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

Antrittsvortrag Masterarbeit

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

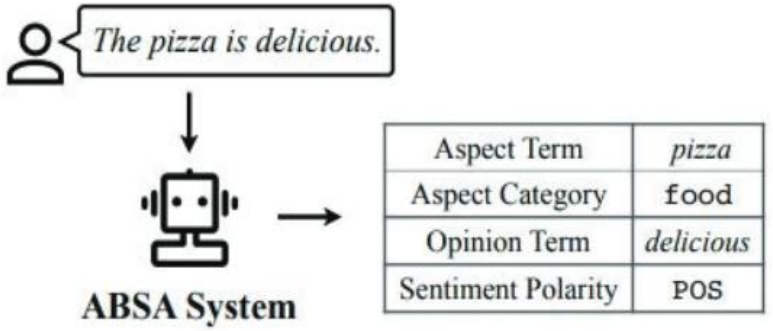
Universität Regensburg

Vorstellung

Niklas Donhauser	6. Semester Master Medieninformatik
Betreuer	Jakob Fehle
Erstgutachter	Prof. Dr. Christian Wolff
Zweitgutachter	Prof. Dr. Udo Kruschwitz
Stand	Literaturrecherche abgeschlossen, Planungsphase der Annotationsstudien

Hintergrund

- **Sentiment Analyse (SA):**
  Bewertung der allgemeinen Stimmung (positiv / negativ / neutral) in Texten [1].
- **Aspekt-basierte Sentiment Analyse (ABSA):**
  Analyse der Stimmung zu bestimmten Aspekten einer Entität (z.B. Eigenschaften, Produktmerkmale) [2].



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

Hintergrund & Verwandte Arbeiten

- Literaturrecherche über Google Scholar, IEEE, ScienceDirect, ACM Digital Library und ACL Anthology.
- Englisch dominiert die verfügbare ABSA-Forschung und somit auch vorhandene Datensätze [4].
- Hoher Aufwand bei manueller Annotation insbesondere bei komplexen Aufgaben [5].
- Die Datenqualität ist entscheidend für das Training genauer, unvoreingenommener und vertrauenswürdiger Modelle für maschinelles Lernen sowie für deren korrekte Auswertung [6].

➢ Der Einfluss der Annotationsqualität und der Annotatoren auf die finale Daten- und Modellqualität ist bei komplexen Aufgaben bislang kaum systematisch untersucht worden.

[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

Verwandte Arbeiten

- Analyse bestehender Strategien zur Sicherung der Annotationsqualität und deren Umsetzung in der Praxis [6].
- CrowdWorkSheet bietet einen strukturierten Rahmen für faire, transparente und qualitativ hochwertige Crowdsourcing-Annotationen [7].
- Die Kombination von Crowdsourcing und LLM-Annotationen mit Label Aggregation erhöht die Annotationsqualität [8, 9].
- Self-Consistency steigert die Robustheit von LLM-Annotationen durch Aggregation mehrerer Reasoning-Pfade [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*.

Zielsetzung

- Entwicklung und Durchführung verschiedener Annotationsstudien durch:
  - Crowdsourcing
  - Large Language Models (Zero-shot / Few-shot)
  - Studierende (Contractors)
  - Experten
- Auswertung der verschieden annotierten Datensätze mit State-of-the-Art Methoden.
  - Untersuchung von Label Aggregation zur Verbesserung der Datenqualität.
  - Evaluation eines Assembly-Ansatzes zur Verbindung von Crowdsourcing und LLM-Annotation.
- Herausarbeitung von Best Practice für solche Studien.

Datensatz

Name: GERestaurant [11]

Größe: 3.078 Sätze (2.153 Trainset, 924 Testset)

Aspekte: 3.149 explizite und 1.165 implizite Aspekte

Aspekt-Kategorien: *food, service, general impression, ambience, price*

Beispiel Satz: *Das Personal war sehr freundlich.*

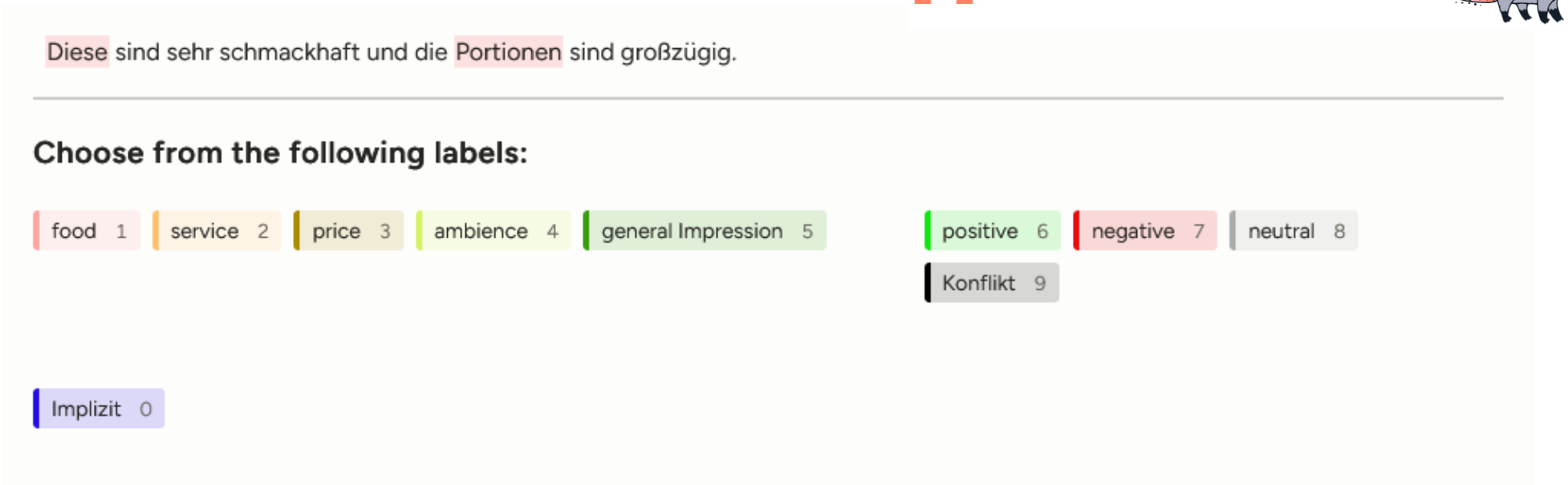
Aspekte:
- Aspekt-Term: Personal
- Aspekt-Kategorie: service
- Aspekt-Polarität: positive

[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

Annotationsstudien

- Verwenden von Label Studio als Annotationstool.[1]
- Neue Annotation vom gesamten Datensatz.
  - Überarbeitung der vorhandenen Annotationsanleitung.
  - Verwenden des Agile Corpus Creation Prozesses [6].

Label Studio



[1] https://labelstud.io/

Annotation: Ground Truth

- Referenzbasis zum Vergleich und zur Bewertung der unterschiedlichen Annotationen.

Vorgehen:

- Überarbeitung des Testdatensatzes (924 Sätze), durch zwei weitere ABSA-Task Experten.
- Überprüfen des Datensatzes auf Konsistenz und fehlerhafte Klassen.
- Iteratives Vorgehen [6].

Annotation: Crowdsourcing

Vorgehen:

- Nutzung des CrowdWorkSheets [7].
- Durchführung einer Pilotstudie [6].
  - Rekrutierung von deutschsprachigen Probanden über Prolific.[1]
  - Prüfung, ob ausreichend Probanden für die Aufgabe auf Prolific rekrutierbar sind.

Annotationsansätze:

- Einfach: Drei Personen annotieren jeweils nur ein Element (Kategorie, Polarität, Aspekt-Phrase) pro Satz.
- Komplex: Eine Person annotiert alle drei Elemente (Kategorie, Polarität, Aspekt-Phrase).
- Mixed: Fokus nur auf Phrasen-Annotationen. Kategorie & Polarität werden durch Modelle vorgegeben.

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

Annotation: Crowdsourcing II

Problematik:

- Teilnehmerzahl: Es ist unklar, ob ausreichend qualifizierte Annotatoren verfügbar sind.
- Technische Umsetzung: Das Prolific Interface ist für aufwendige Annotationen ungeeignet. → Einsatz eines externen Annotationstools notwendig.
- Kosten: 10–20 Cent pro Annotation → bei vollständiger Annotation des Datensatzes rund 215 € - 430 €.

Annotation: Large Language Models

Vorgehen:

- Verwenden von Large Language Models zum Annotieren des Trainingsdatensatzes (GPT-4, GPT 3.5, Mixtral, LLaMA).
- Prompt Engineering:
  - Zero-shot
  - Few-shot
  - Annotationsrichtlinien
- Self-Consistency [10]

Annotation: Studierende und Experten

Vorgehen:

- Durchführung einer Pilotstudie zum Testen des Vorgehens.
- Iteratives Vorgehen:
  - Aufteilung in Batches (circa 200 Sätze)
  - Nachbesprechung der Annotationen
  - Überarbeitung der Guidelines
  - Agreement
- Demografischer Fragebogen

Experten:

- Zwei ABSA-Task Experten überarbeiten den gesamten Trainingsdatensatz mithilfe der vorhandenen Annotation.

Studierende (Contractors):

- Annotatorentraining (Ausführliche Guidelines und Beispielannotationen mit Lösung)
- Annotatorendebriefing (Probleme und Unsicherheiten angeben)

Auswertung

Auswertung der Annotationen:

- Inter-Annotator-Agreement (Krippendorff's α)
- Kontrollfragen

Auswertung Mithilfe von Modellen:

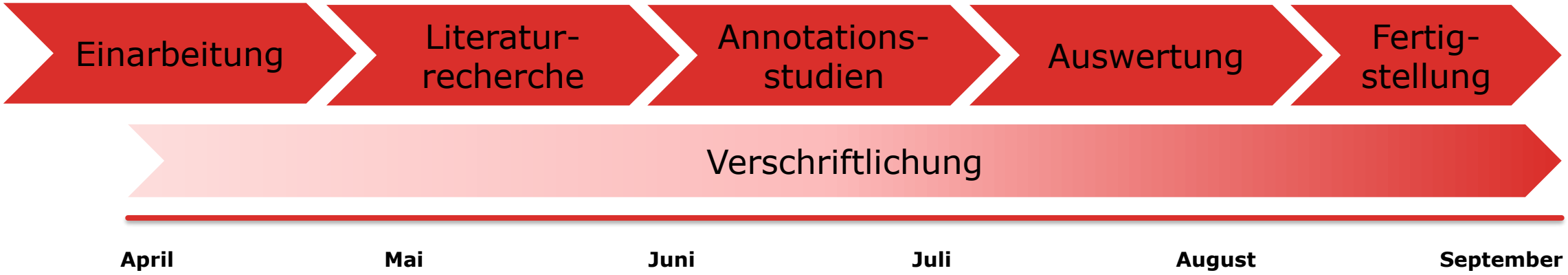
- Einsatz von State-of-the-Art Modellen (MvP, Paraphrase, BERT-CLF) [12, 13, 14]
- Metriken (F1-Score, Accuracy, Recall, Precision)

Annotationsreihenfolge kann Auswertung beeinflussen:

- Die Annotation aller Sentiment-Elemente in einem Schritt erschwert die getrennte Analyse von Teilaufgaben.

[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

Zeitplan


Einarbeitung
Literatur-recherche
Annotations-studien
Auswertung
Fertig-stellung
Verschriftlichung
April
Mai
Juni
Juli
August
September

Zusammenfassung

Hintergrund:
Begrenzte Verfügbarkeit deutschsprachiger ABSA-Datensätze und fehlende Untersuchungen zur Qualität komplexer Annotationen.

Ziel:
Untersuchung des Einflusses von Annotationsarten auf die Qualität von Datensätzen für die Aspekt-basierte Sentiment Analyse.

Annotationsstudien durch:
- Crowdsourcing
- Large Language Models (Zero-shot / Few-shot)
- Studierende (Contractors)
- Experten

Quellen

[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