Implicit Statements in Healthcare Reviews: A Challenge for Sentiment Analysis

Joschka Kersting
Paderborn University
Paderborn, Germany
joschka.kersting@uni-paderborn.de

Frederik S. Bäumer
Bielefeld University of Applied Sciences
Bielefeld, Germany
frederik.baeumer@fh-bielefeld.de

***Abstract*—This paper aims at discussing past limitations set in sentiment analysis research regarding explicit and implicit mentions of opinions. Previous studies have regularly neglected this question in favor of methodical research on standard-datasets. Furthermore, they were limited to linguistically less-diverse domains, such as commercial product reviews. We face this issue by annotating a German-language physician review dataset that contains numerous implicit, long, and complex statements that indicate aspect ratings, such as the physician's friendliness. We discuss the nature of implicit statements and present various samples to illustrate the challenge described.**

***Index Terms*—Sentiment analysis; Natural language processing; Aspect phrase extraction.**

I. Introduction

Natural Language Processing (NLP) is a prominent subdomain of data science that is concerned with automatic processing of text data [1]. Natural language data is a challenge for machines because it is unstructured and contains imprecision, ambiguity, and vagueness [2]. There are characteristics that make standard language an efficient tool in human communication, but at which machine language processing regularly reaches its limits [2], [3]. A thriving topic in NLP and data science is Aspect-Based Sentiment Analysis (ABSA). ABSA "consists of two conceptual tasks, namely an aspect extraction and an aspect sentiment classification" [4], [5]. The aim is to categorize data by aspect and identify the sentiment polarity associated with each aspect. The subject of this analysis can be ratings of any kind, such as product ratings (e.g., cameras) or ratings of services (e.g., physician reviews). A well-known example is the battery of electronic products: While "*The battery of this phone is quite good*" is an explicit statement, '*'The phone lasts all day*" is the same statement but implicitly formulated [6].

As an example for a service review, earlier work [7] analyzes physician reviews and tries to apply a human-like language comprehension. The subject of physician reviews covers healthcare services that have been used by the author or a third party. These evaluative texts are published by users that describe their (dis-)satisfaction with a physician's treatment [8], [9]. A characteristic of these reviews is that they are often shaped by the sensitive physician-patient relationship. Due to this sensitive relationship, which should not be damaged despite review, many authors of reviews may resort to implicit statements in order to conceal the actual assessment somewhat. Implicit statements have the advantage that one does not have to commit oneself and can always deny in case of doubt. An example for implicit phrases is the following:

Example 1.
(1) "With this doctor, you don't just feel like a number."
(1) "Bei diesem Arzt fühlt man sich nicht nur als Nummer."

Example 1 shows that a patient was satisfied with the overall performance. The aspects "*time taken*" and "*friendliness*" are tangent to the positive statement. Both aspects are not explicitly mentioned but can be deduced. For a human reader, the connection arises from the overall context, since "*being a number*" is a phrase in German for feeling "*insignificant*" and "*unknown*". It can also be understood as "*to be treated without regard to personal circumstances*." While sentiment analysis approaches are quite capable of identifying the positive tenor, the domain-specific aspect classes remain unknown. To be able to process these reviews by machine with regard to aspects and associated sentiments, extensive datasets must be created and machine learning methods are trained with these datasets. A great number of previous research studies in this area focused on explicit statements and explicitly excluded implicit statements in some parts [10]. Moreover, studies often used the same datasets provided by Pontiki et al. [11]–[13], as survey papers demonstrate [14]–[16]. Hence, previous research is limited to what the datasets enable it to investigate. For example, the annotation guidelines of Pontiki et al. [17] state that only explicitly mentioned aspects should be annotated and that only one aspect in a sentence should be marked. Hence, researchers may train models that unlearn things that were not marked, due to these artificial boundaries. To make these statements visible by machine, we aim at implicit and explicit rating phrases in user-generated text. In this paper, we want to draw attention to the issue, show related work, and provide ideas to handle it.

This short paper is structured as follows: Section II presents related work for this paper. Based on this, we present examples for implicitness in review texts (Section III) and discuss them (Section IV). Finally, we conclude our work in Section V.

II. Related Work

In this section, we provide the related work with focus on deep learning for NLP (cf. Section II-A), sentiment analysis (cf. Section II-B), and user-generated content (cf. Section II-C).

A. Natural Language Processing

There has been great progress for NLP methods in recent years. Most notably, deep learning has evolved as the go-to method that improved the state-of-the-art in nearly all NLP tasks, such as question answering, sentiment analysis, and others. Here, transformers are the most important development [8], [18], [19]. Transformers apply a number of deep learning layers equipped with attention technology rather than recurrent neural networks [18]. This leads to favorable results and resource efficiency. Most notably, attention can process text sequences as a whole, i.e., it can weight words in a sentence according to their importance for the task it is learning [18], [19]. Recurrent networks such as Long Short-Term Memories (LSTMs) [20], on the other hand, process data sequentially, from the beginning to the end and hence only regard the part they have already seen [18], [19]. Furthermore, transformers have shifted the way neural networks are trained and handled for NLP. That is, large-scale models can be pre-trained on large amounts of raw text on a task that enables the construction of word vectors. These are representations of words, parts of words, or letters. The result is large models that can be shared with others. Industry practitioners can use these models or further train them for their specific data domain. Most notably, transformers are rather fine-tuned as a whole instead of inserting their vectors in other models. This process is called fine-tuning and describes a transformer receiving an additional layer and being trained on a downsstream task such as text classification for sentiment analysis. Here, all layers in the model are trained for this purpose [19], [21].

B. Sentiment Analysis

The second relevant area of research is sentiment analysis. We focus on ABSA in particular because we need to extract relevant statements from texts. Other works deal instead with document or sentence-based sentiment analyses such as full-text classification. ABSA can also be handled this way, but that is not purposeful, because the corresponding text spans must be extracted to enable further analyses for an in-depth knowledge of a text's contents and their explainability, [14]–[16]. This applies also to the distinction of implicit and explicit statements.

To describe ABSA and its components, we introduce the three sub-tasks here: ATE, ACC, and APC. ATE and ACC refer to Aspect Term Extraction and Aspect Class Classification [22]. These are usually conducted together [7] and describe the process of identification and categorization of aspect phrases in texts. APC refers to the Aspect Polarity Classification and describes the process of the sentiment polarity identification [22], e.g., negative or positive sentiment towards a cell phone battery. We need to conduct the first two steps ATE and ACC to extract implicit phrases from text. In contrast to studies such as Kersting & Geierhos [7], [8], we do not aim at topic-related extraction of aspect categories and their corresponding phrases. We deal with implicit and explicit phrases and their distinction. They set up aspect classes, extract them and further analyze them. However, the work also deals with implicit aspect phrases.

Several survey studies [14]–[16] present an overview of ABSA research. As can be seen, most works do not perform ATE, i.e., they do not provide in-depth analyses and go for the sentence or document-level. Moreover, they differ in the datasets they use [13], [23], [24]: Most studies use datasets from commercial review domains, e.g., for products or restaurants. They are not related to healthcare topics or physician reviews. Besides, most neural network approaches are based on rather common layer types such as (bi-)LSTMs or transformers. An example study for ABSA research is the one by De Clercq et al. [24]. They built an ABSA pipeline for social media data contents related to banking, retail, and human resources data. However, this does not deal with implicit statements. Garcia-Pablos et al. [25] present another example. They use topic modeling for finding thematic clusters. The topics found by topic modeling are not intuitive and cannot be clearly delimited for human users [26]. Hence, such approaches are very limited.

C. Physician Reviews

The third domain of relevant research includes physician reviews and research dealing with them. Such reviews serve as a sample domain for NLP research [7], but also are researched themselves [27], i.e., scholars want to investigate the content about healthcare providers and their performance. Physician reviews are published by users on Physician Review Websites (PRWs) sensibly and on the basis of trust [28]. They describe inter-personal issues and aspects such as the friendliness of a healthcare provider. This distinguishes them from commonly used commercial domains. Physician reviews serve as a good example for complex data domains [8]. On PRWs, there are two types of ratings: quantitative ratings such as stars or grades, and qualitative ratings such as texts. Both form a review. Quantitative ratings on PRWs are mostly positive [8] and there are numerous countries covered by PRW.

III. Implicitness in Aspect rich review texts

Having presented the most relevant research areas and selected studies from them, in this section, we present the challenge of implicit statements in service reviews. Here, we use data collected for earlier studies and related work [8], [29]. Hence, we have datasets of three different PRWs from three German-speaking countries. Here, we selected the review texts and split them into sentences using the NLP tool *Spacy*. We further applied some basic quality requirements such as a minimum and maximum length to avoid extremely long sentences spanning over a whole review without punctuation.

The annotation process was organized as follows: First, the sentences were randomized. We avoided annotator bias by not revealing further information about a review or the physician the sentence was written for. The annotation tool used is called *Prodigy*, a web-based tool that helps organizing and saving annotations consistently with multi-user support.

To keep the data quality high, we consistently monitored the annotations among team members. Edge cases were noted and talked through. However, the annotating persons were experienced with the task. We also wrote annotation guidelines where we distinguish explicit and implicit statements. This understanding was applied to the annotations. As a result, we have over 1400 sentences of which ca. 90% contain aspects. About 25% contain implicit aspect phrases and ca. 75% explicit. Despite our efforts, we cannot completely rule out subjectivity and are addressing this issue in future research.

Example 2.
(1) Well, the doctor is a nice person.
(1) Nun, der Doktor ist ein netter Mensch.
(2) When I meet him, he has always a warm smile in his face.
(2) Treffe ich ihn, hat er immer warmes Lächeln im Gesicht.

Sentence (1) in Example 2 does not state an aspect class explicitly, e.g., by saying "*The friendliness is positive.*" The rating towards the physician's friendliness would rather be expressed like in Sentence (2). But here again the questions arise whether naming a word that clearly hints to an aspect class, e.g., "*nice*" to "*friendliness*" is implicit or explicit. Besides, describing the "*warm smile*" is implicit: Here, no aspect class is indicated, but a human reader understands this interpersonal type of communication. As demonstrated, the distinction is not always clear or sharp. This is not uncommon in reviews of medical services, and we explain this with the sensitive doctor-patient relationship. Another explanation why reviewers so often resort to implicit statements may also be the strict rules of the PRWs, which protect against false reviews. Table I presents a number of examples for implicit and explicit aspect phrases as they persist in our dataset. As can be seen the distinction is challenging. To help readers understand our decisions, we accompany each sentence with an explanation. We applied a narrow understanding to implicit aspect phrases in contrast to previous research [7], [8]. They regard each case in which an aspect phrase is not directly mentioned as implicit and compare this to, e.g., Pontiki et al. [13], [17], who focus on directly named aspects. The following list demonstrates our comprehension of implicit and explicit aspect phrases:

Implicit phrases

- Statements apparent only when taken as a phrase or by taking the context as a whole into account, including idioms. An aspect class can be inferred from the phrase.
- Implicit phrases therefore do not contain explicit word choices (see underneath) from the aspect classes.

Explicit phrases

- At least one term of the known aspect classes is given, regardless of the inflectional form or part of speech; synonyms are included.
- It is made clear what is meant in the annotated phrase.

IV. Experiment and Discussion

In this short article, we want to draw particular attention to the need to give special consideration to implicit aspects in the evaluation of services and include relevant elucidating examples. But even the definition of these implicit aspects is not consistent. We have applied a narrow definition of implicit phrases, which makes the annotation task more challenging. Furthermore, the comprehension by Kersting and Geierhos [8] is easier to understand and apply due to its clear nature. However, we regard our approach as more sophisticated and hence useful for research, as easier understanding would limit the analyses. Kersting and Geierhos [8] do not research implicitness, but rather make use of implicit and explicit phrases regardless of whether they belong to either one of them. As an example, "*He is very friendly.*", taken from Table I, would be considered implicit by Kersting and Geierhos [8]. This is because the word "*friendly*" does not name the corresponding aspect phrase "*friendliness*", even though it (clearly) indicates it. However, our understanding is different: The example would be considered explicit, because the aspect class can be identified by human annotators.

Furthermore, to test whether computational models can learn this understanding, we trained several transformer models. These extract and classify implicit and explicit phrases from text based on our data. We conducted the experiment as a tagging task and thus handled two steps in one (ATE, ACC) and applied IO-tags, consistent with other works [8]. Based on Kersting and Geierhos [7], we applied XLM-RoBERTa [21] to the data, expecting to get favorable results. As the early experiment shows, it is possible to extract implicit phrases automatically, e.g., with an F1 score of 0.49 for implicit phrases. The overall accuracy for the model was 0.78, the F1 score 0.70. Words with explicit aspect mentions and irrelevant words using the O-tag were easier for the system to detect. This may be caused by fewer training data for implicit aspect phrases and their nature of being implicit while not having a limited or regularly occurring vocabulary.

Naturally, we conducted more and further experiments. A multi-label multi-class classification on the sentence-level succeeded and achieved good results. The Label ranking average precision (LRAP) [30] training XLM-RoBERTa large is 0.90. This is a very good score as it is close to 1, which would be best (between 0 and 1). Moreover, when training the tagger, other models such as BERT [19] in multiple variations (large, domain-trained, etc.) achieved almost the same scores like XLM-RoBERTa. A large German version of BERT [31] even outperformed XLM-RoBERTa large [21] achieving a macro F1 score of 0.74, an accuracy of 0.81 and an F1 score for implicit phrases of 0.52. Further used parameters were a small batch size (e.g., 4) and few epochs (8).

As can be seen, machine classification and extraction of implicit phrases is possible. However, we still need to further elaborate (and enrich) the definition of implicit aspect phrases. There is some related work in this area, but that is often focused on products and its elaboration does not fully reflect

TABLE I
EXAMPLES OF REVIEW PHRASES

Sentence	Class	Explanation
"*He is very friendly.*"	explicit	What is meant is explicitly made clear in the phrase.
"*He is not at all competent.*"	explicit	The competence is directly mentioned by an adjective.
"*He does not look me in the eye.*"	implicit	Only a human can guess that this is considered rude, because the issue is described in a phrase rather than a single adjective.
"*My wife and I have been going to this doctor for many years.*"	implicit	A trust relationship can only be derived from this statement.
"*I was treated immediately.*"	implicit	A human can guess that there was no waiting time.
"*Through him, a disease was finally detected.*"	implicit	Competence can be derived from the context.

the subtleties that come with the evaluation of services and other inter-personal situations [7, e.g.].

V. CONCLUSION

In this short paper, we have provided insight into implicit statements that occur in German physician reviews. Implicit statements are not a new phenomenon in machine processing of texts, especially not in sentiment analysis. However, implicit statements occur frequently in the review of medical services. This is also due to the sensitive physician-patient relationship. In principle, the challenge seems manageable with machine learning and clean, well-annotated datasets. We will continue to follow this path. the implications to the work of non-German structured languages. Finally, we know that this basic idea is not only valid for German, but that this challenge of implicit statements is also found in other languages. We are therefore eager to develop our approaches beyond German and in other domains. Furthermore, it can be a future research direction to investigate implicit phrases together with sentiment polarity, objectivity and subjectivity.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] I. Zeroual and A. Lakhouaja, "Data science in light of natural language processing: An overview," *Procedia Computer Science*, vol. 127, pp. 82–91, 2018.

[2] K.-U. Carstensen, C. Ebert, C. Ebert, S. Jekat, R. Klabunde, and H. Langer, Eds., *Computerlinguistik und Sprachtechnologie: Eine Einführung [Computational Linguistics and Language Technology: An Introduction]*, 3rd ed. Heidelberg, Germany: Spektrum Akademischer Verlag, 2010.

[3] H. Bußmann, *Lexikon der Sprachwissenschaft [Lexicon of Linguistics]*. Stuttgart: Alfred Kröner Verlag, 2008.

[4] Q. Jiang, L. Chen, R. Xu, X. Ao, and M. Yang, "A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: ACL, 2019, pp. 6280–6285.

[5] M. H. Phan and P. O. Ogunbona, "Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis," in *Proceedings of the 58th Annual Meeting of the ACL*. Online: ACL, 2020, pp. 3211–3220.

[6] K. Schouten and F. Frasincar, "Finding implicit features in consumer reviews for sentiment analysis," in *International Conference on Web Engineering (ICWE)*. Toulouse, France: Springer, 2014, pp. 130–144.

[7] J. Kersting and M. Geierhos, "Human Language Comprehension in Aspect Phrase Extraction with Importance Weighting," in *Natural Language Processing and Information Systems*, ser. LNCS, E. Métais, F. Meziane, H. Horacek, and E. Kapetanios, Eds., vol. 12801. Saarbrücken, Germany: Springer Nature, 2021, pp. 231–242.

[8] ——, "Towards Aspect Extraction and Classification for Opinion Mining with Deep Sequence Networks," in *Natural Language Processing in Artificial Intelligence – NLPinAI 2020*, ser. Studies in Computational Intelligence (SCI), R. Loukanova, Ed. Cham, Switzerland: Springer, 2021, vol. 939, pp. 163–189.

[9] M. Emmert, F. Meier, F. Pisch, and U. Sander, "Physician Choice Making and Characteristics Associated With Using Physician-Rating Websites: Cross-Sectional Study," *Journal of Medical Internet Research*, vol. 15, no. 8, p. e187, 2013.

[10] M. Tubishat, N. Idris, and M. A. Abushariah, "Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges," *Information Processing & Management*, vol. 54, no. 4, pp. 545–563, 2018.

[11] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation*. Dublin, Irland: ACL, 2014, pp. 27–35.

[12] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 Task 12: Aspect Based Sentiment Analysis," in *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, CO, USA: ACL, 2015, pp. 486–495.

[13] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, CA, USA: ACL, 2016, pp. 19–30.

[14] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," *IEEE Transactions on Affective Computing*, 2020.

[15] J. Zhou, J. X. Huang, Q. Chen, Q. V. Hu, T. Wang, and L. He, "Deep Learning for Aspect-Level Sentiment Classification: Survey, Vision, and Challenges," *IEEE Access*, vol. 7, pp. 78 454–78 483, 2019.

[16] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418306456

[17] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval 2016 Task 5: Aspect Based Sentiment Analysis (ABSA-16) Annotation Guidelines," pp. 1–20, 2016.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proceedings of the 31st Conference on Neural Information Processing Systems*. Long Beach, CA, USA: Curran Associates, 2017, pp. 5998–6008.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*. Minneapolis, MN, USA: ACL, 2019, pp. 4171–4186.

[20] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsu-

pervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the ACL.* Online: ACL, 2020, pp. 8440–8451.

[22] T. C. Chinsha and J. Shibily, "A Syntactic Approach for Aspect Based Opinion Mining," in *Proceedings of the 9th IEEE International Conference on Semantic Computing*. Anaheim, CA, USA: IEEE, 2015, pp. 24–31.

[23] A. López, A. Detz, N. Ratanawongsa, and U. Sarkar, "What Patients Say About Their Doctors Online: A Qualitative Content Analysis," *Journal of General Internal Medicine*, vol. 27, no. 6, pp. 685–692, 2012.

[24] O. De Clercq, E. Lefever, G. Jacobs, T. Carpels, and V. Hoste, "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*. Copenhagen, Denmark: ACL, 2017, pp. 136–142.

[25] A. Garcia-Pablos, M. Cuadros, and G. Rigau, "W2VLDA: Almost Unsupervised System for Aspect Based Sentiment Analysis," *Expert Systems with Applications*, vol. 91, pp. 127–137, 2018.

[26] A. Mukherjee and B. Liu, "Aspect Extraction through Semi-Supervised Modeling," in *Proceedings of the 50th Annual Meeting of the ACL*, vol. 1. Jeju, South Korea: ACL, 2012, pp. 339–348.

[27] M. Emmert, U. Sander, A. S. Esslinger, M. Maryschok, and O. Schöffski, "Public Reporting in Germany: the Content of Physician Rating Websites," *Methods of Information in Medicine*, vol. 51, no. 2, pp. 112–120, 2012.

[28] J. Kersting, F. Bäumer, and M. Geierhos, "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 (IoTBDS).* Heraklion, Greece: SCITEPRESS, 2019, pp. 147–155.

[29] J. Kersting and M. Geierhos, "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 (ICAART - NLPinAI 2020).* Valetta, Malta: SCITEPRESS, 2020, pp. 391–400.

[30] Scikit-learn Developers, "sklearn.metrics.label_ranking_average_precision_score – scikit-learn 1.0.2 documentation," https://scikit-learn.org/stable/modules/generated/sklearn.metrics.label_ranking_average_precision_score.html, 2020, accessed 04.04.2022.

[31] B. Chan, S. Schweter, and T. Möller, "deepset/gbert-large - hugging face," https://huggingface.co/deepset/gbert-large, 2022, accessed 2022-04-04.