# NLP coding challenge

# Explorative Datenanalyse

Steps

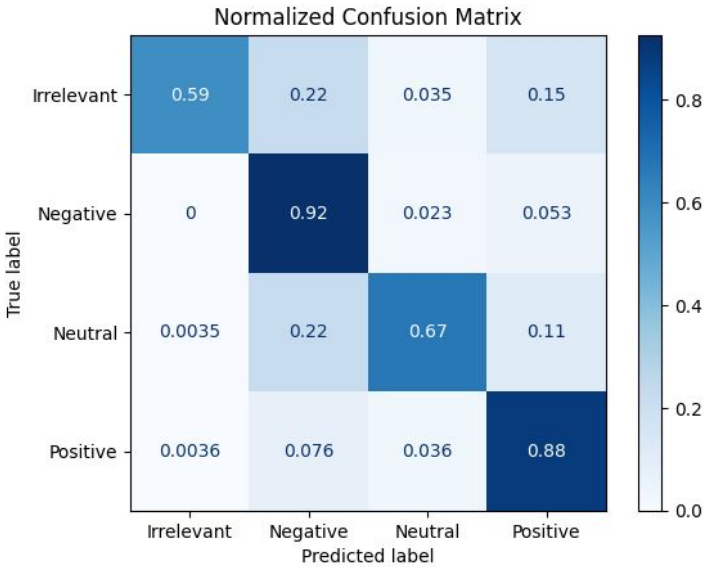
- basic checks
- duplicates, NaNs
- label distribution
- number of products -> same across train val
- sentiment by product
- length of comments (for max length DistilBert)

Evaluation -> F1 macro since not equally distributed between all classes

# Modellbuilding

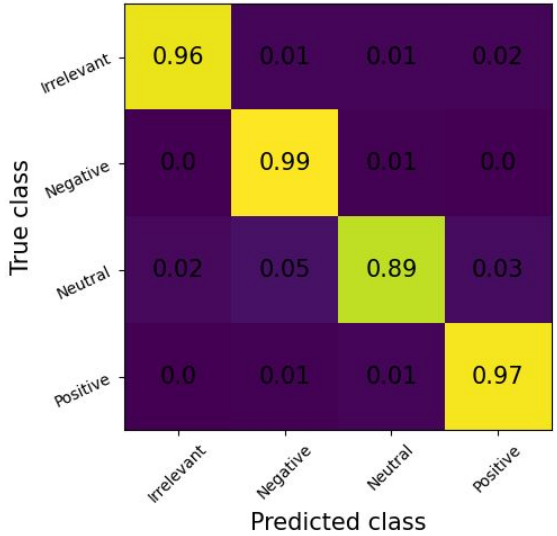
## Naive Bayes (Baseline)

F1 = 0.78




## Distil Bert (bit older "SOTA model")

F1 = 0.95



# ML Engineering

REST-API:

with api.py very simple with request

can be tested with `python src/deploy/request_test.py`

Dockerfile

runs the api.py in a container

Show live