Datensatzverarbeitung
(engl. dataset pipeline)

Datensatzakquisition
(engl. dataset acquisition)

Datensatzsichtung
(engl. dataset inspection)

Datensatzaufbereitung
(engl. dataset preparation)

Datensatzrechenrche
(engl. dataset research)

Datensatzsammlung
(engl. dataset collection)

Datensatzprüfung
(engl. dataset check)

Datensatzauswahl
(engl. dataset selection)

Explorative Datenanalyse
(engl. exploratory data analysis)

In der EDA werden Textdaten untersucht, um Muster, Qualitätsprobleme und Strukturen zu erkennen.

(perplexity.ai)

Datensatzbereinigung
(engl. dataset cleaning)

Datensatzvalidierung
(engl. dataset validation)

Suchstrategie
(engl. search strategy)

Download
(engl. data download)

Speicherung
(engl. data storage)

Datenstrukturanalyse
(engl. data structure analysis)

Analyse der strukturierten Daten
(engl. analysis of structured data)

Analyse der unstrukturierten Daten
(engl. analysis of unstructured data)

Fehlwertbehandlung
(engl. missing value handling)

Duplikatentfernung
(engl. duplicate removal)

Duplikate verzerren NLP-Modelle, daher müssen sie entfernt werden.

Ortsdaten
(engl. location data)

Zeitdaten
(engl. time data)

Bewertungen
(engl. ratings)

Fehlwerterkennung
(engl. missing value detection)

Duplikaterkennung
(engl. duplicate detection)

Textlängenanalyse
(engl. text length analysis)

Rohzeichenfolgen
(engl. raw strings)

Suchbegriffe

- „Beschwerden"
- „Kommentare"
- „NLP"
- „Kundenbeschwerden"
- „Online-Kommentare"
- „Produktbewertungen"

Anforderungen

- organischer Datensatzes

- **lokale Spricherung**

- **Bundesstaaten**

- **Zeitraumbestimmung**
- **Datumsanalyse**

- **NaNs (Not a Number)**
- **NaTs (Not a Text)**

- **Imputation**

„Strategies for handling missing data in text include imputation where missing values are estimated based on existing data and removal where incomplete observations are excluded from the analysis. These strategies help in ensuring the completeness and integrity of the text data [12]."

(Upadhye, 2020, p. 207)

- **Listenweiser Fallausschluss**

```
# Zeilen ohne Text entfernen (0.4% sind vernachlässigbar)
df_clean = df.dropna(subset=['Text'])

# Oder Mindestlänge erzwingen
df_clean = df[df['Text'].str.len() > 30]
```