APIs für Datenerhebung und Inhaltsanalyse

Vorlesung: Methoden der empirischen Kommunikations- und Medienforschung I / Datenerhebung, Wintersemester 2024/2025

Marko Bachl
Freie Universität Berlin

20. 01. 2025

Hallo

ARBEITSSTELLE DIGITALE FORSCHUNGSMETHODEN

dall-e-3, Prompt: A team of communication researchers using digital research methods and computational methods, cyberpunk style

DIESES BILD HABE ICH MIT EINER API ERSTELLT

```r
library(httr2)
library(jsonlite)

key = readLines("openai_key.txt")

req = request("https://api.openai.com/v1/images/generations") |>
  req_headers(
    "Content-Type" = "application/json",
    "Authorization" = paste0("Bearer ", key)
  ) |>
  req_body_json(list(
    model = "dall-e-3",
    prompt = "A team of communication researchers using digital research methods and computational methods, cyb
    n = 1,
    size = "1024x1024"
  )) |>
  req_perform()
```

Nachmachen: 👉 `bsp_dall-e.R` (mit OpenAI-Account, $0.04/Bild)

ARBEITSSTELLE DIGITALE FORSCHUNGSMETHODEN

Real life

PRÄSENTATION UND CODE

Material: Präsentation HTML, Präsentation PDF, Code

AGENDA

1. Was ist eine API?
2. Verbreiteter Einsatz in PuK: Erhebung digitaler Inhalte
   a. Vor und nach der *APIcalypse*
   b. Umsetzung mit `R` und `{httr2}`
3. Neuerer Einsatz in PuK: Nutzung von Cloud-Diensten (z.B. KI)
   a. Zero-shot classification: Kurze Einführung
   b. Zero-shot classification: Umsetzung mit der OpenAI-API

Was ist eine API?

WAS IST EINE API?

- **A**pplication **P**rogramming **I**nterface = Programmierschnittstelle
  - Austausch maschinenlesbarer Daten zwischen verschiedenen Programmen/Computern

- **Web-APIs** nutzen die gleichen Protokolle wie Browser, aber liefern anderen Datenstrukturen
  - Formate sind standardisiert (z.B. XML oder JSON), Inhalte variieren
  - oft nutzen Plattformen für ihre eigenen (Mobil-) Apps ebenfalls APIs

EIN EINFACHES BEISPIEL

Breaking Bad Quotes API

A free API to retrieve some quotes of Breaking Bad, bitch!

☆ Star

> *Shut the f*ck up and let me die in peace.*
>
> *Mike Ehrmantraut*

Breaking Bad Quotes API

EIN EINFACHES BEISPIEL

Anfrage

```
library(httr2); library(jsonlite)
bb_quote = request(base_url = "https://api.breakingbadquotes.xyz/v1/quotes") |> req_perform()
```

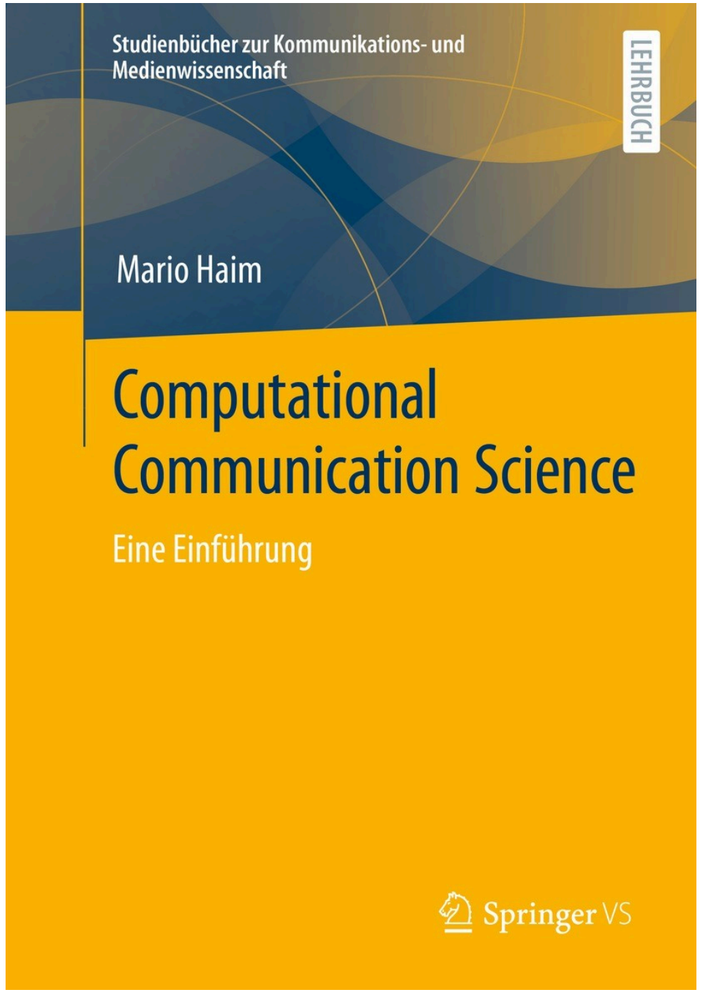
Antwort

```
bb_quote |> resp_body_string() |> prettify()
```

```
[
    {
        "quote": "Because I say so.",
        "author": "Walter White"
    }
]
```

GRUNDBEGRIFFE

NACHLESEN



(Haim, 2023, Kapitel 5.3)

Verbreiteter Einsatz in PuK: Erhebung digitaler Inhalte

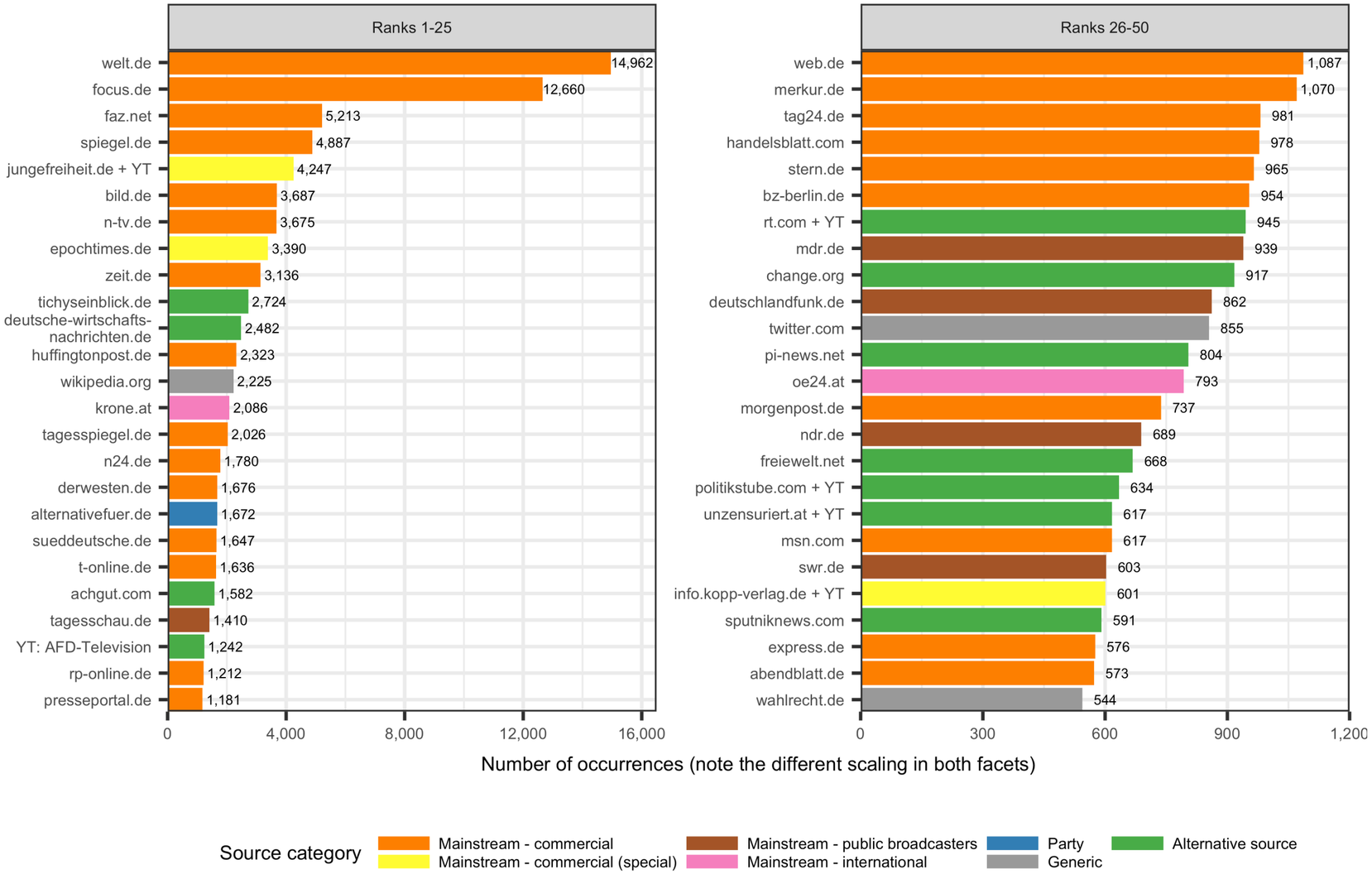
Vor der *APIcalypse*: Kommunikationsspuren auf Social Media

(ALTERNATIVE) MEDIA SOURCES IN AFD-CENTERED FACEBOOK DISCUSSIONS

> The study is based on all posts, comments, and replies on core AfD Facebook pages during the year 2016, as they could be retrieved during the last week of the year. The Facebook Graph API (https://developers.facebook.com/docs/graphapi) was used for the main data collection.

(ALTERNATIVE) MEDIA SOURCES IN AFD-CENTERED FACEBOOK DISCUSSIONS

> the sample consisted of 122 pages, mostly of regional and local sections of the party, of its youth organization Junge Alternative, and of AfD politicians. All posts on these pages (n = 170,033), all comments to the posts (n = 1,455,200), and all replies to the comments (n = 960,077) were retrieved (overall n = 2,585,310).


Ranks 1-25
welt.de 14,962
focus.de 12,660
faz.net 5,213
spiegel.de 4,887
jungefreiheit.de + YT 4,247
bild.de 3,687
n-tv.de 3,675
epochtimes.de 3,390
zeit.de 3,136
tichyseinblick.de 2,724
deutsche-wirtschafts-nachrichten.de 2,482
huffingtonpost.de 2,323
wikipedia.org 2,225
krone.at 2,086
tagesspiegel.de 2,026
n24.de 1,780
derwesten.de 1,676
alternativefuer.de 1,672
sueddeutsche.de 1,647
t-online.de 1,636
achgut.com 1,582
tagesschau.de 1,410
YT: AFD-Television 1,242
rp-online.de 1,212
presseportal.de 1,181
0 4,000 8,000 12,000 16,000
Ranks 26-50
web.de 1,087
merkur.de 1,070
tag24.de 981
handelsblatt.com 978
stern.de 965
bz-berlin.de 954
rt.com + YT 945
mdr.de 939
change.org 917
deutschlandfunk.de 862
twitter.com 855
pi-news.net 804
oe24.at 793
morgenpost.de 737
ndr.de 689
freiewelt.net 668
politikstube.com + YT 634
unzensuriert.at + YT 617
msn.com 617
swr.de 603
info.kopp-verlag.de + YT 601
sputniknews.com 591
express.de 576
abendblatt.de 573
wahlrecht.de 544
0 300 600 900 1,200
Number of occurrences (note the different scaling in both facets)
Source category
Mainstream - commercial
Mainstream - commercial (special)
Mainstream - public broadcasters
Mainstream - international
Party
Generic
Alternative source

(Bachl, 2018)

POST-API-AGE & APICALYPSE

POST-API-AGE & APICALYPSE

- APIs der meisten großen Social-Media-Plattformen mehr oder weniger geschlossen oder für Grundlagenforschung unbezahlbar.
  - Keine (praktikablen) Zugänge zu Facebook, Instagram, Twitter
  - Eingeschränkter Zugang zu TikTok, Reddit, YouTube (aber kaum zu SN-Features)
  - Zugang zu kleineren Plattformen, z.B. BlueSky, Mastodon, Telegram
- Problem: Willkür der Anbieter
- Hoffnung: EU Digital Services Act (DSA)

APIS ARE ALIVE AND WELL

- Trotzdem: APIs bleiben wichtiges Werkzeug für digitale Forschungsmethoden, wenn auch (aktuell) weniger für Social-Media-Forschung
- Datenzugang: Bundestag, MediaWiki Action API (u.a. Wikipedia), Wikimedia REST API, YouTube, Telegram, Tagesschau, The Guardian, DESTATIS, …
- Kommunikation mit Cloud-Diensten (2. Teil der Sitzung)

Umsetzung mit `R` und `{httr2}`

UMSETZUNG MIT `R` UND `{httr2}`

- Beispiel: Aufmerksamkeit für Spitzenkandidat:innen in den letzten drei Wochen — gemessen an den Aufrufen ihrer Wikipedia-Seiten

- Umsetzung mit MediaWiki Action API, Endpoint PageViewInfo
- Nachmachen: 👉 `bsp_wikipedia.R`

https://httr2.r-lib.org/

GENUTZTE PAKETE

```
library(httr2) # Kommunikation mit API über HTTP
library(jsonlite) # JSON-Dateien
library(tidyverse) # Datenmanipulation und Grafik
```

ANFRAGE AN DIE API

```
req = request(base_url = "https://de.wikipedia.org/w/api.php") |>
  req_url_query(!!!list(
    action = "query",
    format = "json",
    prop = "pageviews",
    titles = c("Olaf_Scholz", "Robert_Habeck",
               "Christian_Lindner", "Alice_Weidel",
               "Sahra_Wagenknecht", "Friedrich_Merz"),
    pvipdays = 21),
    .multi = "pipe")
req |>
  req_dry_run()
```

```
GET /w/api.php?
action=query&format=json&prop=pageviews&titles=Olaf_Scholz|Robert_Habeck|Christian_Lindner|Alice_Weidel|Sahra_W
HTTP/1.1
Host: de.wikipedia.org
User-Agent: httr2/1.0.1 r-curl/5.2.1 libcurl/8.7.1
Accept: */*
Accept-Encoding: deflate, gzip
```

ANTWORT DER API

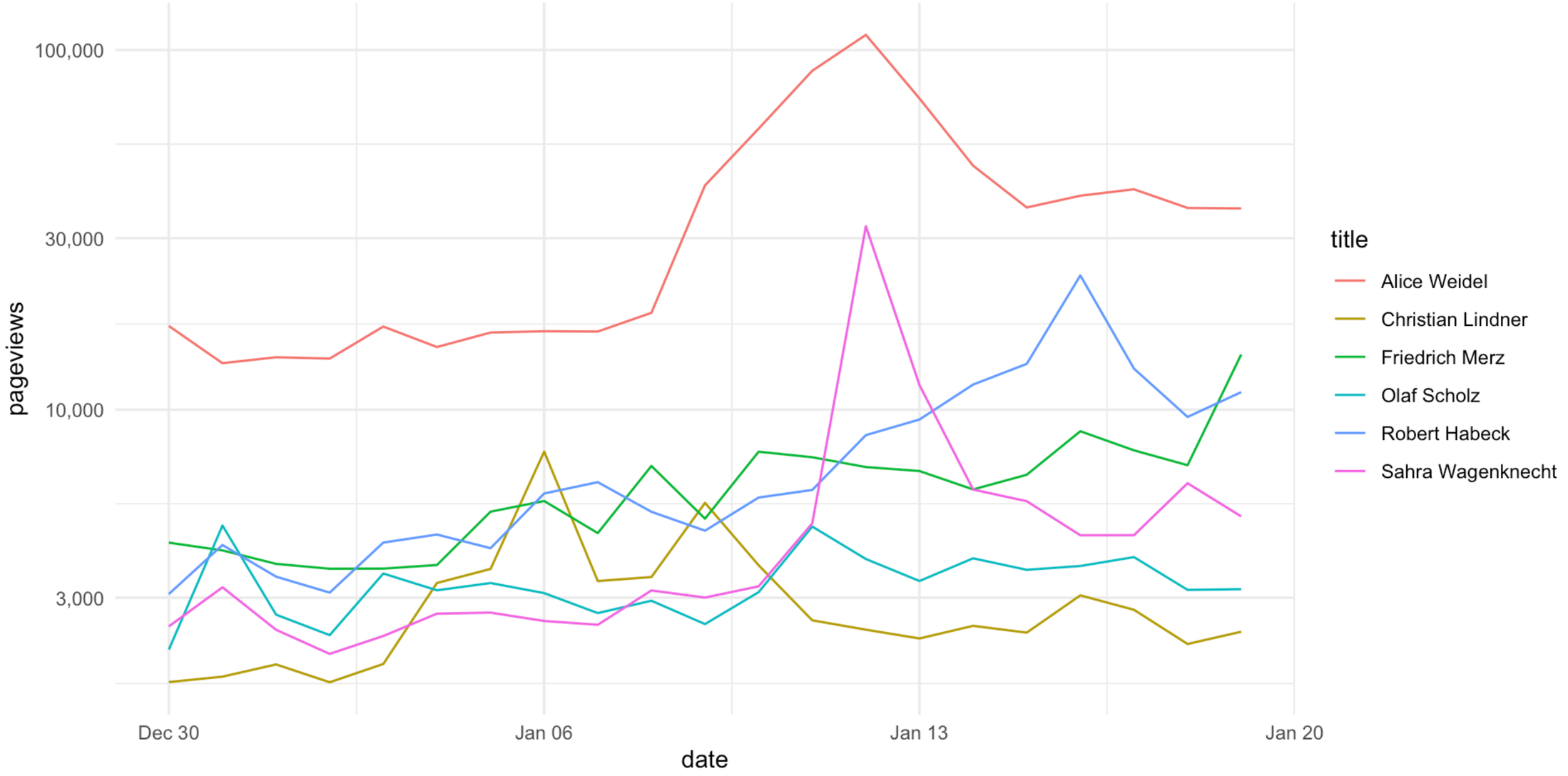
```
resp = req|>
  req_perform()

resp |>
  resp_body_string() |>
  prettify()
```


```
{
    "batchcomplete": "",
    "query": {
        "normalized": [
            {
                "from": "Olaf_Scholz",
                "to": "Olaf Scholz"
            },
            {
                "from": "Robert_Habeck",
                "to": "Robert Habeck"
            },
            {
                "from": "Christian_Lindner",
```

ANTWORT KONVERTIEREN UND PLOTTEN

```
resp |>
  resp_body_json() |>
  _$query |>
  _$pages |>
  map_dfr(as_tibble) |>
  mutate(date = as_date(names(pageviews))) |>
  unnest(pageviews) |>
  ggplot(aes(date, pageviews, color = title)) +
  geom_line() +
  scale_y_log10(labels = scales::label_comma()) +
  theme_minimal()
```

ANTWORT KONVERTIEREN UND PLOTTEN


100,000
30,000
10,000
3,000
pageviews
Dec 30
Jan 06
Jan 13
Jan 20
date
title
Alice Weidel
Christian Lindner
Friedrich Merz
Olaf Scholz
Robert Habeck
Sahra Wagenknecht

UMSETZUNG MIT `R` UND `{httr2}`

- Workflow: API finden, Dokumentation verstehen, Daten abfragen, aufbereiten, analysieren
- Einschränkung der MediaWiki Action API: Nur letzte 60 Tage
- Alternative: Wikimedia REST API 👉 `bsp_wikipedia_rest.R`

Fragen?

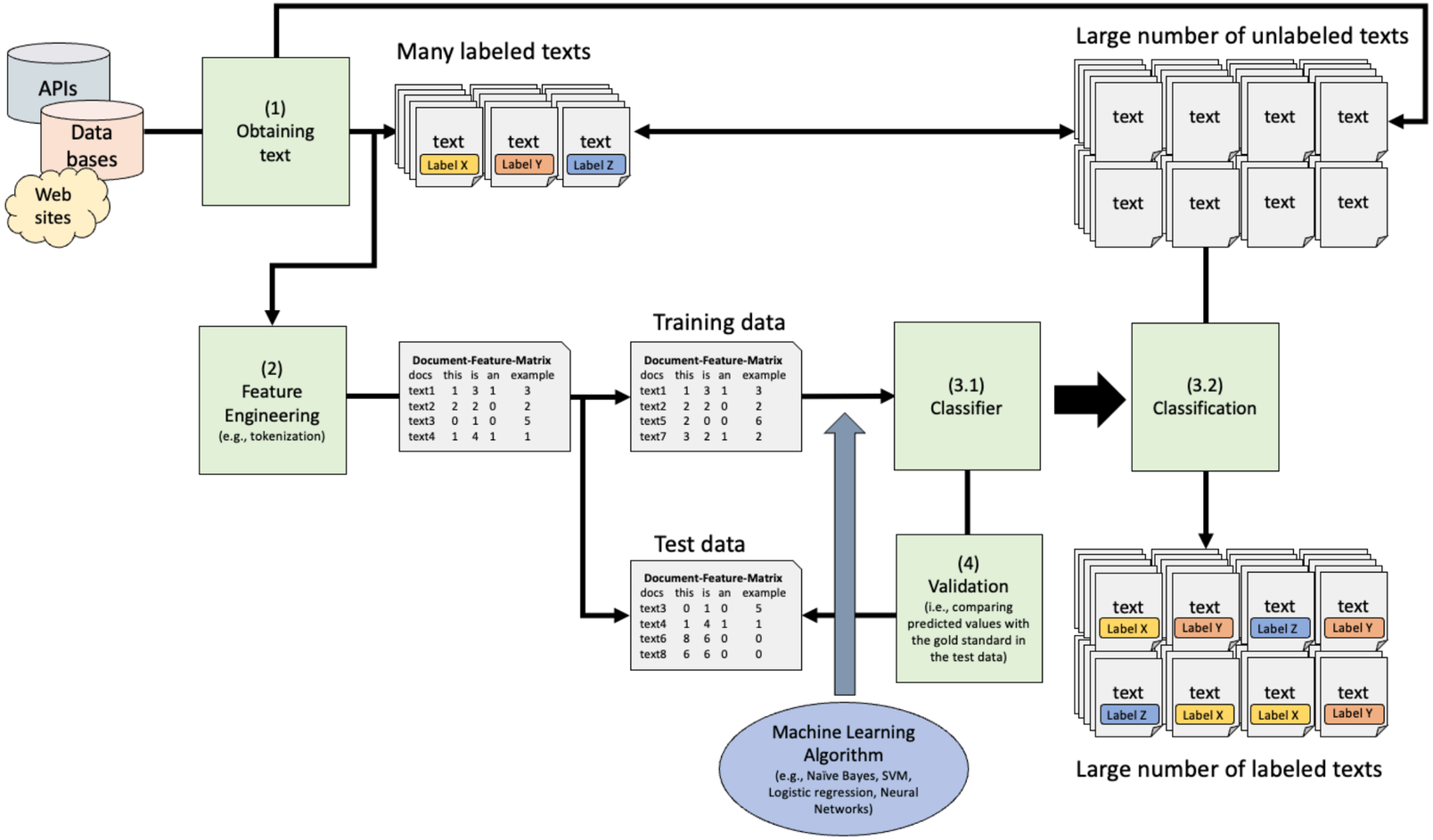
Neuerer Einsatz in PuK: Nutzung von Cloud-Diensten (z.B. KI)

NUTZUNG VON CLOUD-DIENSTEN

- Viele Cloud-Dienste lassen sich über APIs verwenden
- Beispiele aus dem Bereich KI: Huggingface Inference API, OpenAI API, Perspective API
- Workflow ist ähnlich: Anfrage senden, Antwort erhalten
- Unterschiede: Erfordert fast immer Authentifizierung, häufig kostenpflichtig

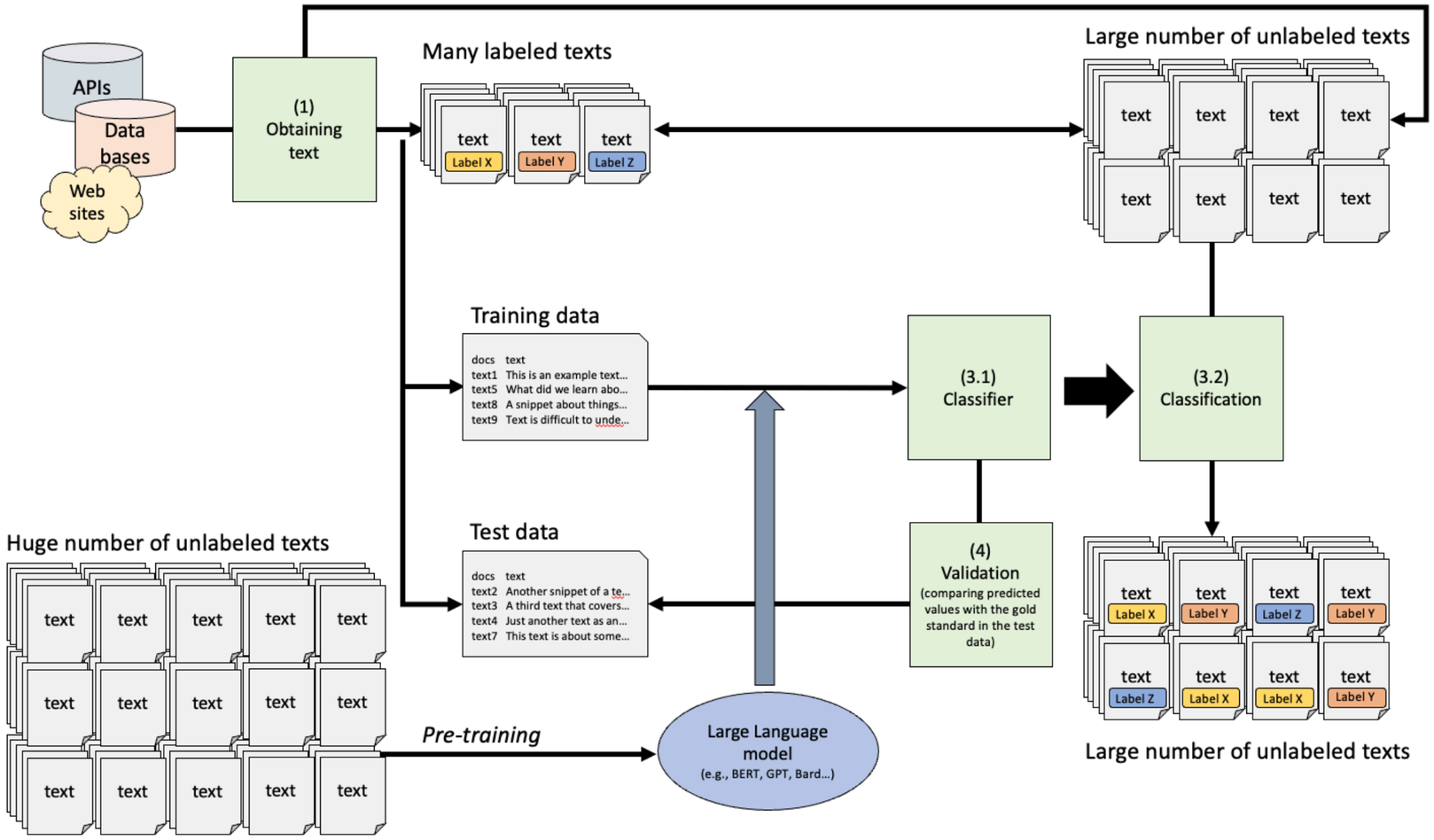
Zero-shot classification: Kurze Einführung

BAG-OF-WORDS MACHINE LEARNING



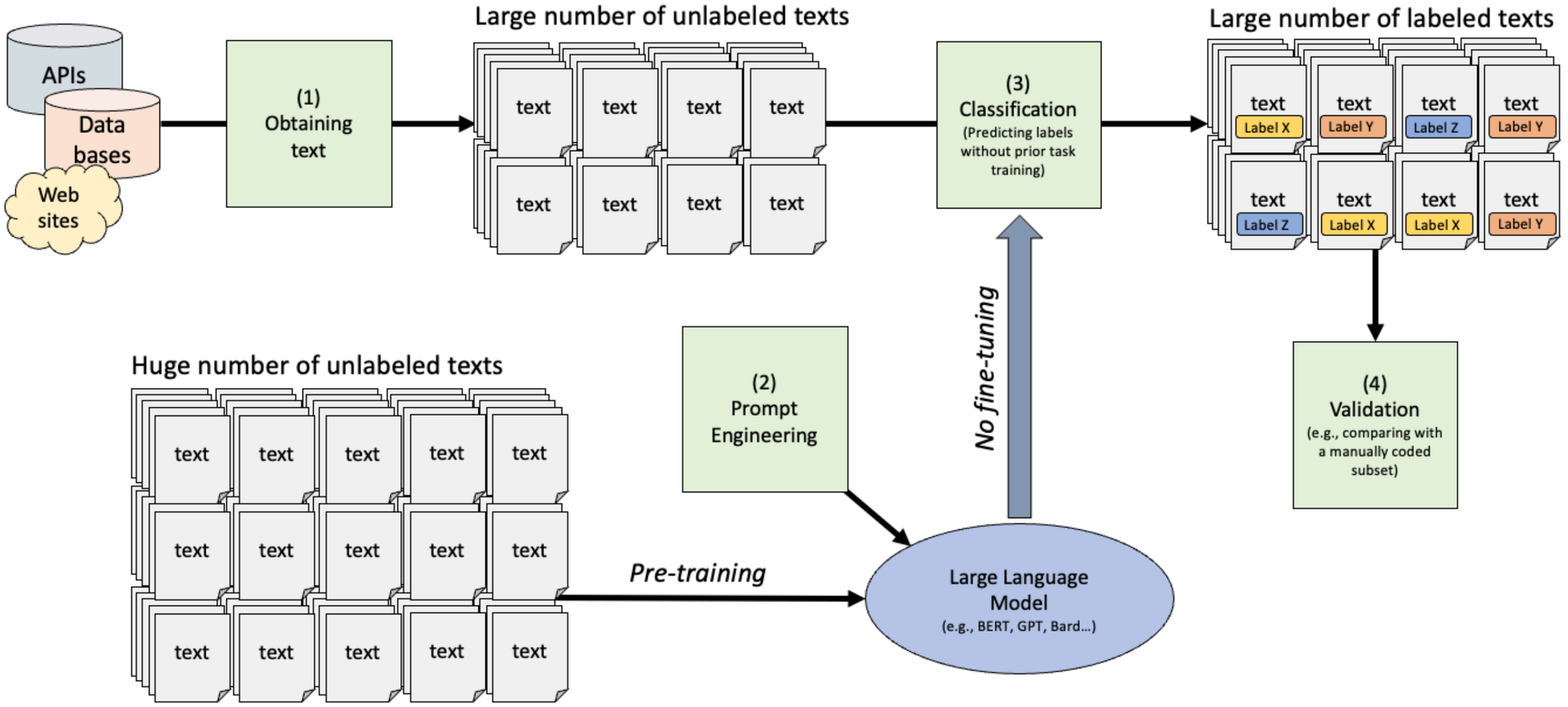
Abbildungen von Philipp K. Masur; Zur Entwicklung von Computational Text Analysis: Bachl & Scharkow (2024)

TRANSFER LEARNING



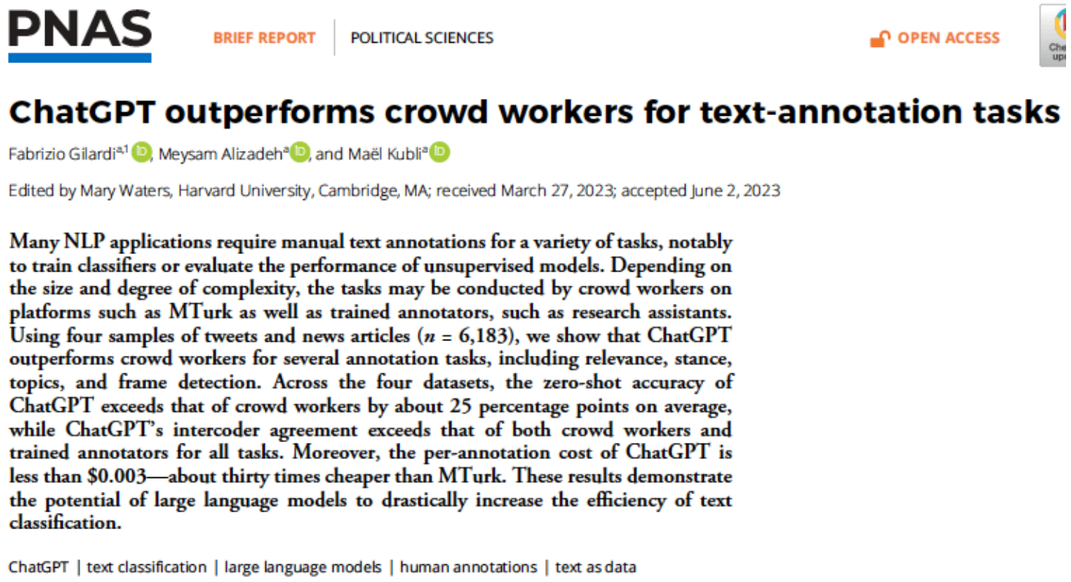
Abbildungen von Philipp K. Masur; Zur Entwicklung von Computational Text Analysis: Bachl & Scharkow (2024)

ZERO-SHOT CLASSIFICATION



Abbildungen von Philipp K. Masur; Zur Entwicklung von Computational Text Analysis: Bachl & Scharkow (2024)

ZERO-SHOT CLASSIFICATION: HYPE (?)



PNAS BRIEF REPORT | POLITICAL SCIENCES OPEN ACCESS

ChatGPT outperforms crowd workers for text-annotation tasks

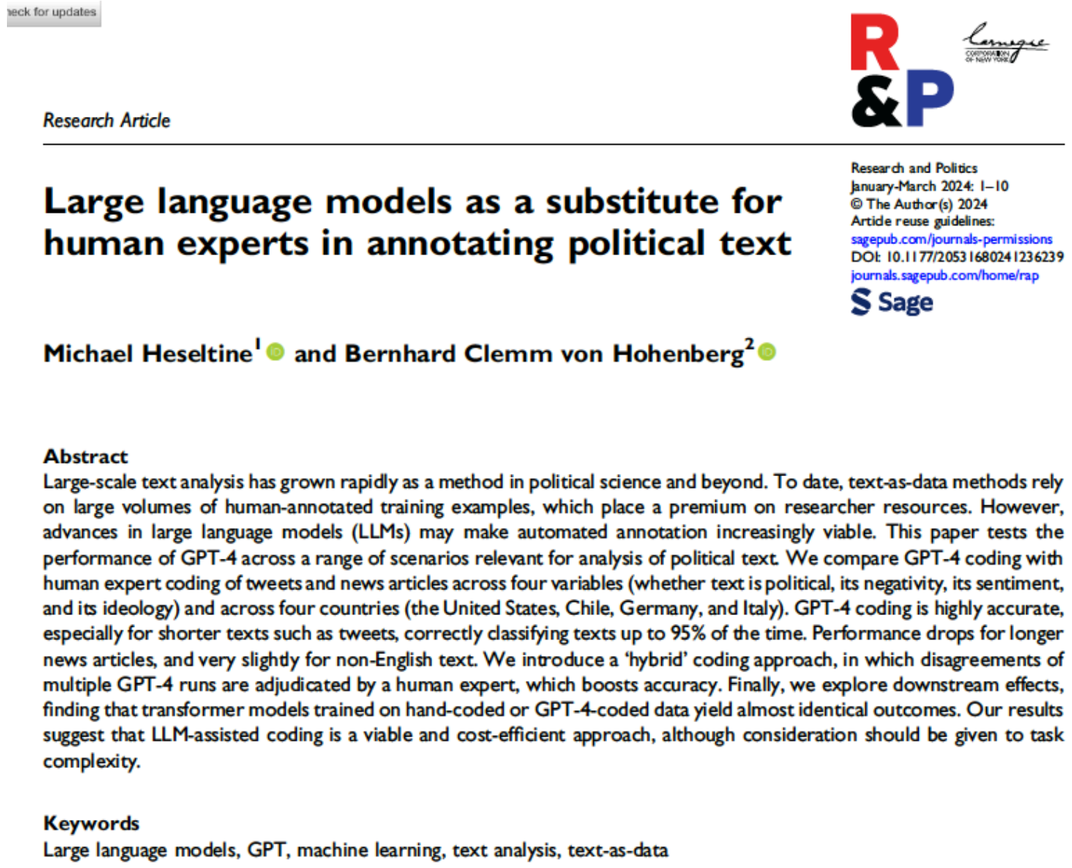
Fabrizio Gilardi[a,1], Meysam Alizadeh[a], and Maël Kubli[a]

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and news articles (n = 6,183), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

ChatGPT | text classification | large language models | human annotations | text as data

(Gilardi et al., 2023)



eck for updates

Research Article

Large language models as a substitute for human experts in annotating political text

Research and Politics
January-March 2024: 1–10

DOI: 10.1177/20531680241236239
journals.sagepub.com/home/rap
Sage

Michael Heseltine[1] and Bernhard Clemm von Hohenberg[2]

Abstract

Large-scale text analysis has grown rapidly as a method in political science and beyond. To date, text-as-data methods rely on large volumes of human-annotated training examples, which place a premium on researcher resources. However, advances in large language models (LLMs) may make automated annotation increasingly viable. This paper tests the performance of GPT-4 across a range of scenarios relevant for analysis of political text. We compare GPT-4 coding with human expert coding of tweets and news articles across four variables (whether text is political, its negativity, its sentiment, and its ideology) and across four countries (the United States, Chile, Germany, and Italy). GPT-4 coding is highly accurate, especially for shorter texts such as tweets, correctly classifying texts up to 95% of the time. Performance drops for longer news articles, and very slightly for non-English text. We introduce a 'hybrid' coding approach, in which disagreements of multiple GPT-4 runs are adjudicated by a human expert, which boosts accuracy. Finally, we explore downstream effects, finding that transformer models trained on hand-coded or GPT-4-coded data yield almost identical outcomes. Our results suggest that LLM-assisted coding is a viable and cost-efficient approach, although consideration should be given to task complexity.

Keywords

Large language models, GPT, machine learning, text analysis, text-as-data

(Heseltine & Clemm von Hohenberg, 2024)

Zero-shot classification: Umsetzung mit der OpenAI-API

ZERO-SHOT CLASSIFICATION: OPENAI-API

- Beispiel: Klassifikation von Inzivilität in Social-Media-Kommentaren
- Klassifikation mit OpenAI GPT-4o und `httr2`
- Nachmachen: 👉 `bsp_zero_shot_openai.R` (mit OpenAI-Account)

Genutzte Pakete

```
library(httr2) # Kommunikation mit API über HTTP
library(jsonlite) # JSON-Dateien
library(tidyverse) # Datenmanipulation und Grafik
```

ERKENNEN VON INZIVILITÄT IN SOCIAL-MEDIA-KOMMENTAREN (STOLL ET AL., 2023)

Kommentar mit mindestens einer der folgenden Eigenschaften gilt als inzivil:

- Vulgäre, unangemessene Sprache, Fluchen
- Beleidigung, Profanität
- Entmenschlichung
- Sarkasmus, Spott, Zynismus
- Negative Stereotype
- Diskriminierung
- Androhung von Gewalt
- Verweigerung von Rechten
- Vorwurf der Lüge
- Erniedrigung, fehlender Respekt, Abwertung

UNTERSUCHUNGSMATERIAL

Wir brauchen ein paar Kommentare zum Testen:

- Einen klar inzivilen Kommentar
- Einen klar nicht inzivilen Kommentar
- Zwei mehrdeutige Kommentare:
  - Einen *nicht* inzivilen Kommentar, der fälschlicherweise als inzivil klassifiziert wird
  - Einen inzivilen Kommentar, der fälschlicherweise als *nicht* inzivil klassifiziert wird

URL FÜR ANFRAGE

```
req = request(base_url = "https://api.openai.com/v1/chat/completions")
req |>
  req_dry_run()
```

```
GET /v1/chat/completions HTTP/1.1
Host: api.openai.com
User-Agent: httr2/1.0.1 r-curl/5.2.1 libcurl/8.7.1
Accept: */*
Accept-Encoding: deflate, gzip
```

KEY ZUR ANMELDUNG BEI OPENAI

> ⚠ Schlüssel und Token niemals öffentlich teilen!

```
key = readLines("openai_key.txt")

req |>
  req_auth_bearer_token(key) |>
  req_dry_run()
```

```
GET /v1/chat/completions HTTP/1.1
Host: api.openai.com
User-Agent: httr2/1.0.1 r-curl/5.2.1 libcurl/8.7.1
Accept: */*
Accept-Encoding: deflate, gzip
Authorization: <REDACTED>
```

PROMPT (1)

Codieranweisung: Was soll KI-Assistent tun?

```
instr = paste(readLines("codieranweisung.txt"), collapse = "\n")
cat(instr)
```

```
Your task is to assist in the classification of social media commments.
You will receive a comment that was posted to a social media plattform.
Your task is to classify the commment as either incivil or civil.

Incivility is defined as a statement that contains any of the following features:
Vulgarity, Inappropriate Language, Swearing, Insults, Name Calling, Profanity,
Dehumanization, Sarcasm, Mockery, Cynicism, Negative Stereotypes, Discrimination,
Threats of Violence, Denial of Rights, Accusations of Lying, Degradation,
Disrespect, Devaluation.

Fill out the provided JSON response form.
First, provide your reasoning to decide whether the search query is incivil or civil.
Then give your classification.
```

(Törnberg, 2024)

PROMPT (2)

Kategoriensystem: Wie soll die Antwort aussehen?

```
response_format = list(
  type = "json_schema",
  json_schema = list(
    name = "social_media_incivility",
    schema = list(
      type = "object",
      properties = list(
        reasoning = list(
          description = "Short text to explain your re
          type = "string"
        ),
        classification = list(
          description = "Classification into incivil c
          type = "string",
          enum = c("incivil", "civil")
        )
      ),
```

```
{
    "type": [
        "json_schema"
    ],
    "json_schema": {
        "name": [
            "social_media_incivility"
        ],
        "schema": {
            "type": [
                "object"
            ],
            "properties": {
                "reasoning": {
```

PROMPT (3)

Codiereinheiten: Was soll klassifiziert werden?

```
cod = readLines("comments.txt")
cat(cod, sep = "\n")
```

```
Die Kartoffelbauern sollen daheim bleiben!
Dein Hund sieht super süß aus.
Dir sollte man bald mal einen Besuch abstatten.
Du geile Sau!
```

ANFRAGE

```
req |>
  req_auth_bearer_token(key) |>
  req_body_json(list(
    model = "gpt-4o",
    messages = list(
      list(role = "system", content = instr),
      list(role = "user", content = cod[1])
    ),
    response_format = response_format,
    temperature = 0,
    max_completion_tokens = 500
  )) |>
  req_dry_run()
```

```
POST /v1/chat/completions HTTP/1.1
Host: api.openai.com
User-Agent: httr2/1.0.1 r-curl/5.2.1 libcurl/8.7.1
Accept: */*
Accept-Encoding: deflate, gzip
Authorization: <REDACTED>
Content-Type: application/json
Content-Length: 1302

{"model":"gpt-4o","messages":[{"role":"system","content":"Your task is to assist in the classification of
social media commments.\nYou will receive a comment that was posted to a social media plattform.\nYour task is
to classify the commment as either incivil or civil.\n\nIncivility is defined as a statement that contains any
```

ANTWORT

```
resp = req |>
  req_auth_bearer_token(key) |>
  req_body_json(list(
    model = "gpt-4o",
    messages = list(
      list(role = "system", content = instr),
      list(role = "user", content = cod[1])
    ),
    response_format = response_format,
    temperature = 0,
    max_completion_tokens = 500
  )) |>
  req_perform()

resp |>
  resp_body_string() |>
  prettify()
```


```
{
    "id": "chatcmpl-ArhOLLklZ9mvUJ8JPvghL5OyrB31g",
    "object": "chat.completion",
    "created": 1737361545,
    "model": "gpt-4o-2024-08-06",
    "choices": [
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "content": "{\"reasoning\":\"The
comment 'Die Kartoffelbauern sollen daheim bleiben!'
translates to 'The potato farmers should stay at
home!' in English. This statement can be interpreted
```

ALLE KOMMENTARE

▶ Code

Kommentar	reasoning	classification
Die Kartoffelbauern sollen daheim bleiben!	The comment suggests that a group of people, in this case, ‘potato farmers’, should stay at home. This could be interpreted as a form of discrimination or devaluation, as it implies that this group is not welcome or should not participate in certain activities. Such statements can be seen as disrespectful or degrading towards the group mentioned.	incivil

Kommentar	reasoning	classification
Dein Hund sieht super süß aus.	The comment is a compliment about someone’s dog being cute. It does not contain any features of incivility such as vulgarity, insults, or disrespect.	civil
Dir sollte man bald mal einen Besuch abstatten.	The comment suggests a visit to the person, which can be interpreted as a threat of violence or intimidation, especially if the context implies a negative or aggressive intent. This makes the comment incivil.	incivil
Du geile Sau!	The phrase ‘Du geile Sau!’ can be considered vulgar and inappropriate language. It uses a term that can be	incivil

ZERO-SHOT CLASSIFICATION

- Generative LLMs und Natural language inference (NLI) (Laurer et al., 2023)
- Rasant entwickelndes Forschungsfeld:
  - Verbesserungen durch weiterentwickelte Modelle
  - Evaluation der Performance: Was geht, was (noch) nicht?
  - Prompt engineering: Welche Codieranweisungen sind besser?
  - Forschungethische Fragen: Biases, Reproduzierbarkeit, Abgängigkeit von proprietären Modellen
- Selbst ausprobieren - es ist nicht so schwer, wie es am Anfang aussieht.

Fragen?

Vielen Dank

Marko Bachl

marko.bachl@fu-berlin.de

LITERATUR

Bachl, M. (2018). (Alternative) media sources in AfD-centered Facebook discussions. *Studies in Communication and Media*, *7*(2), 128–142. https://doi.org/ghhx99

Bachl, M., & Scharkow, M. (2024). *Computational text analysis*. OSF. https://doi.org/10.31219/osf.io/3yhu8

Bruns, A. (2019). After the „APIcalypse“: social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/gf8r25

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/gfs6ng

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Haim, M. (2023). *Computational Communication Science: Eine Einführung*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-40171-9

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, *11*(1), 20531680241236239. https://doi.org/gtkhqr

Laurer, M., Atteveldt, W. van, Casas, A., & Welbers, K. (2023). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 1–17. https://doi.org/10.1017/pan.2023.20

Stoll, A., Wilms, L., & Ziegele, M. (2023). Developing an incivility dictionary for German online discussions – a semi-automated approach combining human and artificial knowledge. *Communication Methods and Measures*, *17*(2), 131–149. https://doi.org/gsnfdn

Törnberg, P. (2024). *Best practices for text annotation with large language models*. arXiv. https://doi.org/gtn9qf