an F1 score of 80%. As we employ our algorithm in a more complex domain, we believe that our study outperforms other studies.

Introduction

Aspect-based Sentiment Analysis (ABSA) is a field that analyzes written evaluations concerning elements of services or products. In general, the field of sentiment analysis is being investigated at an increasing rate, due to the large amounts of data that are available on the internet. In contrast, there are no methods to utilize written texts, such as user reviews: the research mostly covers overall evaluations of full documents, which is not appropriate when it comes to the conflicting polarities of specific properties that rated goods have. ABSA was invented because of lack leading to several studies (Sun, Huang, and Qiu 2019; Tang et al. 2016) and shared tasks (Wojatzki et al. 2017; Pontiki et al. 2015). Up till now, though research has not covered aspects that are only implicitly mentioned by phrases and not explicitly by just nouns.

Domain of Research. The domain of this paper covers physician reviews. Here, ABSA cannot be performed by extracting keywords, due to the implicit nature of the aspect phrases that express trustful and sensitive topics concerning one's health and to the relationship between physician and patient that determines the reviewed services (Bäumer et al. 2017; Kersting, Bäumer, and Geierhos 2019). Still, ABSA studies mostly suggest that nouns are the representative form for aspects in natural language texts, or at least in texts that only use nouns and noun phrases. What is more, studies suggest that these nouns explicitly indicate aspect classes (Pontiki et al. 2016b; Nguyen and Shirai 2015). Most reviews are written about products or common services and thus nouns may seem sufficient.

Two different types of goods are involved here. Experience goods are products or services that can be evaluated only after having experienced them, due to their individual, subjective nature because they are different for each performance. Search goods are those goods that can be interchanged and will be the same every time, such as TVs or smartphones (smartphones have a *battery, memory, etc.*) (Zeithaml 1981).

So far, research conducted in the field of ABSA has mainly been aimed at reviews of products (De Clercq et al. 2017) and services with a limited vocabulary. Reviews from the domain of experience goods also use many nouns: For example, when an experience domain such as a hotel is reviewed, the bed of the breakfast is often mentioned. Services performed by healthcare practitioners are unique by default, in contrast to personally performed services which are commonly rated on the behavior of the employed people (Zeithaml et al. 1990). Reviews for healthcare providers are written on physician review websites (PRWs), such as Ratemds[1] in English and Jameda[2] in German. On these PRWs, users can leave quantitative grades such as stars and also write qualitative comment texts. They expect to be anonymous on these platforms, although the PRW as well as the healthcare provider can clearly identify them through the review text. Many providers therefore feel unfairly treated and choose to use legal options. In general, trust is an important issue when it comes to PRWs (Kersting, Bäumer, and Geierhos 2019; Bäumer et al. 2017). This paper deals with qualitative review texts.

Contributions. ABSA involves three tasks: namely, aspect term extraction, aspect category classification and as-

*Corresponding Author

[1] Available at http://ratemds.com.
[2] Jameda can be accessed at http://jameda.de.

pect polarity classification (De Clercq et al. 2017). This current paper addresses two out of these three tasks, as we omit aspect polarity classification. Our contribution to the field of studying ABSA is through investigating phrases that indirectly hint to rating aspects. Such constructions are complex and long, as they involve insertions and are also not frequent. What is more, we use German instead of English data. The field of physician reviews involves a large number of professions and diseases and has very sensitive, health-related characteristics. Further contributions are our dataset, which consists of raw sentences and a number of manually annotated sentences, and our neural network for aspect phrase extraction, which we have evaluated extensively. With these contributions it is possible to identify phrases and classify them in one step without separation, in comparison to shared tasks such as (Pontiki et al. 2016b).

Our paper is organized as follows: The next section introduces the dataset and the aspect classes, together with our annotated data. We then describe our method and how the neural network was implemented to find German aspect phrases in physician reviews. After that, we discuss and evaluate our findings: namely, our data, domain and system. In the final section, we conclude our paper and point towards future work.

Data

Our dataset is based on German-language physician reviews from three PRWs based in countries where German is primarily spoken: Jameda (Germany), Medicosearch[3] (Switzerland) and Docfinder[4] (Austria). We crawled the data in the summer of 2018 and included review texts, ratings and additional information such as opening hours, the professions of the physician, opening hours, etc. General statistics can be found in Table 1.

The highest number of healthcare providers can be found on Jameda in Germany, while there are fewer on Medicosearch and Docfinder. When considering the number of physicians and reviews, Jameda and Docfinder are frequented more often than Medicosearch. The average ratings are very high. Both Jameda and Medicosearch have a very high number of listed professions, perhaps because they include non-official professions. We deleted reviews written in languages other than.

Annotations were carried out on the sentence level, as this procedure is more efficient, consistent with Pontiki et al. (2016b) and especially applies when the aspect phrases are rather long and complex. For the extraction and classification of aspect phrases, we manually annotated the words in the phrases.

In our study, qualitative methods were used in order to find the aspect categories that can be marked in the sentences. As a basis, we used the categories that users can assign grades for on the PRWs, such as "competence" and "time taken." We discussed the classes in the team and semantically merged similar classes to reach a set of final

[3]Medicosearch can be found at http://medicosearch.ch.

[4]Docfinder can be reached at http://docfinder.at.

[5]Only few data were available.

Table 1: Statistics for German-language PRWs.

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%[5]	No Data
Length (Char.)	383	488	161

categories. For this paper, we chose a first portion of the classes, covering the four categories of "friendliness", "competence", "time taken" and "explanation." In total we annotated 11,237 sentences. In line with studies that perform one step for aspect extraction and aspect target extraction (Zhang, Wang, and Liu 2018), we also combined this step. All annotated classes aim at the physician as the aspect target and we found three different aspect targets in our dataset: the physician, the doctor's office (e.g. the parking situation) and the team. We also constructed an overall evaluation as both the aspect and the target for cases in which patients write sentences such as "Satisfied all round." Three persons (all specialists) were involved in the process: one person annotated, while the two others provided assistance. From the general 11,237 sentences, 6,337 sentences contained an evaluative statement concerning the aspect classes, 4,900 did not. In one sentence, more than one annotation was possible, even for the same category.

The following example shows an English translation of a common review from our German language data: "***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 shown in bold, classes both in bold and in brackets. One issue to handle is that the reviews contain only user-generated content. Reviewers may rate the same aspects with either long phrases and descriptions or with just one word, sometimes more than once in the same sentence. In most cases, however, nouns are not used. This difference is apparent when we compare our work to the studies of Wojatzki et al. (2017) and Pontiki et al. (2016b). For example, our dataset contains more sentences than that of Pontiki et al. (2016b): for example, the English-language laptop topic covers 3,308 sentences and the Dutch-language restaurant topic involves 2,286 sentences. Furthermore, Pontiki et al. (2016a) include only one of possibly several appearances of an aspect phrase.

Method

In this section we present our method for extracting aspect phrases and classifying them in one step by using our annotated data. Before we describe our system, we name steps that did not work, following suggestions from the literature. Liu (2012) proposes the following four methods for conduct-

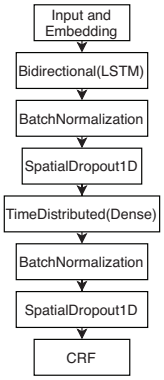

Figure 1: Model Architecture.

ing aspect extraction: (1) extracting frequent noun (phrases), (2) making use of opinion and target relations, (3) applying supervised learning and (4) running topic modeling. We tried all of these ways, but except for supervised learning, all three other approaches failed. This is suggested by relevant literature as well as by experiments we conducted: for example, with topic modeling we found topics that are not separated from each other. The extraction of relations led to no results and the use of frequent nouns and other words extracted very few samples. We also built machine learning algorithms for IOB tagging.

Scholars such as De Clercq et al. (2017) suggest a superiority of IOB tagging. An issue here is that we have long phrases, often with different start words in contrast to named entity recognition, for example. We have collected examples to illustrate this: "*Mr John Doe*" and "*John Doe*" contrasts our German data of "*Dr. Müller hat sich **viel Zeit genommen***" (English translation: "*Dr. Müller **took a lot of time***") and "*Dr. Müller **nimmt sich für seine Patienten viel Zeit***." (English translation: "*Dr. Müller **takes a lot of time for his patients***."; note that in German "*for his patients*" must be annotated too, as it stands in the middle of the phrase.) After omitting the B-tag in order to have only inside (I) and outside (O) tags, our approach started to work properly. We use the I-tags together with the labels. Furthermore, the literature declares that the technology of a Conditional Random Field (CRF) is favorable for sequential labeling tasks. This method is combined with a bidirectional Recurrent Neural Network (RNN) (Toh and Su 2016) for feature extraction. We used this architecture without further features, such as named entities and lemmas, because we have user-generated content with too many mistakes, such as typos. We also conducted experiments with part-of-speech tags and other common features, which did not improve our results. The architecture of our approach is displayed in Figure 1.

As can be seen in Figure 1, our method features a bidirectional Long Short-Term Memory (LSTM), which we regard as a crucial part. It extracts features from the input data in two directions and can thus monitor text content from the beginning and end of a sentence at once. After being processed by the LSTM, a time-distributed dense layer aligns the data that will be considered fully by the CRF: that is, the CRF uses the full sentence for assigning the tags. "BatchNormalization" layers keep the activation down (normalized). Dropout layers prevent overfitting, because our annotated dataset is quite small. The input consists of vectors calculated from the tokens. We trained our system to detect aspect phrases together with their category by using tags such as "I-friendliness" or "O" for a non-relevant word. Having pretrained vectors is important as this causes an increase in the performance of our algorithm. We trained our vectors on all of our sentences, not only on the annotated ones. Interestingly, a dimensionality of 300 performed best because it decreased overfitting and increased recall. The embedding layer in Figure 1 contains all vectors.

Table 2: Results of the Evaluation of our Model.

Measures	**Precision**	**Recall**	**F1 score**
I-explanation	.81	.71	.76
I-friendliness	.75	.74	.75
I-competence	.68	.67	.67
I-time_taken	.85	.80	.82
O	.97	.98	.97
Accuracy			**.95**
Average	**.81**	**.78**	**.80**

We invested time in parameter tuning and testing different model setups, using convolutional layers, for example. Our parameters ended in the best performance with values such as a dropout of 0.3, a small unit size of 30 in the LSTM layer, RMSprop as the optimizer, a small epoch size and a batch size of around 10.

Evaluation and Discussion

This section presents our discussion, demonstrating our evaluation scores in Table 2, which displays precision, recall and F1 score per label as well as the overall accuracy and averages of the named scores. Our system obtains a high accuracy score of 0.95. However, we regard this as less important in comparison to the scores per label, especially the F1 score which is not weighted. The F1 score on average per label is very good, with a value of 0.80 in comparison to Pontiki et al. (2016b) or Wojatzki et al. (2017) who achieve values of roughly 0.50 on their own datasets while having a less complex wording. In contrast, we do not separate the extraction and classification of aspect words and thus avoid the forward propagation of errors. We also trained one model for finding all categories at once, while other scholars such as Toh and Su (2016) have trained separate models for each category. Still, we seem to achieve better evaluation values than they do. Also, by training just one model we avoided overlapping aspect phrases for different categories.

We can see that the precision scores are better than the recall values, which might be caused by the rather small number of annotated sentences for training. During the process of parameter tuning, overfitting to the existing training set was an issue to be dealt with. Improving the recall means that the model finds aspect phrases in unknown data better, as recall implies that more of the existing phrases are found. We expect that a rather balanced precision and recall scores

will be favorable for future applications of the model. When taking the F1 scores 0.76, 0.75, 0.67, 0.82 and 0.97 into account, we judge the recall values of 0.67 to 0.80 (and 0.98 for label "O") as good, especially for the domain and data. However, the accuracy has a high score of 0.95, which is enhanced by the comparatively very high appearance of the label "O". This leads us to the decision to primarily use the F1 score.

A direct comparison to other studies and models is not possible as we evaluated our model on our own dataset, not on others. However, comparing our model and dataset to other studies, including shared tasks with models and datasets, suggests that our methods and data are superior. Yet, numerical scores may not be perfect and can lead to incorrect conclusions, which leads us to consider the possibility of manual evaluations as an additional evaluative feature. We wrote several example sentences that, as we regard it, were difficult for an automatic system to recognize or even close to the edge. Our system nevertheless performed well and found all aspect phrases and classified them correctly.

Conclusion

In this paper we briefly summarized the domain of ABSA and pointed out topics that had not been addressed in previous research. We also presented our dataset and the annotated aspect categories: namely, "friendliness", "competence", "time taken" and "explanation." All of these categories apply to the domain of physician reviews and experience goods. Our study involved 11,237 annotated sentences for these categories. In the future, we want to enlarge this by annotating more categories, also with new aspect targets such as the team of a physician. We then introduced our approach for extracting and classifying aspect phrases with implicit mentions from raw text. We also presented our good performance scores and critically reviewed them. In comparison to similar studies from ABSA research, in our work we integrate aspect phrase extraction and classification, outperforming others such as Pontiki et al. (2016b) even though we use our data and domain. Future work will also deal with polarity classification.

Acknowledgments

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 R. Mülfarth, Frederik S. Bäumer and Marvin Cordes for their support with the data collection.

References

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,*, volume 756, 77–89. Druskininkai, Lithuania: Springer.

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. ACL.

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. SCITEPRESS.

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

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. ACL.

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. 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. ACL.

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*, o. S. 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. 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. 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.