Annotation Quality and Its Influence on Aspect-Based Sentiment Analysis: A Case Study on German Restaurant Reviews

Master thesis presentation

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

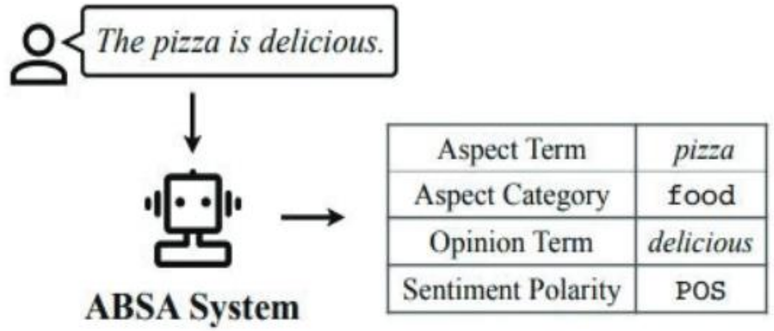
Universität Regensburg

Introduction

Niklas Donhauser	6th semester Master Media Informatics
Supervisor	Jakob Fehle
First reviewer	Prof. Dr. Christian Wolff
Second reviewer	Prof. Dr. Udo Kruschwitz
Current status	Literature research completed, planning phase of the annotation studies

Background

- **Sentiment Analysis (SA):**
  Evaluation of the general sentiment (positive / negative / neutral) in texts [1].
- **Aspect-based Sentiment Analysis (ABSA):**
  Analysis of the sentiment towards certain aspects of an entity (e.g. properties, product features) [2].



Aspect Term	*pizza*
Aspect Category	food
Opinion Term	*delicious*
Sentiment Polarity	POS

Sentiment Elemente [3]

[1] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature. https://hyse.org/pdf/SentimentAnalysis-and-OpinionMining.pdf

[2] Chauhan, G. S., Nahta, R., Meena, Y. K., & Gopalani, D. (2023). Aspect based sentiment analysis using deep learning approaches: A survey. Computer Science Review, 49, 100576. https://doi.org/10.1016/j.cosrev.2023.100576

[3] Singhi, V., Chauhan, C., & Soni, P. K. (2024, April). Exploring Progress in Aspect-based Sentiment Analysis: An In-depth Survey. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 1-10). IEEE. https://doi.org/10.1109/I2CT61223.2024.10543612

Background & Related Work

- Literature search via Google Scholar, IEEE, ScienceDirect, ACM Digital Library and ACL Anthology.
- English dominates the available ABSA research and thus also existing datasets [4].
- High effort for manual annotation, especially for complex tasks [5].
- Data quality is crucial for training accurate, unbiased and trustworthy machine learning models and for their correct evaluation [6].

➢ The influence of annotation quality and annotators on the final data and model quality has hardly been systematically investigated for complex tasks.

[4] Chebolu, S. U. S., Dernoncourt, F., Lipka, N., & Solorio, T. (2023, November). A review of datasets for aspect-based sentiment analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 611-628). https://doi.org/10.18653/v1/2023.ijcnlp-main.41

[5] Li, G., Wang, H., Ding, Y., Zhou, K., & Yan, X. (2023). Data augmentation for aspect-based sentiment analysis. *International Journal of Machine Learning and Cybernetics*, *14*(1), 125-133.

[6] Klie, J. C., Castilho, R. E. D., & Gurevych, I. (2024). Analyzing dataset annotation quality management in the wild. Computational Linguistics, 50(3), 817-866. https://doi.org/10.48550/arXiv.2307.08153

Related Work

- Analysis of existing strategies to ensure annotation quality and their implementation in practice [6].
- CrowdWorkSheet provides a structured framework for fair, transparent and high-quality crowdsourcing annotations [7].
- The combination of crowdsourcing and LLM annotations with label aggregation increases annotation quality [8, 9].
- Self-consistency increases the robustness of LLM annotations by aggregating multiple reasoning paths [10].

[7] Díaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V., & Denton, R. (2022, June). Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2342-2351). https://doi.org/10.1145/3531146.3534647

[8] He, Z., Huang, C. Y., Ding, C. K. C., Rohatgi, S., & Huang, T. H. K. (2024, May). If in a crowdsourced data annotation pipeline, a gpt-4. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-25). https://doi.org/10.1145/3613904.3642834

[9] Li, J. (2024, April). A comparative study on annotation quality of crowdsourcing and LLM via label aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6525-6529). IEEE.

[10] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Objective of the work

- Development and implementation of various annotation studies by:
  - Crowdsourcing
  - Large Language Models (Zero-shot / Few-shot)
  - Students (Contractors)
  - Experts
- Evaluation of the various annotated datasets using state-of-the-art methods.
  - Investigation of label aggregation to improve data quality.
  - Evaluation of an assembly approach to combine crowdsourcing and LLM annotation.
- Development of best practice for such studies.

Dataset

Name: GERestaurant [11]

Size: 3,078 Sentences (2.153 Trainset, 925 Testset)

Aspects: 3,149 explicit and 1,165 implicit aspects

Aspect categories: *food, service, general impression, ambience, price*

Example: *Das Personal war sehr freundlich.*
(The staff were very friendly.)

Aspects:

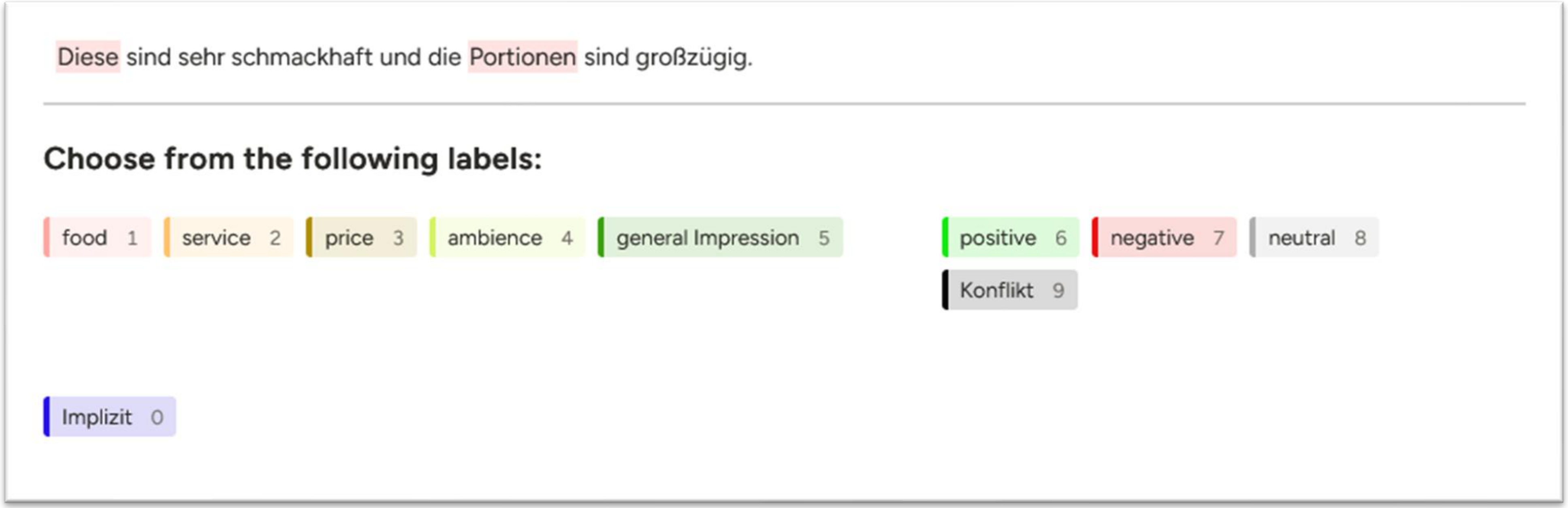
- Aspect term: Personal (staff)
- Aspect category: service
- Aspect polarity: positive

[11] Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024, September). GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024) (pp. 123-133). https://doi.org/10.48550/arXiv.2408.07955

Annotations studies

- Using Label Studio as an annotation tool.[1]
- New annotation of most of the dataset.
  - Revision of the existing annotation instructions.
  - Using the Agile Corpus Creation process [6].

Label Studio



Diese sind sehr schmackhaft und die Portionen sind großzügig.

Choose from the following labels:

[1] https://labelstud.io/

Annotation: Ground Truth

- Reference basis for comparing and evaluating the different annotations.

Procedure:

- Revision of the test dataset (924 records) by two additional ABSA task experts.
- Checking the dataset for consistency and incorrect classes.
- Iterative procedure [6].

Annotation: Crowdsourcing

Procedure:

- Use of the CrowdWorkSheet [7].
- Conducting a pilot study [6].
  - Recruitment of German-speaking test subjects via Prolific.[1]
  - Check whether sufficient test subjects can be recruited for the task on Prolific.

Annotation approaches:

- Simple: Three people annotate only one element (category, polarity, aspect phrase) per sentence.
- Complex: One person annotates all three elements (category, polarity, aspect phrase).
- Mixed: Focus only on phrase annotations. Category & polarity are specified by models.

[1] https://www.prolific.com/

Annotation: Crowdsourcing II

Problem:

- Number of participants: It is unclear whether enough qualified annotators are available.
- Technical implementation: The Prolific interface is unsuitable for complex annotations. → Use of an external annotation tool necessary.
- Costs: 10-20 cents per annotation → around € 215 - € 430 for complete annotation of the dataset.

Annotation: Large Language Models

Procedure:

- Use of Large Language Models to annotate the training dataset (GPT-4, GPT 3.5, Mixtral, LLaMA).

Prompt Engineering:

  - Zero-shot
  - Few-shot
  - Annotation guidelines
- Self-Consistency [10]

Annotation: Students and Experts

Procedure:

- Conduct a pilot study to test the procedure.
- Iterative procedure:
  - Division into batches (approx. 200 sentences)
  - Debriefing of the annotations
  - Revision of the guidelines
  - Agreement
- Demographic questionnaire

Experts:

- Two ABSA task experts revise most of the training dataset using the existing annotation.

Students (Contractors):

- Annotator training (detailed guidelines and example annotations with solution)
- Annotator debriefing (indicate problems and uncertainties)

Evaluation

Evaluation of the annotations:

- Inter-Annotator-Agreement (Krippendorff's α)
- Control questions

Evaluation with the help of models:

- Use of state-of-the-art models (MvP, paraphrase, BERT-CLF) [12, 13, 14]
- Metrics (F1 score, accuracy, recall, precision)

[12] Gou, Z., Guo, Q., & Yang, Y. (2023, July). *MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4380–4397). https://doi.org/10.18653/v1/2023.acl-long.240

[13] Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., & Lam, W. (2021). *Aspect sentiment quad prediction as paraphrase generation*. *arXiv preprint* arXiv:2110.00796. https://doi.org/10.48550/arXiv.2110.00796

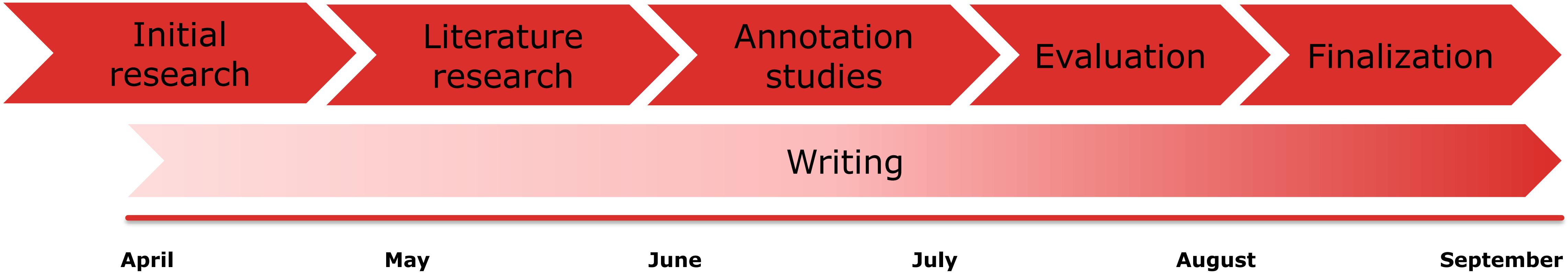
[14] Fehle, J., Münster, L., Schmidt, T., & Wolff, C. (2023, September). *Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews*. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)* (pp. 202–218). https://aclanthology.org/2023.konvens-main.21.pdf

Evaluation II

Annotation order can influence evaluation:

- Example: *Das Personal war sehr freundlich.*
- All-in-One Annotation: Subjects annotate all three elements (phrase, category, polarity) and conclusions are drawn about subtasks. afterwards.
    - Annotation: Personal, Service, Positive -> Service, Positive
- Split-by-Subtask: Subjects annotate the individual subtasks (term, category, polarity and categorization, polarity) separately.
    - Annotation I: Personal, Service, Positive
    - Annotation II: Service, Positive

Schedule


Initial research
Literature research
Annotation studies
Evaluation
Finalization
Writing
April
May
June
July
August
September

Summary

Background:
Limited availability of German-language ABSA datasets and lack of studies on the quality of complex annotations.

Aim:
Investigation of the influence of annotation types on the quality of datasets for aspect-based sentiment analysis.

Annotation studies by:

- Crowdsourcing
- Large Language Models (Zero-shot / Few-shot)
- Students (Contractors)
- Experts

Sources

[1] Liu, B. (2022). Sentiment Analysis and Opinion Mining. Springer Nature. https://hyse.org/pdf/SentimentAnalysis-and-OpinionMining.pdf

[2] Chauhan, G. S., Nahta, R., Meena, Y. K., & Gopalani, D. (2023). Aspect based sentiment analysis using deep learning approaches: A survey. Computer Science Review, 49, 100576. https://doi.org/10.1016/j.cosrev.2023.100576

[3] Singhi, V., Chauhan, C., & Soni, P. K. (2024, April). Exploring Progress in Aspect-based Sentiment Analysis: An In-depth Survey. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 1–10). IEEE. https://doi.org/10.1109/I2CT61223.2024.10543612

[4] Chebolu, S. U. S., Dernoncourt, F., Lipka, N., & Solorio, T. (2023, November). A review of datasets for aspect-based sentiment analysis. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 611–628). https://doi.org/10.18653/v1/2023.ijcnlp-main.41

[5] Li, G., Wang, H., Ding, Y., Zhou, K., & Yan, X. (2023). Data augmentation for aspect-based sentiment analysis. International Journal of Machine Learning and Cybernetics, 14(1), 125-133.

[6] Klie, J. C., Castilho, R. E. D., & Gurevych, I. (2024). Analyzing dataset annotation quality management in the wild. Computational Linguistics, 50(3), 817–866. https://doi.org/10.48550/arXiv.2307.08153

[7] Díaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V., & Denton, R. (2022, June). CrowdWorkSheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 2342–2351). https://doi.org/10.1145/3531146.3534647

[8] He, Z., Huang, C. Y., Ding, C. K. C., Rohatgi, S., & Huang, T. H. K. (2024, May). If in a crowdsourced data annotation pipeline, a GPT-4. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1–25). https://doi.org/10.1145/3613904.3642834

[9] Li, J. (2024, April). A comparative study on annotation quality of crowdsourcing and LLM via label aggregation. In ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 6525–6529). IEEE.

[10] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

[11] Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024, September). GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024) (pp. 123–133). https://doi.org/10.48550/arXiv.2408.07955

[12] Gou, Z., Guo, Q., & Yang, Y. (2023, July). *MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4380–4397). https://doi.org/10.18653/v1/2023.acl-long.240

[13] Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., & Lam, W. (2021). *Aspect sentiment quad prediction as paraphrase generation*. *arXiv preprint* arXiv:2110.00796. https://doi.org/10.48550/arXiv.2110.00796

[14] Fehle, J., Münster, L., Schmidt, T., & Wolff, C. (2023, September). *Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews*. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)* (pp. 202–218). https://aclanthology.org/2023.konvens-main.21.pdf