What is there

- Scraping
    - Python script
- Topic extraction
    - STM model (Simon & Patrick)
- Sentiment Analysis
    - Training data
        - Manually annotated tweets (~ 200)
    - Feature extraction
        - DFM of n-grams (clean of umlauts, symbols, stopwords; stemmed)
        - Hashtags, emojis
    - Models
        - Dictionary-based
            - Unigram-based model with third-party dictionaries
        - ML-based
            - Full `mlr3` pipeline that performs preprocessing, training, tuning and evaluation in a nested resampling framework
            - Models and metrics easily customized

What remains open

- Scraping
  - R script - already implemented in parts but suspended for the time being
- Topic extraction
  - Alternative topic models
  - Incorporation of hashtags
- Sentiment Analysis
  - Training data
    - Probably a lot more required – worth the effort?
    - Third-party data for pre-training (e.g., GermEval2017)?
  - Feature extraction
    - Lemmatization – no good `R` option for German found yet; even worth it?
    - POS tagging
    - Incorporation of emojis
    - Comparison of different weighting schemes
  - Models
    - Dictionary-based
      - n-gram models
      - Custom dictionaries (e.g., by expansion)
    - ML-based
      - Definition of evaluation metric (inner + outer)
      - Model selection