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

Abschlussvortrag Masterarbeit

Niklas Donhauser
Lehrstuhl für Medieninformatik
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

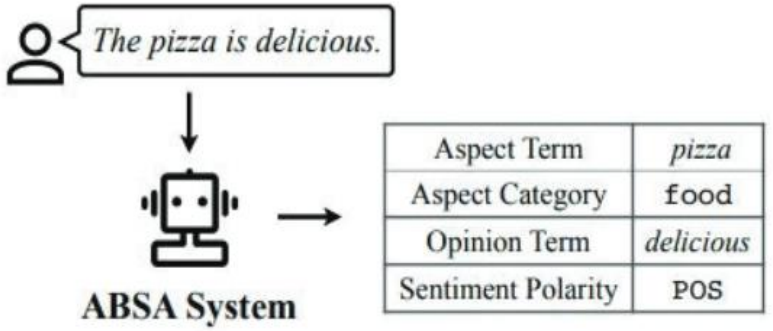
Universität Regensburg

Vorstellung

Niklas Donhauser	6. Semester Master Medieninformatik
Betreuer	Jakob Fehle
Erstgutachter	Prof. Dr. Christian Wolff
Zweitgutachter	Prof. Dr. Udo Kruschwitz
Stand	Annotationsstudie Crowdsourcing, Auswertung der Daten, Verschriftlichung

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

Zielsetzung

- Entwicklung und Durchführung verschiedener Annotationsstudien durch:
  - Crowdsourcing über Prolific[1]
  - Large Language Models (LLM)
  - Studierende
  - Experten
- Auswertung der verschieden annotierten Datensätze mit State-of-the-Art (SOTA) Methoden.
- Herausarbeitung von Best Practices für solche Studien.

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

Datensatz & Tasks

- Ausgangsdatensatz: GERestaurant [7]
- Neues Subset: 1904 Sätze (1.000 Trainset, 904 Testset)
- Aspekt-Kategorien: *Essen, Service, Gesamteindruck, Ambiente, Preis*
- Beispiel Satz: *Das Personal war sehr freundlich.*

Aspect Category Sentiment Analysis (ACSA)

- Aspekt-Kategorie: Service
- Aspekt-Polarität: Positiv

Target Aspect Sentiment Detection (TASD)

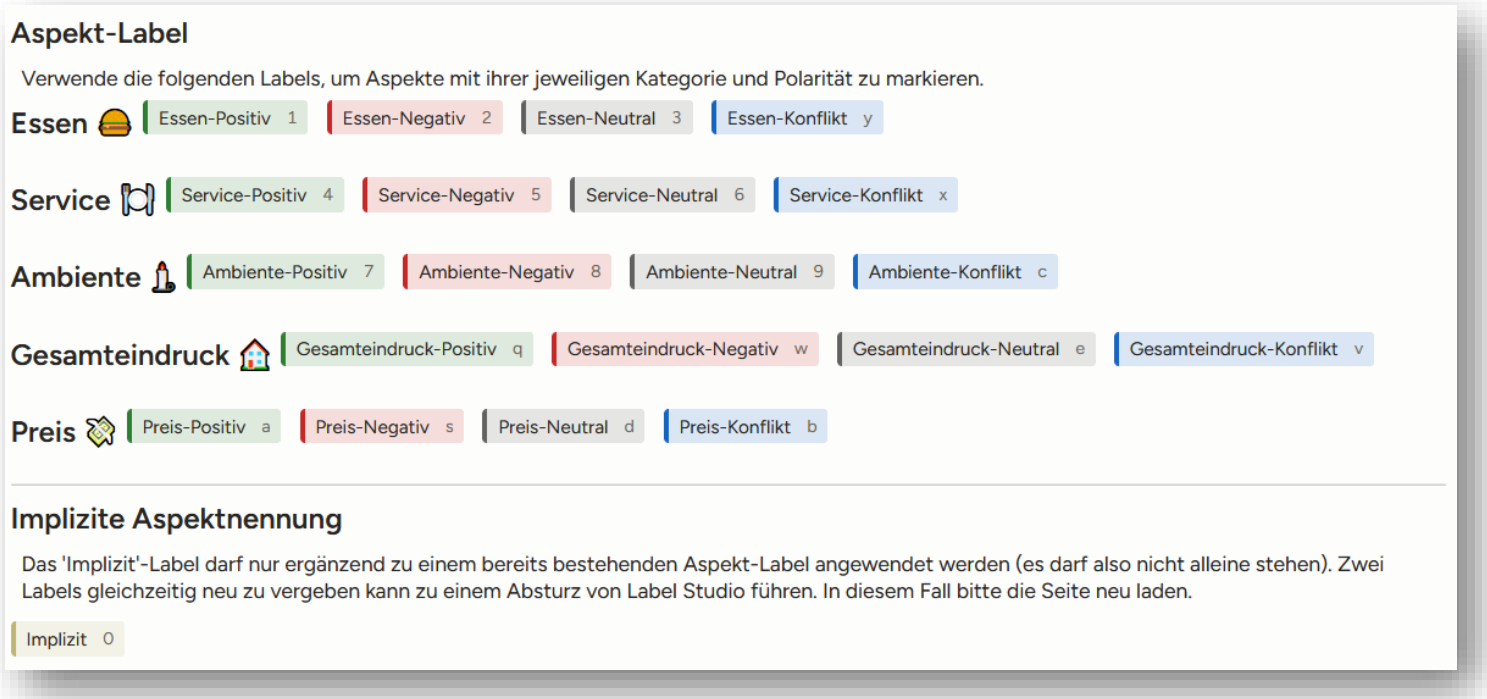
- Aspekt-Kategorie: Service
- Aspekt-Polarität: Positiv
- Aspekt-Term: Personal

[7] 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

Label Studio

- Verwenden von Label Studio als Annotationstool.[1]
- Neue Annotation vom 1000 zufällig ausgewählten Sätzen aus dem originalen Trainset.
  - Überarbeitung der vorhandenen Annotationsanleitung.
  - Verwenden des Agile Corpus Creation Prozesses [6].



[1] https://labelstud.io/

[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

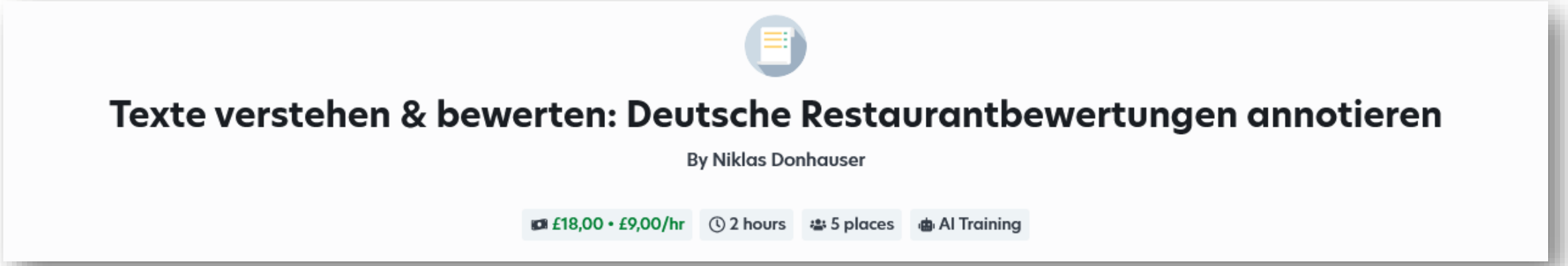
Annotation: Ground Truth

- Evaluationsset: Referenzbasis zum Vergleich und zur Bewertung der unterschiedlichen Annotationen.
- 2 ABSA-Task-Experten haben alle 924 Texte annotiert.
  - Aufteilung in 5 gleich große Batches.
  - Feedback-Runden nach jedem Batch.
  - Bei Texten ohne Übereinstimmung wurde gemeinsam ein Label bestimmt.
- Inter-Annotator Agreement (IAA)
  - TASD: Micro-F1 71.60
  - ACSA: Micro-F1 87.20 | Krippendorff's alpha 67.40

	Positive		Negative		Neutral		Total	
Aspect Category	Expli.	Impli.	Expli.	Impli.	Expli.	Impli.	Expli.	Impli.
AMBIENCE	83	7	24	14	2	0	109	21
FOOD	260	22	159	29	37	3	456	54
GENERAL	23	96	18	89	1	9	42	194
PRICE	14	0	45	11	5	4	64	15
SERVICE	146	21	94	56	3	0	243	77
Total	526	146	340	199	48	16	914	361

Annotation: Crowdsourcing

- Verwendung von Prolific[1] zum Rekrutieren der Probanden.
- Teilnehmer pro Task: 15 (TASD), 15 (ACSA)
- Durchführung der Annotationsstudie in dieser Woche.
- Pilotstudie (abgeschlossen)
  - 5 Batches (à 200 Sätze), um die TASD Task zu testen.
  - 15 Teilnehmer (5 erfolgreiche Annotationen, 10 Abbrüche)



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

Annotation: Large Language Models

- Verwenden von LLMs zum Annotieren des Trainingsdatensatzes.
- Zwei separate Durchgänge für TASD bzw. ACSA.
- Ansatz & Parameter [8]
  - Gemma 3 27B
  - 30-shot
  - Self-Consistency[9] 5 Generationen mit unterschiedlichen Seeds
  - Temperatur 0.8
  - Dauer: Circa 10h
- Inter-Annotator Agreement zwischen den verschiedenen LLM-Durchläufen
  - TASD: Micro-F1 90.17
  - ACSA: Micro-F1 97.20 | Krippendorff's alpha 93.25

[8] Hellwig, N. C., Fehle, J., Kruschwitz, U., & Wolff, C. (2025, May). *Do we still need human annotators? Prompting large language models for aspect sentiment quad prediction*. arXiv. https://doi.org/10.48550/arXiv.2502.13044
[9] 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*.

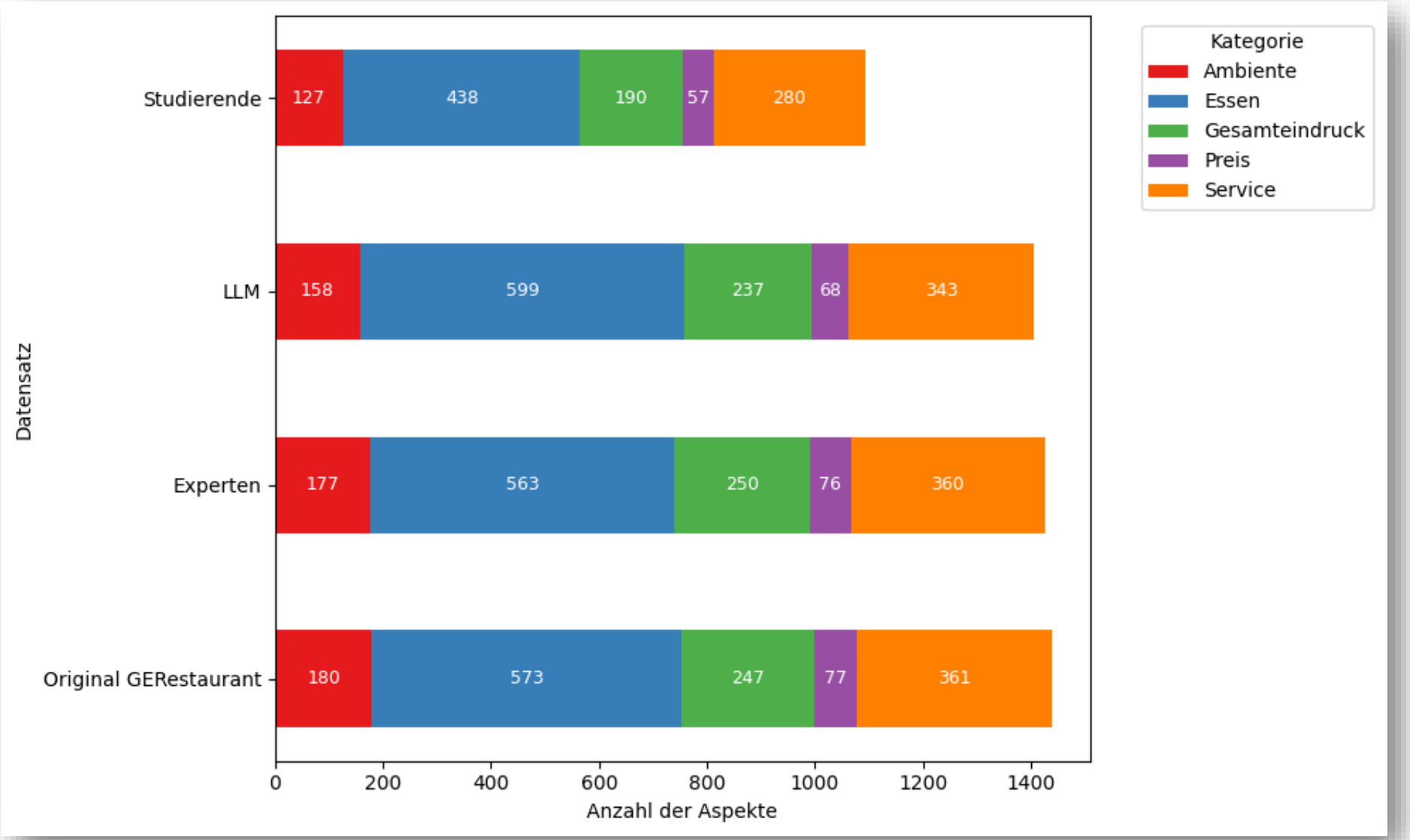
Annotation: Studierende

- Annotatoren haben einen Medieninformatik Hintergrund.
- Anwerbung über das Versuchspersonen Forum.
- 32 Studierende (15 bei TASD, 17 bei ACSA)
  - Fehlerhafte Annotation in zwei Batches (ACSA) → erneute Annotation dieser.
- Jeder Annotator bearbeitet 200 Sätze.
- Vergütung: 2 Versuchspersonenstunden.
- Dauer der Studie: circa 3 Wochen.
- Inter-Annotator Agreement
  - TASD: Micro-F1 50.29
  - ACSA: Micro-F1 82.31 | Krippendorff's alpha 65.88

Annotation: Experten

- Überarbeitung des Trainingsdatensatzes durch einen ABSA-Task-Experten.
- Ursprüngliche Annotation: Annotation durch eine Person, mit anschließender Überprüfung durch eine weitere Person.
- Die ursprünglichen Annotationen werden als Grundlage für die Überarbeitung gewählt.
    - Anpassung der Annotationsanleitung.
    - Überprüfung auf Konsistenz.
- Kein IAA, da nur ein Experte.
- Änderungen:
    - 1333 Triplets sind identisch geblieben.
    - 194 Triplets wurden geändert.

Datensatzvergleich



Auswertung

- Klassifizierung (ACSA)
  - BERT-CLF [10]
  - Hier-GCN [11]
- Textgenerierung (TASD)
  - MvP [12]
  - Paraphrase [13]
- LLM (Beispiele wurden aus den jeweiligen Datensätzen entnommen für die Tasks ACSA & TASD)
  - Few-Shot Prompting [14]
  - Instruction Fine-Tuning [15]
- Hyperparameter (Batchgröße, Epochen, Lernrate) wurden nach der Referenzimplementierung in [16] übernommen.

[10] 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
[11] Cai, H., Tu, Y., Zhou, X., Yu, J., & Xia, R. (2020, December). Aspect-category based sentiment analysis with hierarchical graph convolutional network. In Proceedings of the 28th international conference on computational linguistics (pp. 833-843).
[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] Šmíd, J., Přibáň, P., & Král, P. (2025). LLaMA-based models for aspect-based sentiment analysis. arXiv preprint arXiv:2508.08649.
[15] Simmering, P. F., & Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. arXiv preprint arXiv:2310.18025.
[16] Fehle, J., Donhauser, N., Kruschwitz, U., Hellwig, N. C., & Wolff, C. (2025, September). German Aspect-based Sentiment Analysis in the Wild: B2B Dataset Creation and Cross-Domain Evaluation. In *21st Conference on Natural Language Processing (KONVENS 2025)* (Vol. 9, p. 213).

Ergebnisse: ACSA

- Auswertung: Tupel werden nicht doppelt gezählt (z. B. *Service-Positiv, Service-Positive* → *Service-Positiv*).
- Expertendatensatz liefert die höchsten Werte: Alle Methoden erzielten auf dem Expertendatensatz die höchsten Micro- und Macro-F1-Werte.
- LLMs übertreffen klassische Modelle: LLaMA FT und Gemma FS schneiden generell besser ab als BERT-CLF und Hier-GCN.

Aspect Category Sentiment Analysis (ACSA)

Dataset	Students		LLM		Experts	
Method	Micro	Macro	Micro	Macro	Micro	Macro
BERT-CLF	77.79	74.50	77.44	74.65	**78.26**	75.17
Hier-GCN	79.21	76.73	79.13	77.17	**79.78**	78.13
Gemma FS	86.05	84.22	85.60	83.83	**86.29**	84.62
LLaMA FT	85.73	83.42	84.85	82.02	**86.39**	83.12

Ergebnisse: TASD

- Auch hier erreichen alle Methoden mit dem Expertendatensatz die höchsten Micro- und Macro-F1 Werte.
- LLaMA FT erzielt durchgehend die höchsten Werte, mit überdurchschnittlich großem Abstand auf dem Expertendatensatz.
- Gemma FS zeigt auf dem LLM-Datensatz sehr gute Performance, diese ist jedoch nicht auf den Expertendatensatz übertragbar.

Target Aspect Sentiment Detection (TASD)

Dataset	Students		LLM		Experts	
Method	Micro	Macro	Micro	Macro	Micro	Macro
Paraphrase	57.33	52.10	57.37	53.06	**61.65**	56.24
MvP	56.83	55.89	60.65	56.95	**64.01**	59.42
Gemma FS	62.28	59.06	65.58	62.30	**63.38**	58.74
LLaMA FT	69.33	65.50	66.24	63.20	**71.47**	67.40

Erkenntnisse

Annotation

- Strukturierte Anleitungen und Interfaces wichtig:
  - Klare und gut strukturierte Annotationsanleitungen und Interfaces sind entscheidend, um Fehler während der Annotation zu vermeiden.
- LLM-Annotationen nahe an Experten-Annotation:
  - Few-Shot + Self Consistency LLMs liefern qualitativ hochwertige Daten, oft nahe an Expertendaten.

Auswertung

- Experten-Annotation steigern die Qualität:
  - Alle Methoden erzielen auf den Expertendatensatz die höchsten Micro- und Macro-F1-Werte.
- LLMs übertreffen klassische Modelle:
  - LLM-Ansätze schneiden meist besser ab als klassische Klassifikations- oder Textgenerierungsmodelle.

Limitation & Future Work

Limitation

- Expertenannotation liefert hohe Qualität, dies ist jedoch sehr zeitaufwendig.
- Erkenntnisse nur in einer Domäne (Restaurant Reviews).
- Kostenfaktor bei der Crowd-Annotation (700 Euro).

Future Work

- Domain-Transfer & Robustheit:
  - Test auf anderen Datensätzen oder Domänen, um Generalisierung zu evaluieren.
- LLM-Vorannotation mit Expertenprüfung:
  - LLMs generieren erste Labels, Experten validieren diese.

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 (Few-shot)
- Studierende
- Experten

Ergebnisse
- Mit geringem Aufwand können bereits gute Ergebnisse erzielt werden, aber die Leistung variiert je nach Aufgabe (ACSA / TASD), sodass die Methode stets auf Basis der Anforderungen und Ausgangslage ausgewählt werden sollte.

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] 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
[8] Hellwig, N. C., Fehle, J., Kruschwitz, U., & Wolff, C. (2025, May). Do we still need human annotators? Prompting large language models for aspect sentiment quad prediction. arXiv. https://doi.org/10.48550/arXiv.2502.13044
[9] 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*.
[10] 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
[11] Cai, H., Tu, Y., Zhou, X., Yu, J., & Xia, R. (2020, December). Aspect-category based sentiment analysis with hierarchical graph convolutional network. In Proceedings of the 28th international conference on computational linguistics (pp. 833-843).
[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] Šmíd, J., Přibáň, P., & Král, P. (2025). LLaMA-based models for aspect-based sentiment analysis. arXiv preprint arXiv:2508.08649.
[15] Simmering, P. F., & Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. arXiv preprint arXiv:2310.18025.
[16] Fehle, J., Donhauser, N., Kruschwitz, U., Hellwig, N. C., & Wolff, C. (2025, September). German Aspect-based Sentiment Analysis in the Wild: B2B Dataset Creation and Cross-Domain Evaluation. In *21st Conference on Natural Language Processing (KONVENS 2025)* (Vol. 9, p. 213).