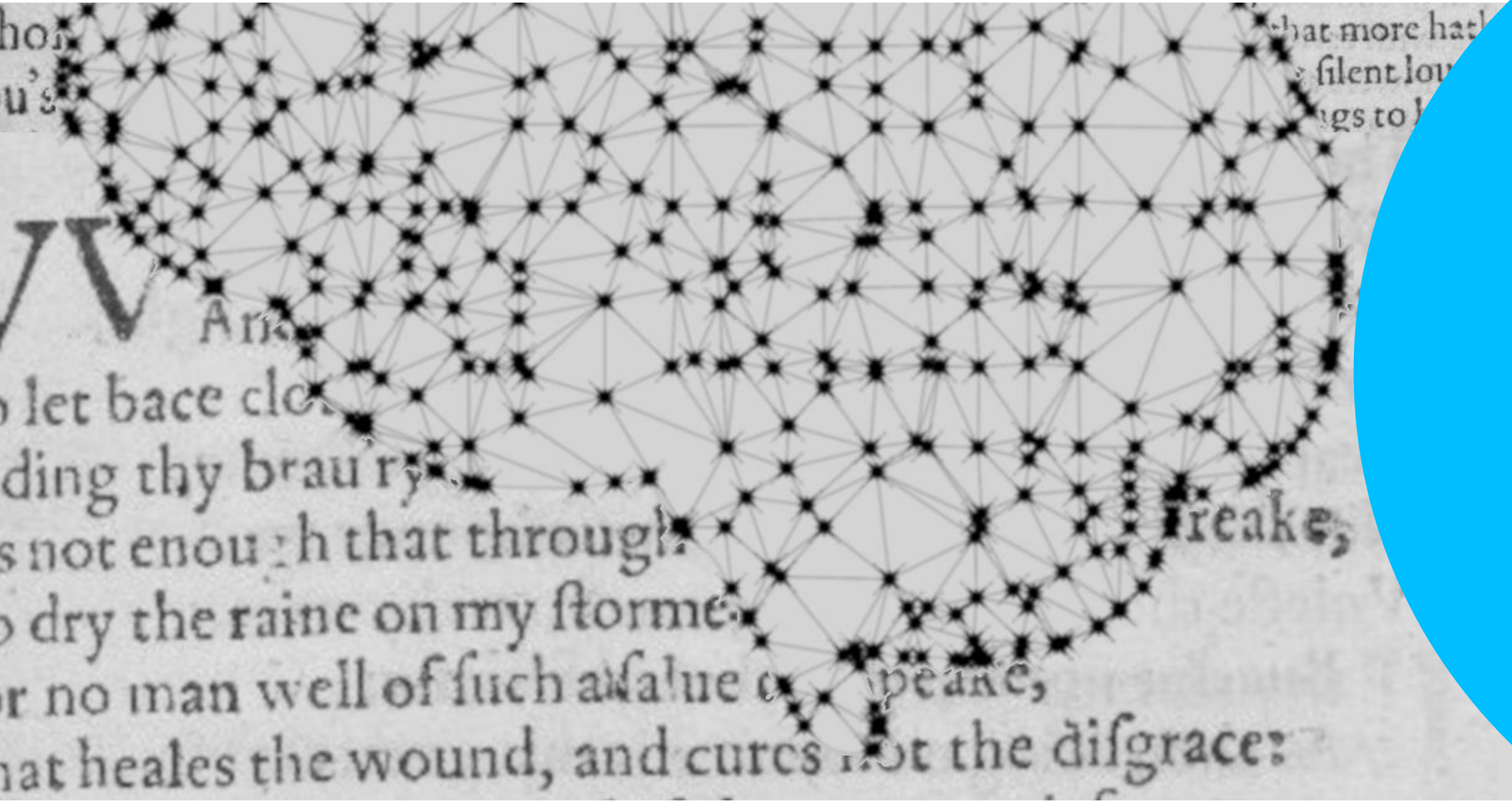
University of Stuttgart
Institut für Maschinelle Sprachverarbeitung


that more hath
silent lou
ngs to
An
let bace clo
ding thy brau ry
s not enough that through
dry the raine on my ſtorme
r no man well of ſuch a ſalue can ſpeake,
hat heales the wound, and cures not the diſgrace:
reake,

Author Classification in Poetry

Katrin Schmidt & Carlotta Quensel

Team Lab NLP | SS2021 | 14-06-2021

Author classification

Goal

- Stylometrics: what age, education, gender, social class has the author
  - Find specific values for one author
- Author prediction: classification task with known classes
  - Which of 5 known authors has written this text
- Our goal: find a method which works well with poetry
  - Style choices might reflect the medium more than the author
  - Special features (number of verses, meter, anaphora, …)
  - Choice of data is limited

Author classification

Data

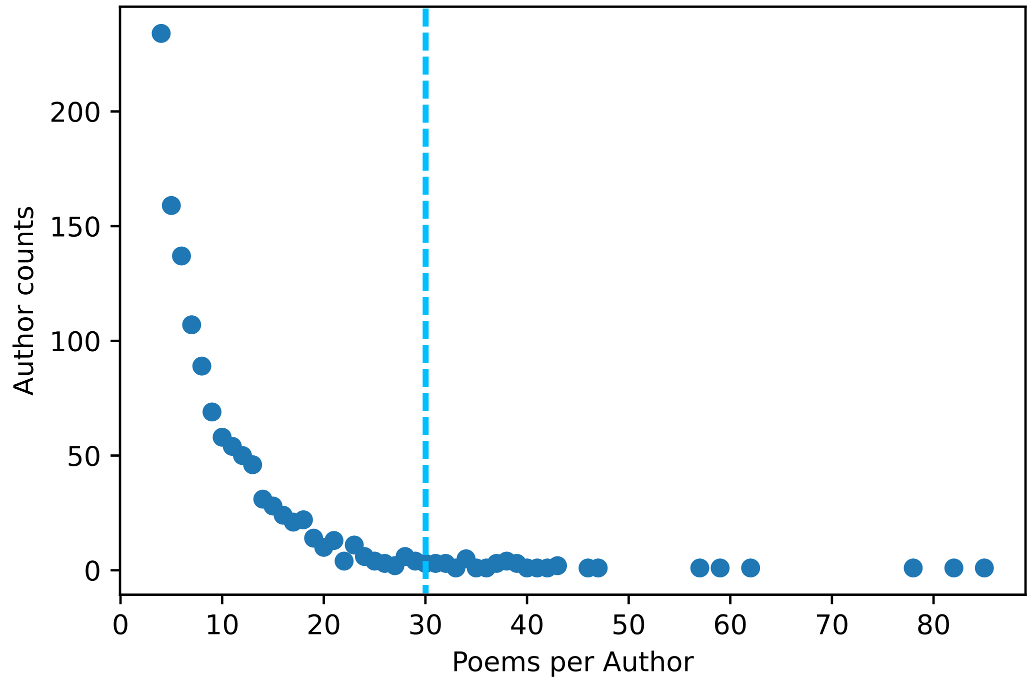


***Fig 1** – Poem distribution per author*

Poetry Foundation (founded in 2003)

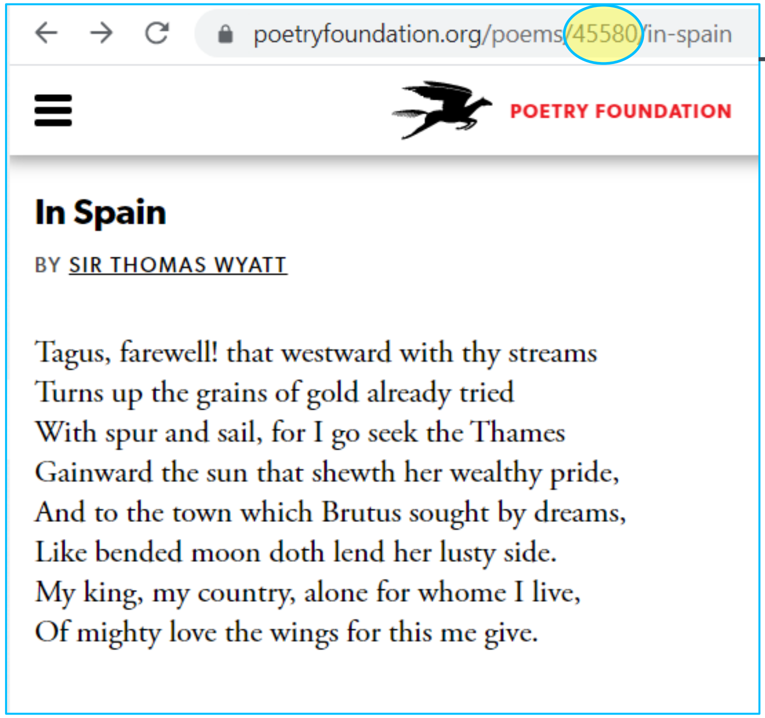
- 3 309 authors
- 15 567 poems

We decided on the 30 most prolific authors

- More datapoints per class
- Lower computational effort
- 1 569 poems

Author classification

Data

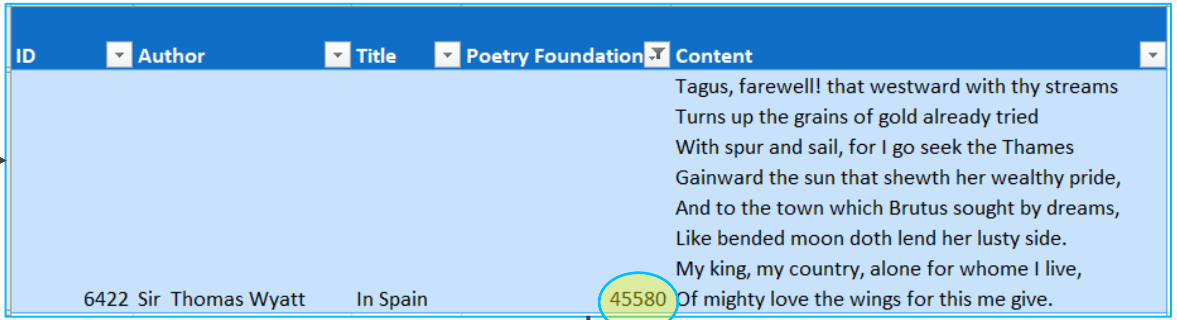


In Spain

BY SIR THOMAS WYATT

Tagus, farewell! that westward with thy streams
Turns up the grains of gold already tried
With spur and sail, for I go seek the Thames
Gainward the sun that shewth her wealthy pride,
And to the town which Brutus sought by dreams,
Like bended moon doth lend her lusty side.
My king, my country, alone for whome I live,
Of mighty love the wings for this me give.

CSV from kaggle



ID	Author	Title	Poetry Foundation	Content
6422	Sir Thomas Wyatt	In Spain	45580	Tagus, farewell! that westward with thy streams Turns up the grains of gold already tried With spur and sail, for I go seek the Thames Gainward the sun that shewth her wealthy pride, And to the town which Brutus sought by dreams, Like bended moon doth lend her lusty side. My king, my country, alone for whome I live, Of mighty love the wings for this me give.

Tokenization with NLTK

```
{'Sir  Thomas Wyatt' : [[‘tagus', ',', 'farewell', '!', 'that',
'westward', ',', 'with', 'thy', 'streams', ',', ‘turns', 'up',
'the', 'grains', 'of', 'gold', 'already', 'tried', ',', ‘with',
'spur', 'and', 'sail', 'for', ‘i', 'go', 'seek', 'the',
‘thames', ',', ‘gainward', 'the', 'sun', 'that', 'show', "'",
…]}
```

Bag-of-words vectors

```
((0,0,0,1,0,0,1,1,0,0,0,1,0,0,1, …), 'Sir Thomas Wyatt')
```

Approach

Baseline method

- Maximum Entropy Classifier

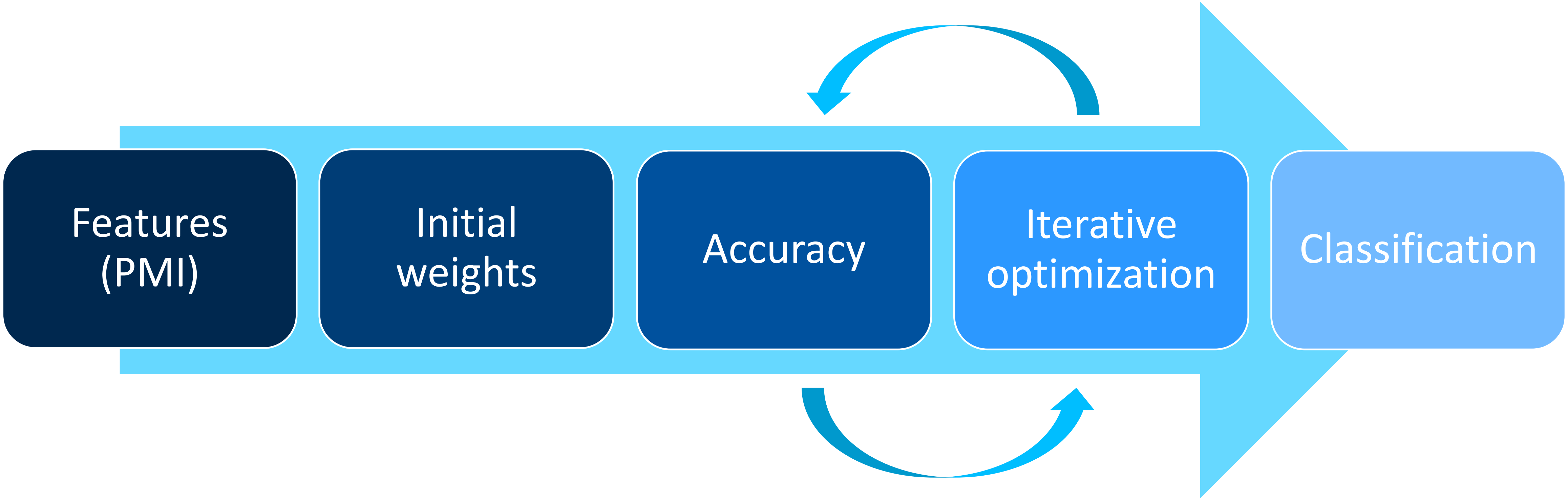
$$\boxed{p_\lambda(y|\bar{x}) = \frac{\exp \sum_i \lambda_i \cdot f_i(y,\bar{x})}{\sum_{y'} \exp \sum_i \lambda_i \cdot f_i(y',\bar{x})}}$$

  - Feature f_i: data properties paired with a label
    - e.g.: $f_1(y,\bar{x}) = \begin{cases} 1, & y = \text{Shakespeare} \wedge \text{thou} \in \bar{x} \\ 0, & \text{otherwise} \end{cases}$
  - Weight λ_i for each feature to represent the importance of the feature

⇨Training by optimizing the weights

Approach

Baseline method



Features (PMI)

Initial weights

Accuracy

Iterative optimization

Classification

- Pointwise mutual information
- 30 features per author

- Randomized between -10 and 10

- Count correct classifications with current weights

- Update weights by adding the gradient

- Choose label with highest $p_\lambda(y|\bar{x})$ for each label

Evaluation

Quantitative

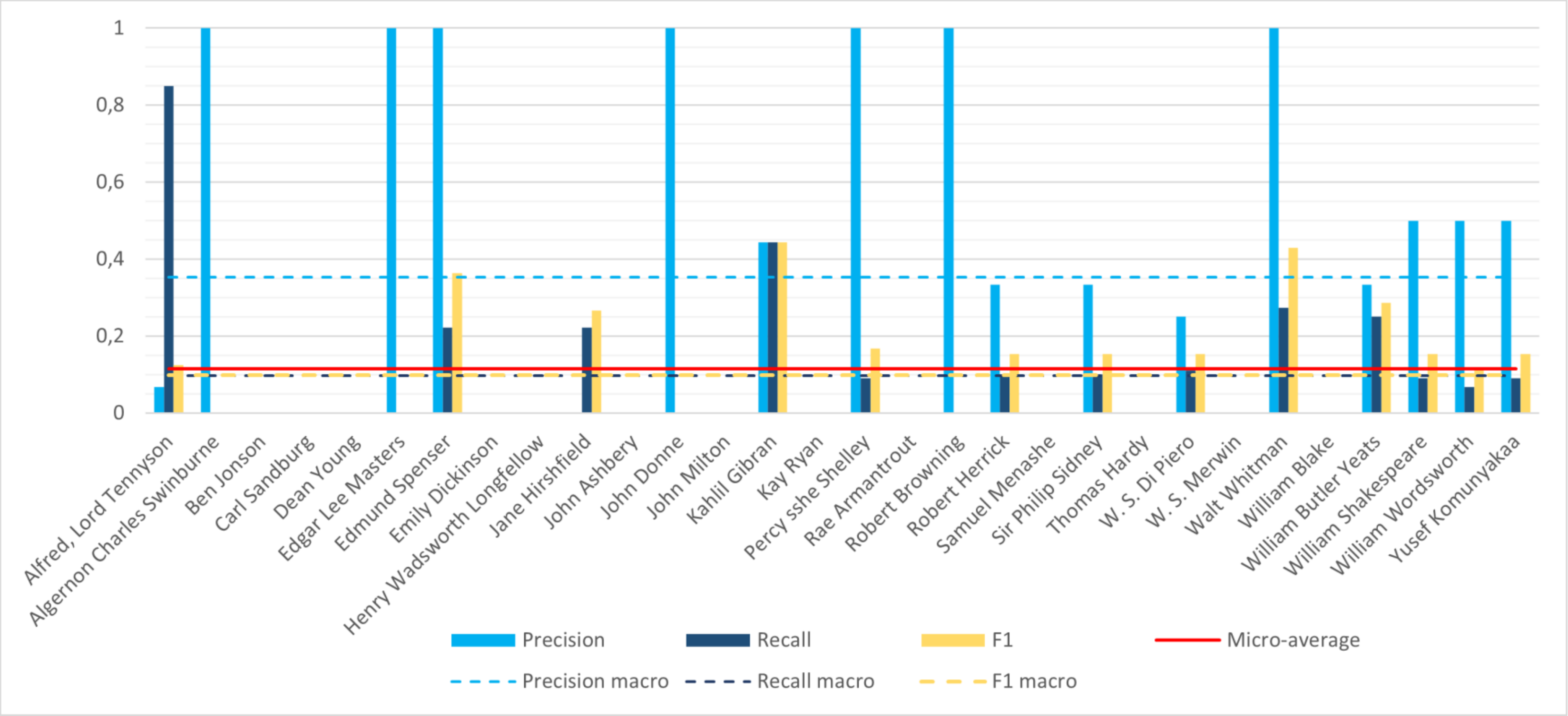


***Fig 2** – Evaluation on test data (338 poems) for 30 authors, trained on 959 poems*

- Accuracy much worse on the test data
  - test: 0.11
  - training: 0.49
- Many authors never predicted

Evaluation

Qualitative

- Many authors never predicted
  - ? Bad features for this author
  - ? Data not evenly distributed

- Features are equally important (random weights approximate each other)
  - ? Different model or feature selection

- The alphabetically first author is chosen too often
  - Every author has the same probability
  - ? Choose most prolific author instead

Next steps

- Research Question

  ❓ Which features are inherent to poetry writers

  ❓ Are there inter-dependencies between the features

- Advanced method
  - Extending the features ⇨ until 07/05
    - Number of stanzas and verses ⇨ Katrin
    - Rhyming and rhyme schemes ⇨ Carlotta
  - Non-linear model for sparse features and interaction
    - Multi-layer NN ⇨ until 07/12

University of Stuttgart
Germany

Thank you!

Katrin Schmidt
katrin.schmidt@ims.uni-stuttgart.de

Carlotta Quensel
carlotta.quensel@ims.uni-stuttgart.de