Gregor Willenbrock
Institut für Kommunikationswissenschaft

Mini-Workshop: Zero-Shot mit der OpenAI-API

Primer für Zero-Shot-Klassifikation mit llms // Dresden, 02.12.2025

Bisheriges Vorgehen bei automatischer Textklassifikation (ohne llms)

Diktionärsbasierte Ansätze

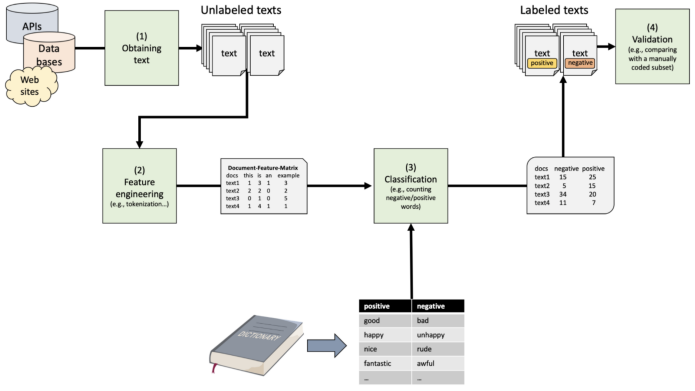

Abbildung von Masur (2024).

Supervised Textklassifikation

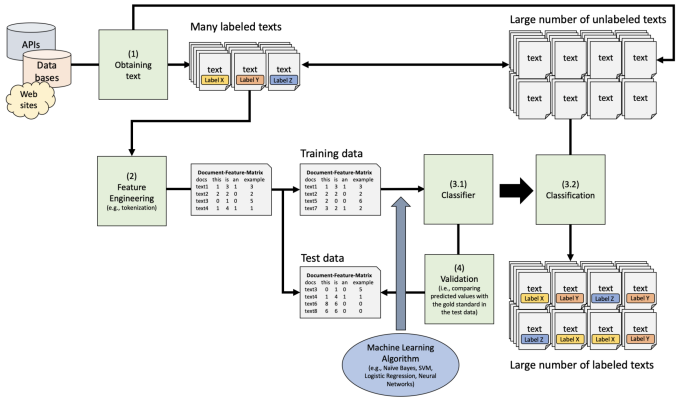

Abbildung von Masur (2024).

Klassifikation mit Word Embeddings (Pretrained Models)

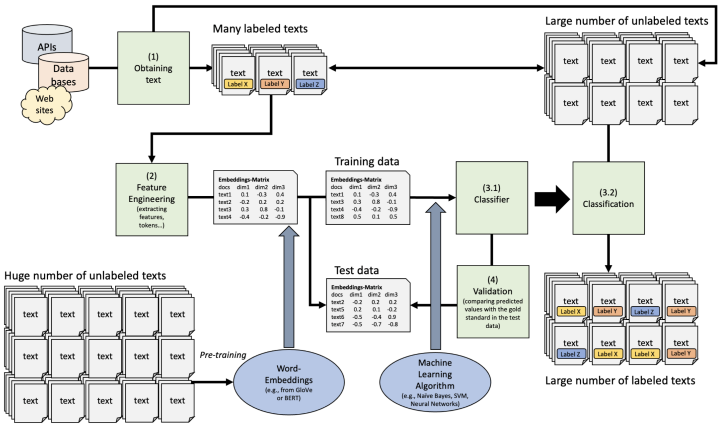

Abbildung von Masur (2024).

Jetzt: Klassifikationen direkt mit LLMs

Klassifikation mit LLMs

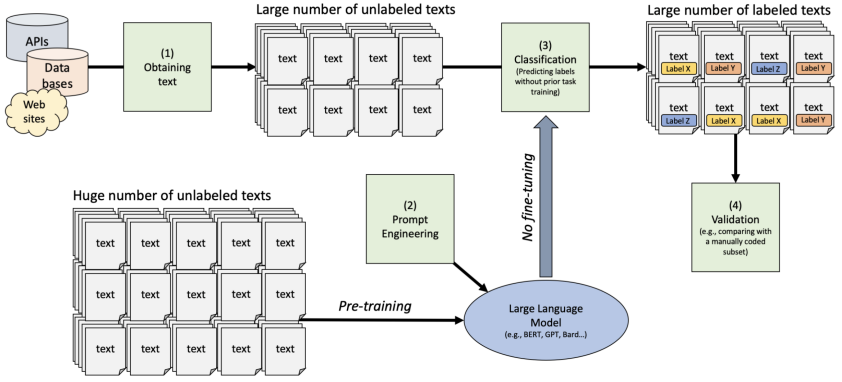

Abbildung von Masur (2024).

Arten von “Promptengineering”

	Zero-Shot Classification	One-Shot Classification	Few-Shot Classification
Definition	Providing no examples	Providing one example	Providing a few examples
Example prompt	In the following social media posts, look for instances where people offer constructive feedback. 1 = Constructive Feedback present 0 = Constructive Feedback absent	In the following social media posts, look for instances where people offer constructive feedback. 1 = Constructive Feedback present 0 = Constructive Feedback absent ***Example:*** You should always keep two offers on the table before accepting an offer.” ***Classification:*** 1	In the following social media posts, look for instances where people offer constructive feedback. 1 = Constructive Feedback present 0 = Constructive Feedback absent ***Example:*** You should always keep two offers on the table before accepting an offer.” ***Classification:*** 1 ***Example:*** Haha, that’s life.” ***Classification:*** 0
Strengths	The model is free in how it interprets the prompt, which can mean it better choses from the range of possibilities	The model has an example to draw from, but because it is just one, it usually doesn’t limit its flexibility	The model has clear-cut rules with regard to how to code the texts. Helps in streamlining the codes
Weaknesses	Can lead to unwanted codes or completely unreliable coding	Less open than zero-shot	Too many examples can constrain the models ability to generalize beyond the examples

Abbildung von Masur (2024).

Qualität der Codierung (im Vergleich zur menschlichen)

- Schwierige Studienlage: Viele Preprints, häufig alte Modelle und hauptsächlich OpenAI-zentriert (vgl. auch Ollion et al., 2023)
- Qualität variiert, aber häufig akzeptabel
  - `gpt4-turbo` Crohnbachs $\alpha = .78$ bei Pilny et al. (2024)
  - `gpt4-turbo` bei unterschiedlichen Klassifikationsaufgaben (bspw. Likert-Skala zum Vorhandensein unterschiedlicher Emotionen) in unterschiedlichen Sprachen durchschnittlich eine *Accuracy* von .682 (Rathje et al., 2024)
- Meistens besser als andere Machine-Learning-Ansätze
  - GPT (`gpt4-turbo`) übertrifft die Genauigkeit von Dictionary-Methoden (Ollion et al., 2023; Rathje et al., 2024)
  - `gpt4-turbo` funktioniert gut mit Zero-Shot, allerdings ist es nicht unbedingt besser als fine-tuned Modelle wie BERT. (Kristensen-McLachlan et al., 2025; Rathje et al., 2024)
- Interne Konsistenz ist stark abhängig vom Prompt (auch bei kleinen Änderungen wie z.B. “classify” vs. “rate”) (Reiss, 2023) und vom Inhalt/Kontext des zu codierenden Materials (Gielens et al., 2025)

Ethische Bedenken zur Codierung mit LLMs

- Im Fazit der Studien ist häufig eine Warnung vor unreflektierter/ unkontrollierter Nutzung, aber keine Warnung vor der Nutzung per se zu finden (Gielens et al., 2025; Ollion et al., 2023; Törnberg, 2024)
- OpenAI (und ähnliche Anbieter) nutzt Eingaben fürs weitere Training: Problematisch bei sensiblen, privaten oder urheberrechtlich geschützten Daten (Ollion et al., 2023; Rathje et al., 2024; vgl. auch Spirling, 2023)
- Laufende Modellanpassungen erschweren Replizierbarkeit, später ggf. auch Reproduzierbarkeit (Kristensen-McLachlan et al., 2025)
- Modellbias bleibt eine schwer absehbare zusätzliche Problemdimension (bspw. Gupta et al., 2023)

FOSS als Lösung

Nach Spirling (2023) können offene Modelle diese Bedenken aus dem Weg räumen. Notwendige Ressourcen: Hardwareinfrastruktur und Know-how.

Gielens et al. (2025): *“Goodbye human annotators? Content analysis of social policy debates using ChatGPT”* I

System Prompt

Persona: You are a professional researcher named Jakub. You are an expert on qualitative content analysis. You are always focussed and rigorous. Task Description: Analyse [language] [document_type] for arguments related to [policy_name]. [policy_description]. The analysis will identify whether [document_type] contain arguments for or against [policy_name].

Gielens et al. (2025): *“Goodbye human annotators? Content analysis of social policy debates using ChatGPT”* II

> *For each [document_type], provide a classification for each argument in an HTML table. Do not include the text of the [document_type] in the table. Only report the classification values. The HTML table has 5 rows, one per [document_type]. The HTML table has 10 columns, one per argument. The elements of the table are “0” and “1”. Indicate “1” if the [document_type] discusses aspects of the specified argument and “0” is the [document_type] does not discuss the specific argument. Here is an example of the required output format: [example_output]*

Gielens et al. (2025): *“Goodbye human annotators? Content analysis of social policy debates using ChatGPT”* III

User Prompt

Determine whether a [document_type] discusses each of the following ten arguments: [[arguments]] [document_type] contain an argument if the author opposes the argument and also when the author argues in favour of the argument. [policy_name] need not be mentioned explicitly in the [document_type] to relate to the argument. A [document_type] can discuss more than one argument. You will now be provided with 5 [document_type] separated by a new line. [[documents]]

Gielens et al. (2025): *“Goodbye human annotators? Content analysis of social policy debates using ChatGPT”* IV

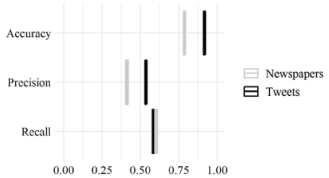

Abbildung 1: Gesamtgüte

- **Accuracy** = (TN +TP) / (TN +TP + FN + FP) = Anteil aller korrekt klassifizierten Fälle (positiv und negativ) an allen Fällen
- **Precision** = TP / (TP+FP) = Anteil der als positiv klassifizierten Fälle, die wirklich positiv sind
- **Recall/Sensitivität** = TP / (TP+FN) = Anteil der tatsächlich positiven Fälle, die korrekt als positiv erkannt wurden

Gielens et al. (2025): *“Goodbye human annotators? Content analysis of social policy debates using ChatGPT”* V

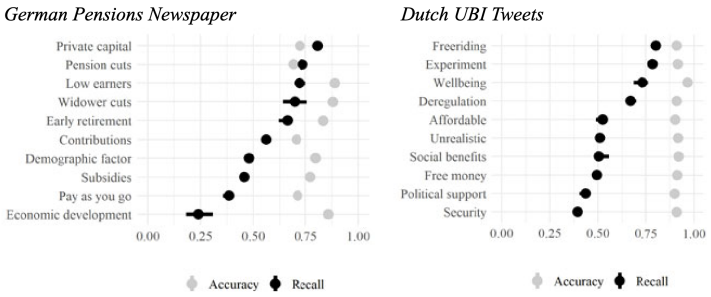

Abbildung 2: Güte nach Argument

Qualität der Codierung (im Vergleich zur Ground Truth)

Törnberg (2025): *“[llms] Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages”* I

- Aufgabe: Tweets von US-Parlamentarier:innen nach der politischen Zugehörigkeit der Verfassenden codieren.
- Daten aus einer bestehenden Datenbank, in der die Parteizugehörigkeit bekannt ist (“ground truth”).
- GPT-4 objektiv mit menschlichen Codierenden und alternativen ML-Methoden vergleichen.
- Komplexes Material:
  - leicht einordenbare Parteipropaganda oder Angriffe auf Gegner
  - Botschaften, deren Intention und Zielpublikum interpretiert werden müssen
  - inhaltlich neutrale oder private Nachrichten, die politisch kaum zuordenbar sind

Törnberg (2025): *“[llms] Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages”* II

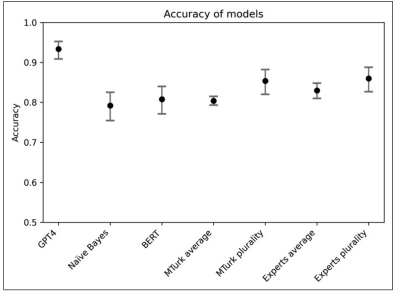

Abbildung 3: Törnberg (2025): Accuracy im Modellvergleich mit 95%-Konfidenzintervall.

Weiter ins Thema

- Fine-Tuning (open-source) llms: Alizadeh et al. (2025)
- Model Bias: Gupta et al. (2023)
- Zero-shot Best Practices: Törnberg (2024)

Quellen I

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2025). Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning. *Journal of Computational Social Science*, *8*(1), 17. https://doi.org/10.1007/s42001-024-00345-9

Gielens, E., Sowula, J., & Leifeld, P. (2025). Goodbye Human Annotators? Content Analysis of Social Policy Debates Using ChatGPT. *Journal of Social Policy*, 1–20. https://doi.org/10.1017/S0047279424000382

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2023). *Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs*. https://doi.org/10.48550/ARXIV.2311.04892

Kristensen-McLachlan, R. D., Canavan, M., Kárdos, M., Jacobsen, M., & Aarøe, L. (2025). Are Chatbots Reliable Text Annotators? Sometimes. *PNAS Nexus*, *4*(4), pgaf069. https://doi.org/10.1093/pnasnexus/pgaf069

Quellen II

Masur, P. (2024). *Computational Analysis of Digital Communication* [Github-Repository]. Course at Vrije Universiteit Amsterdam. https://github.com/masurp/VU_CADC

Ollion, E., Shen, R., Macanovic, A., & Chatelain, A. (2023, Oktober 4). *ChatGPT for Text Annotation? Mind the Hype!* https://doi.org/10.31235/osf.io/x58kn

Pilny, A., McAninch, K., Slone, A., & Moore, K. (2024). From Manual to Machine: Assessing the Efficacy of Large Language Models in Content Analysis. *Communication Research Reports*, *41*(2), 61–70. https://doi.org/10.1080/08824096.2024.2327547

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT Is an Effective Tool for Multilingual Psychological Text Analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Reiss, M. V. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark*. https://doi.org/10.48550/ARXIV.2304.11085

Spirling, A. (2023). Why Open-Source Generative AI Models Are an Ethical Way Forward for Science. *Nature*, *616*(7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4

Quellen III

Törnberg, P. (2024). *Best Practices for Text Annotation with Large Language Models*. https://doi.org/10.48550/ARXIV.2402.05129

Törnberg, P. (2025). Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. *Social Science Computer Review*, *43*(6), 1181–1195. https://doi.org/10.1177/08944393241286471