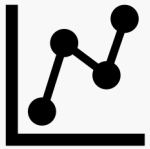
News classification

An end-to-end ML example.

Maximilian Engelhardt
maximilianengelhardt@mail.de

Goal and focus of this project

Focus on best practices to ensure robustness and soundness

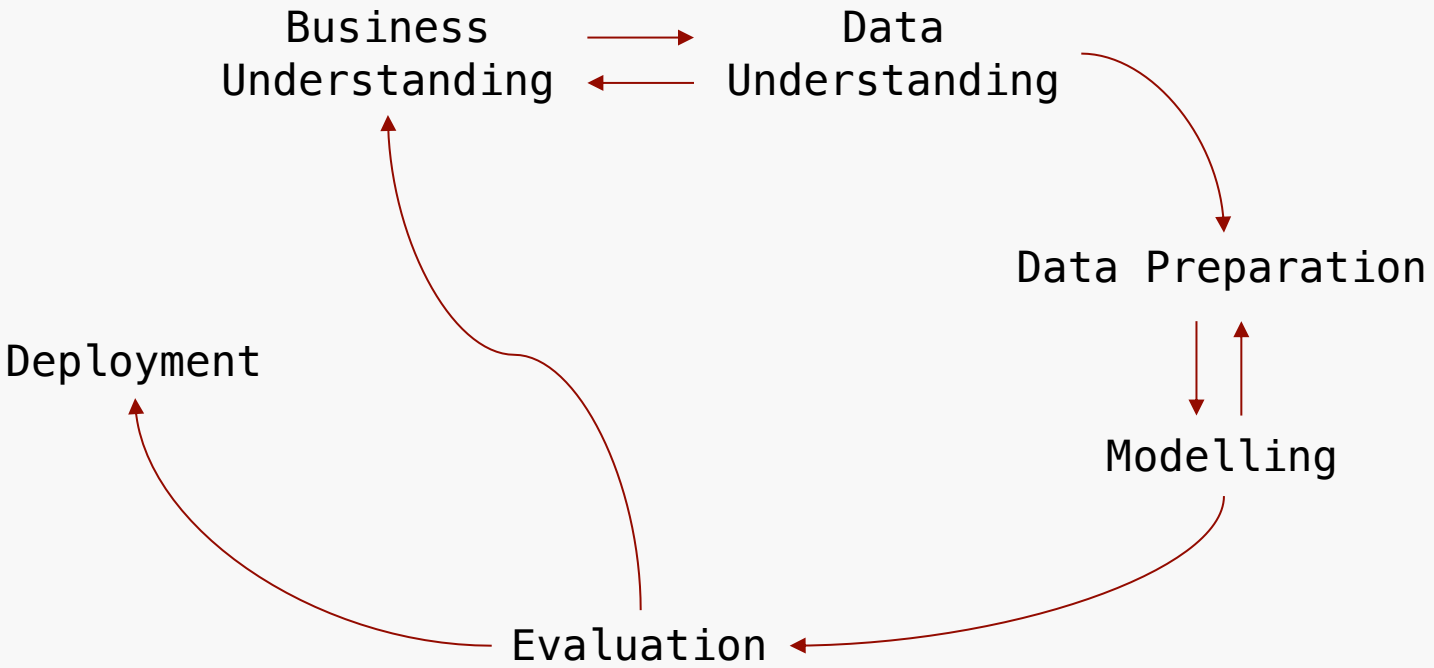


Build and serve a news title classifier.

curl –X GET localhost:8000/predict?title=“Trump refuses to leave White House”

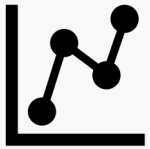
Overall end-to-end approach

CRISP-DM as the general iterative process model



Data Understanding

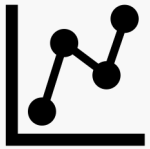
Title is the most prominent feature for not perfectly balanced classes



ID	TITLE	URL	PUBLISHER	CATEGORY	STORY	HOSTNAME	TIMESTAMP
...	US open: Stocks fall after Fed official hints at accelerated tapering	...	IFA Magazine	b	...	...	1394470371550
...	Hunger Games trumps Hobbit at MTV Movie Awards	...	Stuff.co.nz	e	...	...	1397459812062

Data Preparation

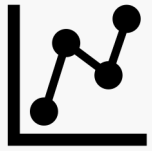
Text is converted to a numerical representation using BOW and TFIDF



"Hunger Games trumps Hobbit at MTV Movie Awards" -> $\boldsymbol{x} = \langle x_1, \ldots, x_m \rangle$

Modeling

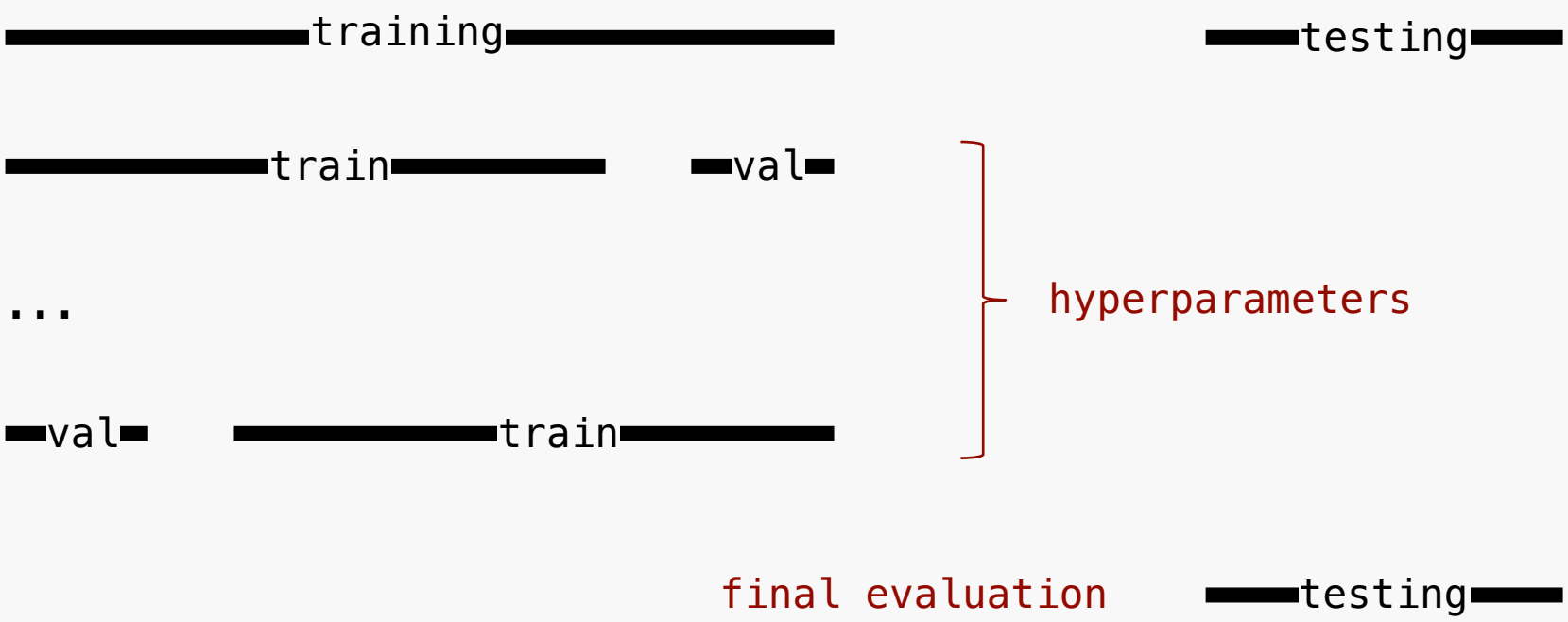
Multinomial Naïve Bayes is chosen as a fast and reliable classifier



$$p(y|x_1, \ldots, x_n) \propto \prod p(x_i|y)\, p(y)$$

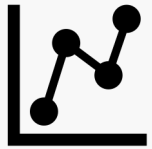
Modeling

Overall workflow for modeling and evaluation



Modeling

The hyperparameter search file summarizes the modeling approach



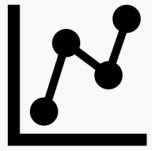
```yaml
random_seed: 42 # for reproducibility
test_size: 0.5 # size of held-out test set

parameters: # hyperparameters for tuning
        vectorizer__max_features: # number of features for bag of words
                - 20000
                - 5000
        tfidf__use_idf: # whether to use tfidf
                - True
        naivebayes__alpha: # naive bayes smoothing parameter
                - 1.0

search: # parameters for the grid search
        n_jobs: -1 # parallel processes (-1 means maximum available)
        k_splits: 3 # number of splits
        metric: "balanced_accuracy" # metric for selecting best parameters
```

Evaluation

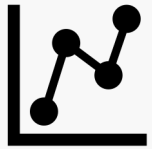
Class specific precision and recall metrics



	precision	recall	f1-score
business	0.895	0.908	0.902
science and technology	0.949	0.968	0.959
entertainment	0.959	0.859	0.906
health	0.898	0.898	0.898
weighted average	0.922	0.922	0.922

Deployment

The classifier is deployed using fastAPI and Docker

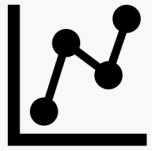


```
$ docker run -p 8000:8000 --name ing-service ing
```

Visit http://localhost:8000/docs

Out of scope

State-of-the-art NLP and full scale MLOps



State-of-the-art NLP e.g. using Transfer Learning and Transformers

Full model management and tracking e.g. using mlflow Model API