Möglichkeiten und Grenzen der KI-gestützten Annotation am Beispiel von Emotionen in Lyrik

Kröncke, Merten
merten.kroencke@uni-goettingen.de
Universität Würzburg, Deutschland
ORCID: 0000-0003-2717-0598

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Göttingen, Deutschland
ORCID: 0000-0001-6944-6113

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Universität Göttingen, Deutschland
ORCID: 0000-0001-5833-0414

Winko, Simone
simone.winko@phil.uni-goettingen.de
Universität Würzburg, Deutschland
ORCID: 0000-0002-1006-7925

Einleitung

Die rasante Entwicklung der Verarbeitung natürlicher Sprache durch neuronale Netze hat in den letzten 11 Jahren auch die Arbeitsweisen der digitalen Geisteswissenschaften deutlich verändert. Die Entwicklung neuronaler Architekturen hat zwei Ansätze ermöglicht, die auch heute das NLP bestimmen: 1. ‘Finetuning’: Ein Sprachmodell wird auf vielen Daten vortrainiert und dann auf deutlich weniger Daten für eine bestimmte Aufgabe feinjustiert. 2. ‘Chat’: Sehr große Sprachmodelle werden auf sehr vielen Daten vortrainiert und dann in einem zweiten Schritt auf die Kommunikation mit Anwendern hin eingerichtet. Der Finetuning-Ansatz hat sich schnell in den Digital Humanities durchgesetzt. Allerdings ist er mit vergleichsweise hohen Kosten verbunden, da die Leistungsfähigkeit für das Finetuning stark mit der Anzahl der Trainingsbeispiele korreliert. Deswegen ist die Verwendung von sehr großen Sprachmodellen ohne eine größere Anzahl von Trainingsbeispielen (*zero-shot* oder *few-shot prompting*) besonders interessant, schließlich muss in einem solchen Kontext nur eine kleine Menge von Testdaten annotiert werden. Eine Antwort auf die Frage, ob in einem Forschungsprojekt der klassische Finetuning-Ansatz durch *zero-shot* oder *few-shot prompting* in sehr großen Sprachmodellen ersetzt werden kann, ist nicht einfach, da die Antwort von der Komplexität der Aufgabenstellung ebenso abhängt wie vom Zeitpunkt, zu dem man die Frage stellt: Die Finetuning-Ansätze entwickeln sich ebenso weiter wie die sehr großen Sprachmodelle. Dazu kommen pragmatische Fragen: Hat die Arbeitsgruppe Zugriff auf die technische Infrastruktur, die man für das Finetuning von großen Sprachmodellen benötigt? Hat sie die finanziellen Ressourcen, um die kommerziellen Modelle für umfangreiche Annotationsaufgaben zu verwenden? Unser Beitrag will eine (wenn auch nur temporär gültige) Antwort für einen bestimmten Bereich liefern, die Annotation von Emotionen in literarischen Texten, und dadurch zugleich an der Diskussion darüber mitwirken, wie in den DH eine Antwort auf jene Frage gefunden werden kann.

Grundlage für unsere Arbeit sind die Annotationen von lyrischen Texten im Rahmen des DFG-Projekts *The Beginnings of Modern Poetry* (https://uni-goettingen.de/moderne-lyrik/). Die annotierten Texte haben wir drei großen Sprachmodellen mit der Aufgabe vorgelegt, jeweils eine Strophe mit Blick auf das Vorkommen von sechs Emotionsgruppen zu annotieren. Dabei haben wir nach einigen Vorstudien systematisch zwei Aspekte variiert: kurzer vs. langer Prompt und einfach randomisiertes vs. stratifiziert randomisiertes Sampling.

Forschungsstand

Die Möglichkeit, ChatGPT und verwandte Dienste zur Annotation von Daten zu verwenden, wurde sehr schnell erkannt. Ding et al. 2023 beobachten bei wenigen und klar definierten Labeln gute bis sehr gute Resultate, sehen allerdings auch eine deutliche Varianz abhängig vom Prompt. Ähnlich optimistische Ergebnisse haben Gilardi et al. 2023 erzielt. Törnberg 2023 vergleicht ChatGPT4s Annotationen von Tweets – wird eine republikanische oder eine demokratische Position vertreten? – mit denen von Expert:innen und von Arbeitern von Mechanical Turk und kommt zu dem Ergebnis, dass die Ergebnisse von ChatGPT deutlich besser und konsistenter sind als die der beiden menschlichen Gruppen. Reiss 2023 warnt allerdings davor, ChatGPT als Annotationswerkzeug ohne manuelle Datenvalidierung zu verwenden, da das System sehr empfindlich auf die Manipulation einzelner Wörter und Einstellungen reagiert. Rebora et al. 2023 vergleichen für die Aufgabe der Sentiment Analysis ChatGPT mit einem Finetuning-Modell und kommen zu dem Ergebnis, dass letzteres immer noch bessere Ergebnisse liefert (ähnlich Wang 2023).

Die Diskrepanzen zwischen den Ergebnissen lassen sich durch drei Aspekte gut erklären: Wie schwierig ist die Aufgabe? Named Entity Recognition ist einfacher als Sentiment Analyse usw. Welches Modell wurde verwendet? Alle Aufsätze, die wir gesichtet haben, verwenden (auch) ChatGPT, aber OpenAI bietet zu einem Zeitpunkt un-

terschiedliche Modelle mit unterschiedlichen Leistungsniveaus an – und die Modelle werden laufend aktualisiert. Was ist der Referenzpunkt des Vergleichs? Zum einen geht es um die Frage, ob man menschliche Annotator:innen durch große Sprachmodelle ersetzen kann, zum anderen darum, ob ChatGPT & Co. die Leistungsfähigkeit von Finetuning-Modellen erreichen.

Für die Promptgestaltung haben wir uns an den Empfehlungen von https://www.promptingguide.ai/ orientiert. Nach dem Überblick von Vatsal und Dubey 2024 gibt es keine klare Empfehlung zum Prompting bei Emotionsannotationen.

Ressourcen

Das Untersuchungskorpus besteht aus Texten in Lyrikanthologien, die sich auf Gedichte von Zeitgenoss:innen konzentrieren. Die Anthologien stammen aus der Zeit von 1859 bis 1919 und enthalten mehr als 6000 Gedichte, von denen 1412 manuell annotiert wurden (vgl. Winko et al. 2022a, Winko et al. 2022b).

Die Emotionsannotation zielt darauf ab, die im Text gestalteten Emotionen (und nicht die Emotionen der Leser:innen) zu erfassen. Möglich war, einer Textstelle genau eine, aber auch keine oder mehrere Emotionen zuzuweisen. Genutzt wurde ein Set von 40 diskreten Emotionen, darunter zum Beispiel Hoffnung, Sehnsucht oder Hass. Die Annotationseinheiten sind Wörter bzw. Wortfolgen (vgl. Kröncke et al. 2022). Da für viele einzelne Emotionen nur eine sehr geringe Zahl von Annotationen vorliegt, werden die Emotionen nachträglich zu sechs Gruppen zusammengefasst, orientiert an Shaver et al. 1987: Liebe (Love), Freude (Joy), Trauer (Sadness), Erregung/Überraschung (Agitation), Angst (Fear) und Wut (Anger).

Für das maschinelle Lernen wurde der Task leicht angepasst. Um die Komplexität der Aufgabe und den Rechenaufwand zu reduzieren, haben wir eine bestimmte Segmentierung vorgegeben, nämlich die Einheit ‘Strophe’. Die Multi-Label-Klassifikation basiert auf dem (mithilfe einer großen Zahl manueller Annotationen trainierten) Modell SauerkrautLM-7B-HerO und wurde für die sechs Emotionsgruppen nach Shaver et al. 1987 durchgeführt.

Sowohl das Korpus als auch das Annotationsverfahren und das maschinelle Lernen haben wir an anderen Stellen bereits ausführlicher erläutert (Konle et al. 2022, Konle et al. 2024).

Die folgenden Experimente basieren auf zwei Samples aus dem Gesamtkorpus: Zum einen verwenden wir ein einfach randomisiertes Sample (350 Strophen), das die im Korpus *de facto* vorhandenen Häufigkeitsverhältnisse der Emotionsklassen widerspiegeln soll, zum anderen ein stratifiziert randomisiertes Sample (ebenfalls 350 Strophen: 50 x jede der 6 Emotionsgruppen + 50 Strophen ohne Emotion), das eine gewisse Mindesthäufigkeit pro Emotionsgruppe garantiert, aber durch die Kookkurrenz von Emotionsgruppen ebenfalls nicht zu einer Gleichverteilung führt.[1]

Tabelle 1: Anzahl von Strophen mit Emotionsgruppe je Korpussample

	total	Agitation	Fear	Anger	Sadness	Joy	Love	No Emotion
Simple random Sample	350	23	12	14	92	89	94	133
Stratified random Sample	350	64	64	68	125	124	124	50

Da eine Strophe mehrere Emotionsgruppen enthalten kann, ist die Summe der Strophen pro Emotionsgruppe in einer Zeile größer als 350. Aus dem gleichen Grund gibt es im Fall des Stratified random Sample stets mehr als 50 Strophen pro Emotionsgruppe.

Methoden

In allen Experimenten lassen wir Chat-Modelle Fragen zu den Emotionen in lyrischen Texten beantworten. Der Task ist der gleiche, den bereits das Finetuning-Modell SauerkrautLM-7B-HerO übernommen hat, also die Zuweisung von keiner, einer oder mehreren der sechs Emotionsgruppen nach Shaver et al. 1987 zu einzelnen Strophen. Wir verwenden drei (kommerzielle) Modelle: GPT4o (OpenAI), Claude (Anthropic) und Gemini (Google).

In unseren Experimenten testen wir einen kurzen und einen langen Prompt. Der kurze Prompt enthält keine Erläuterungen der Emotionskonzepte und keine Annotationsbeispiele, der lange Prompt schon. Die Gestaltung des langen Prompts ist durch verschiedene Vorexperimente informiert. Ziel ist unter anderem, die (ansonsten zu große) Häufigkeit, mit der Emotionen zugewiesen werden, zu reduzieren. Wir setzen explizites CoT-Prompting ein und weisen darauf hin, dass hinreichend starke Emotionsindikatoren vorliegen müssen. Insgesamt ergeben sich vier Experimente:

1. Simple random Sampling. Short Prompt
2. Simple random Sampling. Long Prompt
3. Stratified random Sampling. Short Prompt
4. Stratified random Sampling. Long Prompt

Ergebnisse

Tabelle 2 und 3 geben Auskunft über die Performance der Emotionserkennung. Für das Modell Claude 3.5 Sonnet können wir in der Variante Stratified random Sample – Long Prompt noch keine Ergebnisse mitteilen, da unsere langen Prompts bei Anthropic wiederholt zur Überschreitung des rate limits geführt haben. Wir planen, das Experiment nachzuholen.

Tabelle 2: Performance der Emotionserkennung (Simple random Sample) (F1)

	Agitation	Fear	Anger	Sadness	Joy	Love
SauerkrautLM-7B-HerO	.74	.83	.82	.7	.7	.73
GPT4o (Short Prompt)	0.0	0.24	0.43	0.62	0.5	0.57
Gemini 1.5 (Short Prompt)	0.0	0.16	0.3	0.57	0.53	0.6
Claude 3.5 Sonnet (Short Prompt)	0.0	0.18	0.35	0.57	0.49	0.62
GPT4o (Long Prompt)	0.19	**0.34**	**0.5**	0.63	0.5	0.66
Gemini 1.5 (Long Prompt)	0.23	0.3	0.43	0.66	0.54	0.61
Claude 3.5 Sonnet (Long Prompt)	**0.25**	0.33	0.38	**0.67**	**0.56**	**0.72**

Tabelle 3: Performance der Emotionserkennung (Stratified random Sample) (F1)

	Agitation	Fear	Anger	Sadness	Joy	Love
SauerkrautLM-7B-HerO	.74	.83	.82	.7	.7	.73
GPT4o (Short Prompt)	0	0.65	0.46	0.67	0.48	0.63
Gemini 1.5 (Short Prompt)	0.0	0.62	0.52	0.65	0.5	0.65
Claude 3.5 Sonnet (Short Prompt)	0.0	0.55	0.46	0.63	0.53	0.66
GPT4o (Long Prompt)	0.19	**0.66**	**0.65**	**0.71**	**0.6**	**0.68**
Gemini 1.5 (Long Prompt)	**0.21**	0.62	0.5	0.67	0.52	0.66
Claude 3.5 Sonnet (Long Prompt)	/	/	/	/	/	/

bei hier für Claude 3.5 Sonnet in der Long-Prompt-Version keine Daten verfügbar sind).

Die bisherigen Beobachtungen haben sich am F1-Score orientiert. Unterscheidet man Precision und Recall, werden weitere Befunde sichtbar. Das gilt nicht zuletzt für die Erkennung von solchen Strophen, die *keine* Emotion enthalten (vgl. Tabelle 4, exemplarisch für das Simple random Sample).

Tabelle 4: Erkennung von Strophen ohne Emotion (Simple random Sample)

	Strophen	Strophen ohne Emotion laut manueller Annotation	Strophen ohne Emotion laut Modell	F1	Precision	Recall
GPT4o (Short Prompt)	350	133	69	0.58	0.86	0.44
Gemini 1.5 (Short Prompt)	350	133	60	0.46	0.73	0.33
Claude 3.5 Sonnet (Short Prompt)	350	133	54	0.53	0.93	0.38
GPT4o (Long Prompt)	350	133	102	0.67	0.77	0.59
Gemini 1.5 (Long Prompt)	319	127	82	0.67	0.85	0.55
Claude 3.5 Sonnet (Long Prompt)	350	133	57	0.57	0.95	0.41

Diskussion

Die Performance aller Short-Prompt-Modelle bleibt hinter den Finetuning-Ergebnissen zurück, wenn auch unterschiedlich deutlich. Andere Studien sind auf Basis anderer Daten und anderer Tasks zu ähnlichen Ergebnissen gekommen (etwa Rebora et al. 2023).

Die Long-Prompt-Modelle performen besser als die Short-Prompt-Modelle. Ein besonders deutliches Beispiel ist die Emotionsgruppe Agitation, die von den Short-Prompt-Modellen gar nicht erkannt wird, möglicherweise weil der Begriff in unserem Annotationsdesign ein spezifisches Konzept bezeichnet, das mit der alltagssprachlichen Bedeutung des Worts ‘Agitation’ wenig gemein hat. [2] In einigen Fällen reicht die Performance der besten Long-Prompt-Modelle fast an die Finetuning-Ergebnisse heran, z. B. im Fall von ‘Love’, oder zieht gleich, etwa im Fall von ‘Sadness’.

Im Stratified random Sample performen die Modelle entweder ungefähr gleich gut oder deutlich besser (Anger, Fear) als im Simple random Sample. Relevant dafür ist allerdings, dass Anger und Fear im Simple random Sample nur selten vorkommen, weshalb die entsprechenden Ergebnisse nicht allzu belastbar sind.

Zwischen den drei Modellen GPT4o, Gemini 1.5 und Claude 3.5 Sonnet zeigen sich je nach Sample und je nach Prompt einige Unterschiede. Im Simple random Sample ist die Performance von GPT4o oder Claude 3.5 Sonnet am besten, im Stratified random Sample von GPT4o (wo-

Für alle Modelle und für alle Prompts gilt, [3] dass bei der Erkennung von Emotionslosigkeit die Precision höher als der Recall ist. Der Befund hängt damit zusammen, dass die Modelle den Strophen häufiger Emotionen (bzw. seltener *keine* Emotionen) als menschliche Annotator:innen zuschreiben. Erklärungsrelevant dürfte sein, dass manuell ‘sparsam’ annotiert werden sollte, auch mit Rücksicht auf das Inter-Annotator-Agreement. Die Modelle sind nicht an diese Annotationspraxis gebunden, wenngleich der lange Prompt sie (wie beabsichtigt) in diese Richtung zu lenken scheint, immerhin steigt hier der Recall, besonders bei GPT4o und Gemini 1.5.

Das wichtigste Ergebnis unserer Studie ist, dass die sehr großen Sprachmodelle auch bei einer komplexen Aufgabe wie der Annotation von Emotionen teilweise das Niveau von Finetuning-Modellen erreichen, aber die Ergebnisse abhängig von der Kategorie stark und in schwer zu prognostizierender Weise schwanken. Die große und nur teilweise transparente Varianz in Abhängigkeit von der Promptgestaltung gilt es ebenfalls zu berücksichtigen.

Zahlreiche weitere Studien sind denkbar. Aufschlussreich wäre, die Experimente auch für die 40 Einzelemotionen (statt nur für die 6 Emotionsgruppen) durchzuführen. Zudem lassen sich die Prompts anpassen, etwa insofern als den Modellen eine bestimmte Rolle zugewiesen wird (“Du bist Expertin für …”, “Du bist eine Person des 19. Jahrhunderts” usw.). [4] Des Weiteren wäre zu testen, ob sich die Performance ändert, wenn der Task modifiziert oder anders modelliert wird, zum Beispiel inklusive Segmentierung (die Teil der manuellen Annotation ist) und/oder als Reihe binärer Klassifikationen. Die binäre Klassifikation haben wir exemplarisch getestet. Es zeigte sich, dass das

Modell nun *seltener* (statt wie im bisherigen Setup häufiger) als menschliche Annotator:innen Emotionsgruppen zuweist. [5] Der Befund deutet abermals auf die große Varianz des Modellverhaltens hin. Schließlich wäre informativ, das Inter-Annotator-Agreement zwischen menschlichen Annotator:innen mit dem Agreement zwischen Sprachmodellen zu vergleichen.

Dass die Performanz je nach Kategorien und Prompts stark variiert, veranlasst uns zu folgendem Schluss: Auch wenn die Sprachmodelle ständig verbessert werden, wird man wohl auf absehbare Zeit nicht ohne die Entwicklung von Annotationsguidelines und die Annotation von ausreichend Testdaten auskommen.

Fußnoten

1. Daten und Code: https://github.com/MertenKroencke/DHd2025_LLMs.
2. Die Emotionsgruppe ‘Agitation’ besteht aus vier Subemotionen: (unspezifische) Emotionalität, Überraschung, Spannung, Aufregung. Aufregung – diese Subemotion entspricht sprachlich am ehesten ‘Agitation’ – macht in den Samples weniger als 10% der Agitation-Annotationen aus. Bei Agitation handelt es sich zu über 80% um (unspezifische) Emotionalität.
3. Gemini 1.5 hat bei einigen Strophen auf den Prompt keine Antwort gegeben. Diese Strophen schließen wir aus der Auswertung aus.
4. Vergleichbares könnte auch bei der Gestaltung des System Prompts verwendet werden. Das systematische Variieren der Temperatur hat dagegen nach Törnberg keinen größeren Effekt (Törnberg 2023).
5. Binäre Klassifikation mit ChatGPT4o (short prompt) F1-Score: Agitation: 0.5, Anger: 0.56, Fear: 0.75, Joy: 046, Love: 0.66 , Sadness: 0.76. In allen Fällen (bis auf Sadness) weist das Modell den Strophen die Emotionsgruppe seltener als menschliche Annotator:innen zu.

Bibliographie

Ding, Bosheng, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li und Lidong Bing. 2023. “Is GPT-3 a Good Data Annotator?” arXiv. http://arxiv.org/abs/2212.10450 .

Gilardi, Fabrizio, Meysam Alizadeh und Maël Kubli. 2023. “ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks.” In Proceedings of the National Academy of Sciences 120. 30: e2305016120. https://doi.org/10.1073/pnas.2305016120 .

Konle, Leonard, Merten Kröncke, Fotis Jannidis und Simone Winko. 2022. “Emotions and Literary Periods.” In *DH2022: Responding to Asian Diversity. Conference Abstracts, July 25–29, 2022, Tokyo, Japan* , 278–281.

Konle, Leonard, Merten Kröncke, Fotis Jannidis und Simone Winko. 2024. “On the Unity of Literary Change. The Development of Emotions in German Poetry, Prose, and Drama between 1850 and 1920 as a Test Case.” In *CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark* [eingereicht].

Kröncke, Merten, Fotis Jannidis, Leonard Konle und Winko, Simone. 2022. “Annotationsrichtlinien Emotionsmarker und Emotionen.” Zenodo. https://doi.org/10.5281/zenodo.6021152 .

Rebora, Simone, Marina Lehmann, Anne Heumann, Wei Ding und Gerhard Lauer. 2023. “Comparing ChatGPT to Human Raters and Sentiment Analysis Tools for German Children’s Literature.” In CHR 2023: Computational Humanities Research Conference, December 6–8, 2023, Paris, France , 333–343.

Reiss, Michael V. 2023. “Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark”. arXiv. https://doi.org/10.48550/arXiv.2304.11085 .

Shaver, P., J. Schwartz, D. Kirson und C O’Connor. 1987. “Emotion Knowledge: Further Exploration of a Prototype Approach.” In Journal of Personality and Social Psychology 52.6: 1061–1086.

Törnberg, Petter. 2023. “ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning”. arXiv. https://doi.org/10.48550/arXiv.2304.06588 .

Vatsal, Shubham und Harsh Dubey. 2024. “A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks.” arXiv. https://doi.org/10.48550/arXiv.2407.12994 .

Winko, Simone, Konle, Leonard, Kröncke, Merten und Fotis Jannidis. 2022a. “Lyrik-Anthologien 1850-1910.” Zenodo. https://doi.org/10.5281/zenodo.6053952 .

Winko, Simone, Konle, Leonard, Kröncke, Merten und Fotis Jannidis. 2022b. “Korpusbeschreibung der Lyrik-Anthologien 1850-1910.” Zenodo. https://doi.org/10.5281/zenodo.6204787 .