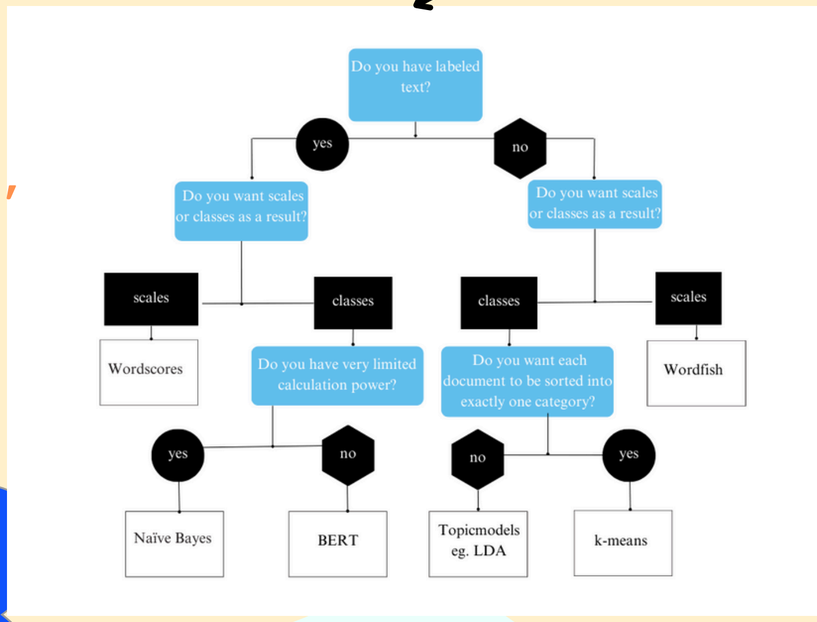
MACHINE LEARNING IN QTA

MODEL CHOICE

research methods used for your goal

data type?

What is my use case, data format, resources?



PREPROCESSING

1. turn data into dataframe
2. turn dataframe into corpus
3. turn into dfm, dtm, matrix dependent on model choice

What format do I need? What features do I keep?

= input text in right format

remove urls, emojis etc.

manipulate text features

- remove punctuation, numbers, symbols, stopwords
- stem or lemmatize

HYPERPARAMETERS

= parameters you have to determine

determine through

- literature
- evaluation methods

How can I optimise my model?

- iterations how often the algorithm should run
- choosing the number of topics/clusters "k"
- examples
  - algorithm e.g. sorting algorithm

EVALUATION & VISUALISATION

- reliability
  - -delivers the same result in repeated measurements
  - set.seed()
  - reliability of hand-coded text
- validity
  - -whether a measure measures what it should
  - face-, construct validity
  - convergent-, statistical-, semantic-, predictive validation
- `ggplot2` framework