Feature Selection in Authorship Attribution: Ordering the Wordlist

•••

Maciej Eder (Institute of Polish Language PAS, Kraków) @MaciejEder

Joanna Byszuk (Institute of Polish Language PAS, Kraków) @jbyszuk

DH Conference 2019 in Utrecht

Research problem

Features in authorship attribution

- Linguistic/stylistic characteristics of texts allowing for authorship identification
- Lexical
  - Word frequencies
- Syntactic
  - Frequencies of grammatical categories
- Prosodic
- … and much more (e.g. char n-grams)

Features in machine learning

- A well-researched topic of feature selection
  - dimension reduction
  - shrinkage
  - penalization
  - …
- We *don't* aim at selecting a subset of features

The aim of the study

- Rearranging the set of lexical features (wordlist)
- Some deeper linguistic understanding of the most distinctive features
- Discover if words efficient in classification share any linguistic properties

What we know

- Grammatical words are strong predictors

  (Mosteller & Wallace, 1964)

- Grammatical words occupy the top of the frequency list

  (Zipf, 1948)

- Therefore: top *N* words are strong predictors

  (common practice, 1980s–)

What we don’t know

- Where is the cut-off point where frequent words don’t discriminate anymore
- (= how many MFWs to take?)
- What is the discrimination power of the features down the list

Most Frequent Words

- MFWs = words ordered according to their frequencies
- = mean TF (term frequencies)
- Prioritizes common grammatical words
- Hapax legomena at the bottom of the list
- Proper nouns (names) somewhere in between

TF-IDF

- Term Frequency / Inverse Document Frequency
- Commonly used in information retrieval
- A way of extracting "keywords"
- It prioritizes words *important* for particular texts
- It prioritizes proper nouns
- Grammatical words usually at the bottom

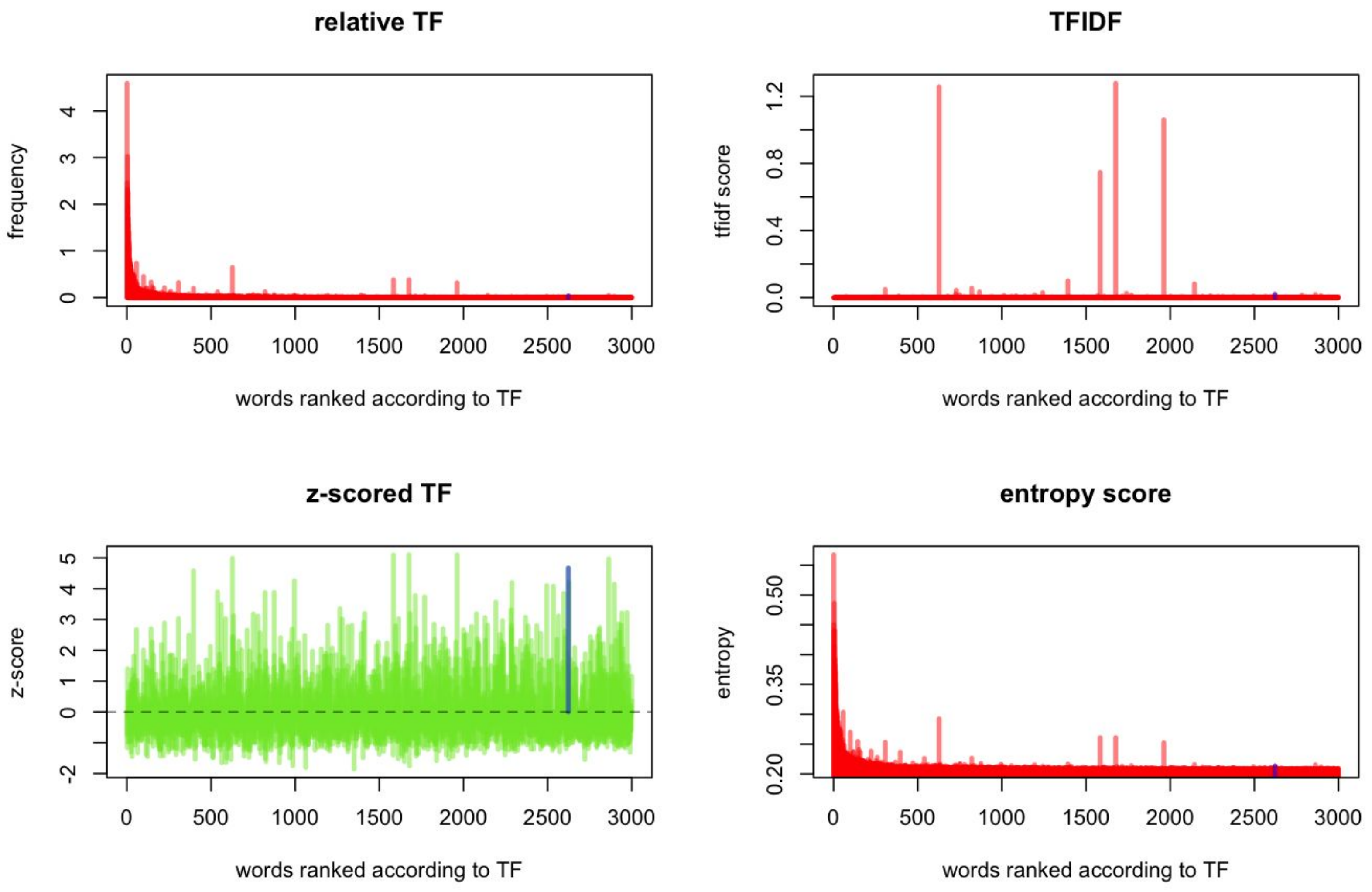
Coefficient of Variation

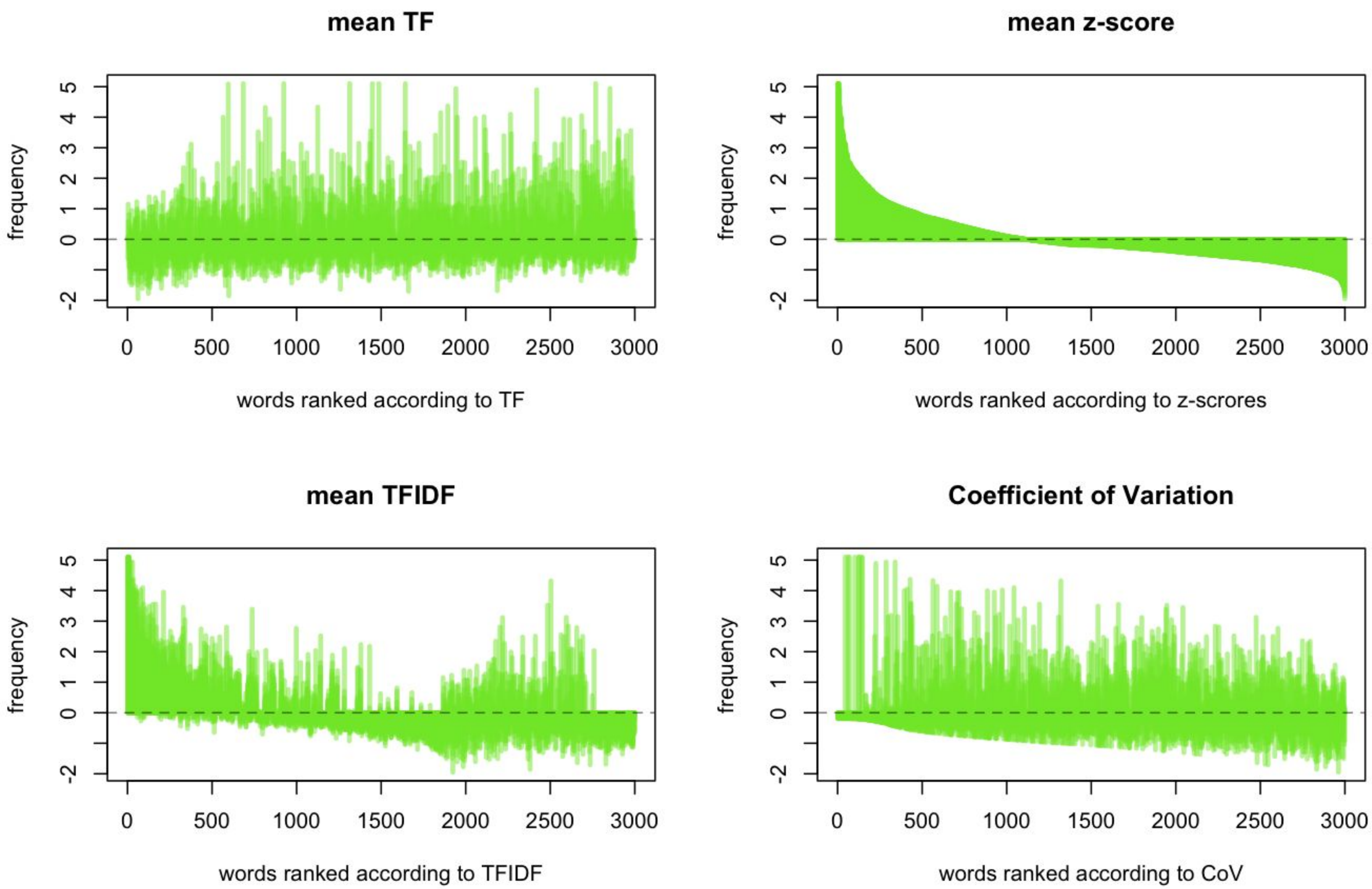
- Variance – the degree to which a given feature (a word) varies in a corpus
- A word of the biggest variance = potential discriminator
- However: variance depends on frequency
- Therefore: coefficient of variation:

$$\mathrm{CoV}_i = \sigma_i / \mu_i$$

Ordering ≠ Weighting

- Features (word frequencies) might be transformed (weighted) differently
  - Term Frequency (= no weighting)
  - Z-scores (cf. Evert et al., 2017, etc.)
  - Term Frequency / Inverse Document Freq.
  - Mutual Information
- This study is *not* about weighting evaluation


relative TF
frequency
0
1
2
3
4
words ranked according to TF
0
500
1000
1500
2000
2500
3000
TFIDF
tfidf score
0.0
0.4
0.8
1.2
words ranked according to TF
z-scored TF
z-score
-2
0
1
2
3
4
5
words ranked according to TF
entropy score
entropy
0.20
0.35
0.50
words ranked according to TF


mean TF
frequency
words ranked according to TF
mean z-score
frequency
words ranked according to z-scrores
mean TFIDF
frequency
words ranked according to TFIDF
Coefficient of Variation
frequency
words ranked according to CoV
0
500
1000
1500
2000
2500
3000
-2
0
1
2
3
4
5

Ordering ≠ Weighting

	TF	z-scores	TF-IDF	z-scored TF-IDF
mean TF (=MFWs)	✗	✓	✗	✗
TF-IDF	✗	✓	✗	✗
variance	✗	✓ (?)	✗	✗
CoV	✗	✓	✗	✗

Data

Dataset

A Small Collection

100 Polish (3 texts per author, one additional)

100 English (3 texts per author, one additional)

– canon literary texts, similar in topic and dates of creation

Preliminary tests also for French and German

Method

Experimental setup

- Supervised classification
- Leave-one-out cross validation
- kNN classifier (k=1), *aka* Delta
  - Also: SVM, NSC
- Cosine Delta distance measure
- A set of 10 subsequent features tested:

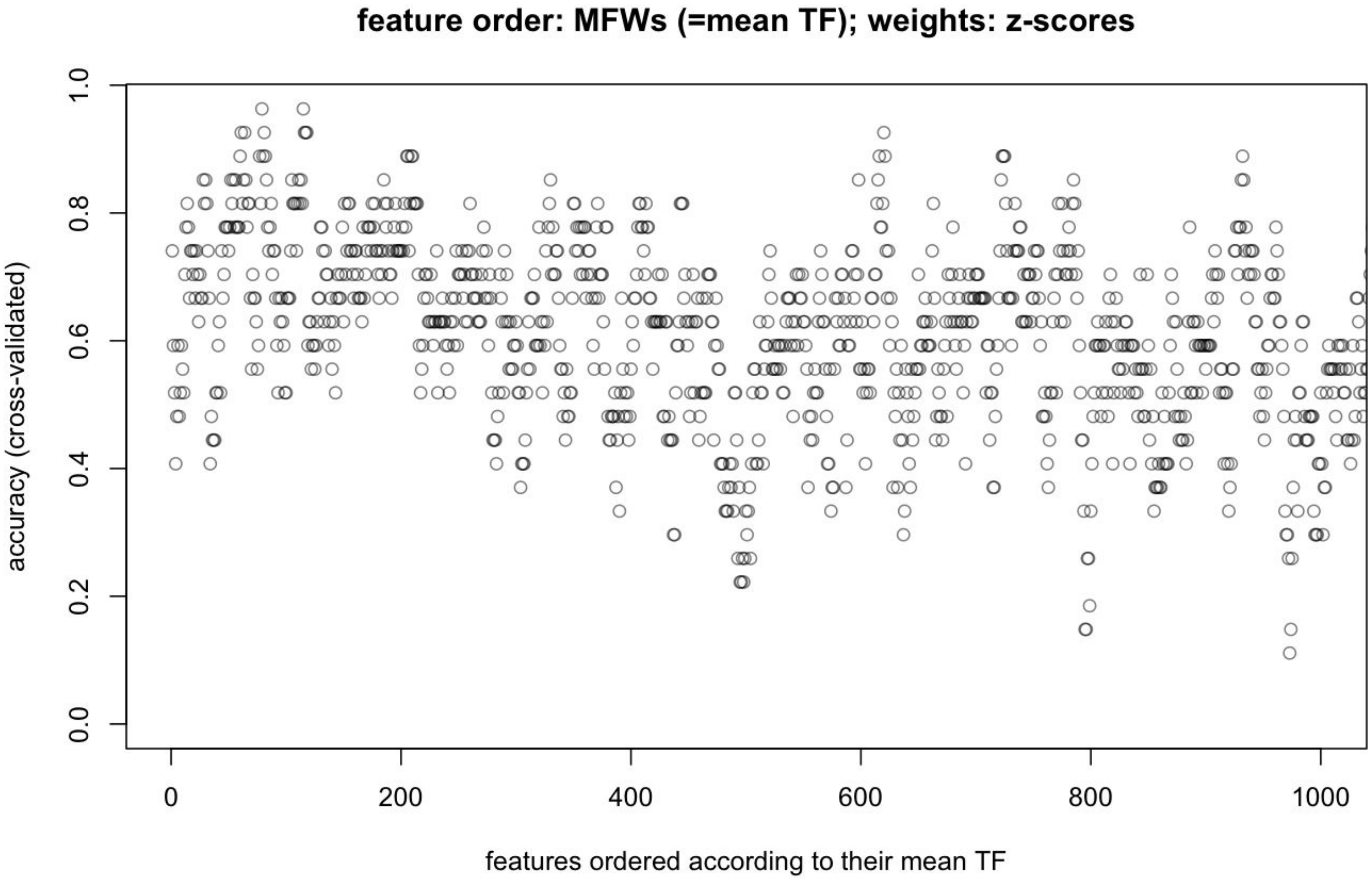
$$F_k = \{w_i, w_{i+1}, \ldots, w_{i+9}\}$$

N subsequent features

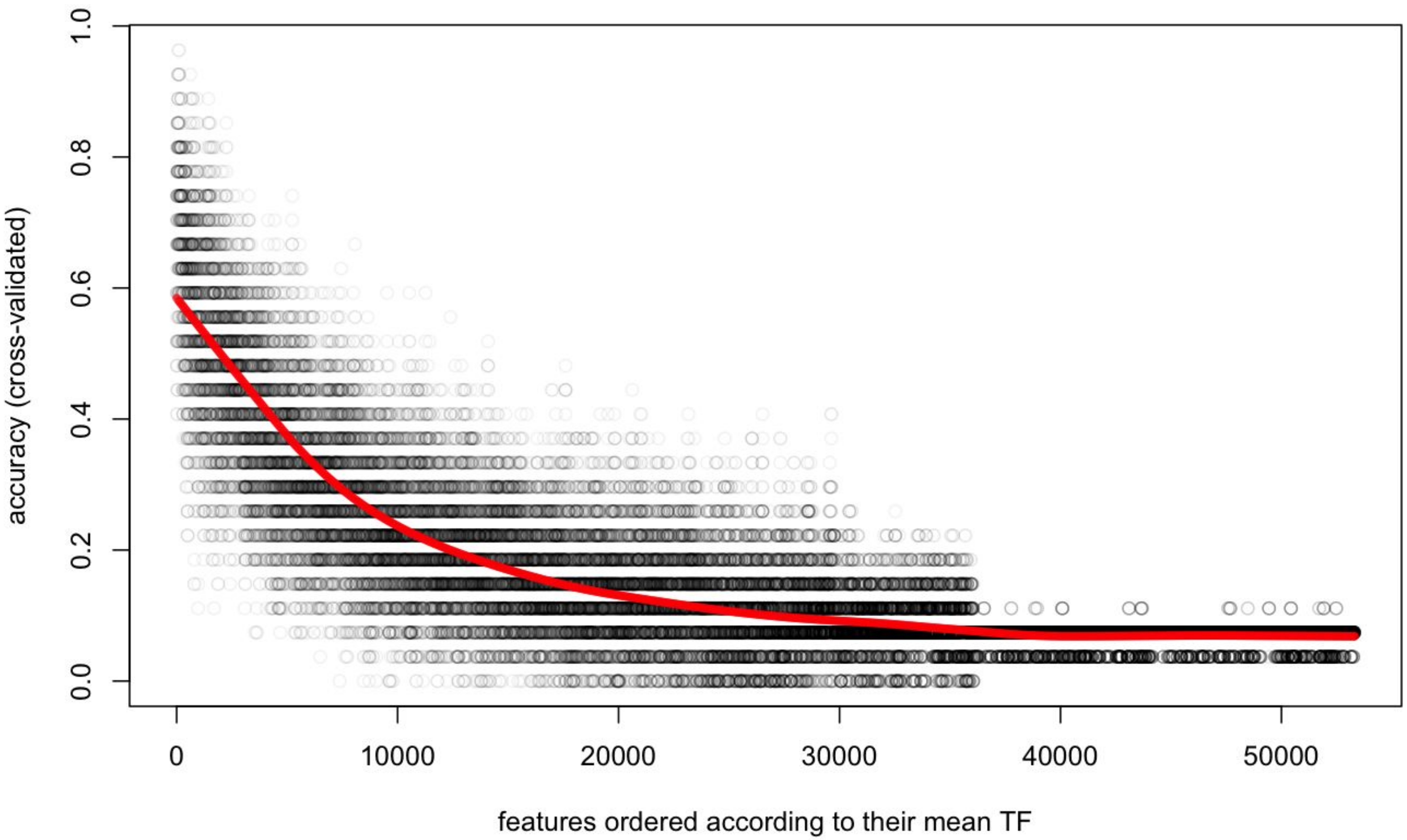
the and to of I a in that he was it you her …

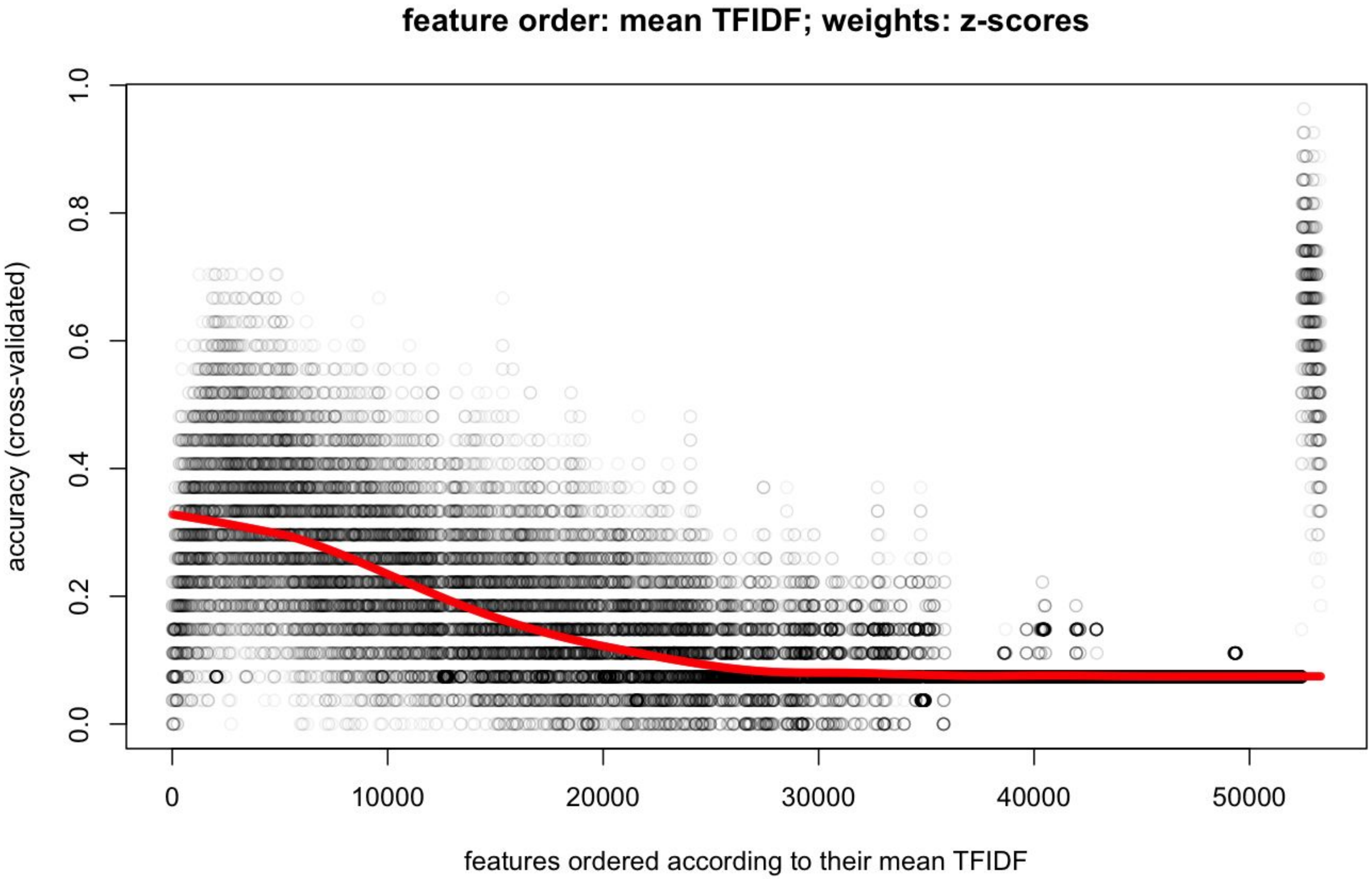
1. the and to of I [# 1]
2. and to of I a [# 2]
3. to of I a in […]
4. of I a in that […]
5. … … … … … [# *n*]

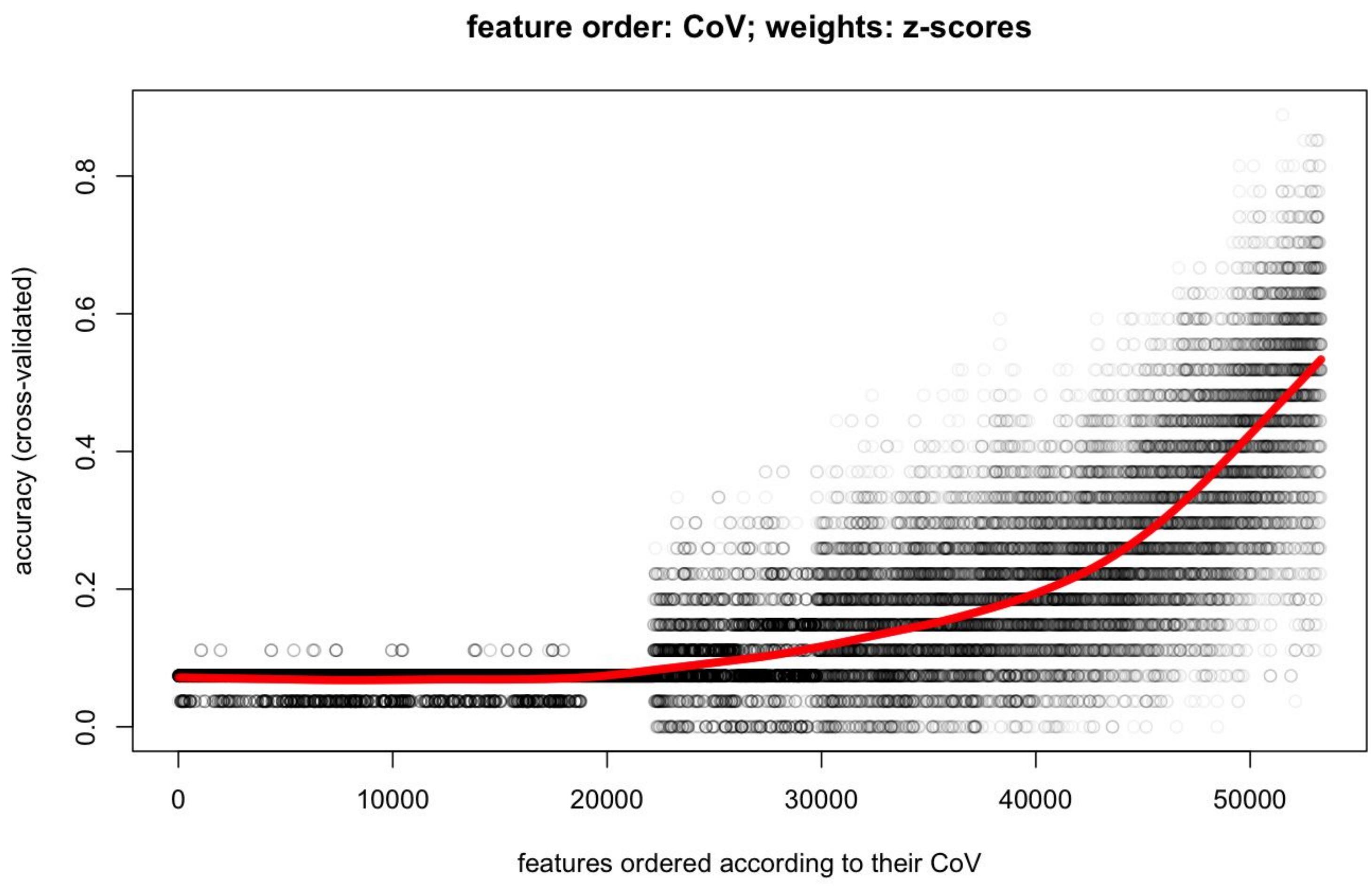
Results


feature order: MFWs (=mean TF); weights: z-scores
accuracy (cross-validated)
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
features ordered according to their mean TF

feature order: MFWs (=mean TF); weights: z-scores



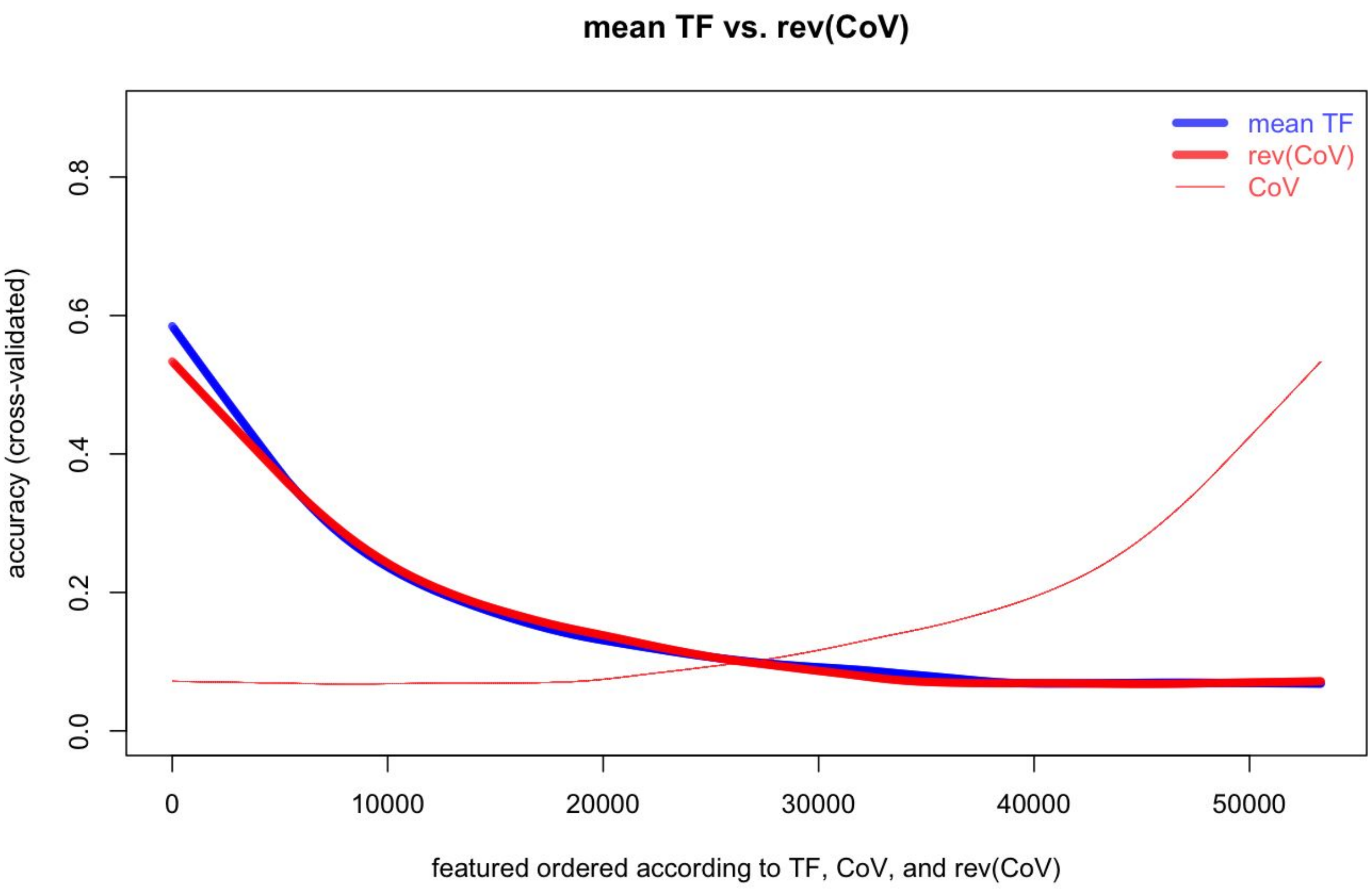

feature order: mean TFIDF; weights: z-scores
accuracy (cross-validated)
0.0
0.2
0.4
0.6
0.8
1.0
0
10000
20000
30000
40000
50000
features ordered according to their mean TFIDF


feature order: CoV; weights: z-scores
accuracy (cross-validated)
0.0
0.2
0.4
0.6
0.8
0
10000
20000
30000
40000
50000
features ordered according to their CoV

First observations

- MFW – working great (phew, we knew!)
- CoV – unexpectedly well, in fact…
- CoV aggregating good features even better than MFW!

So what if…

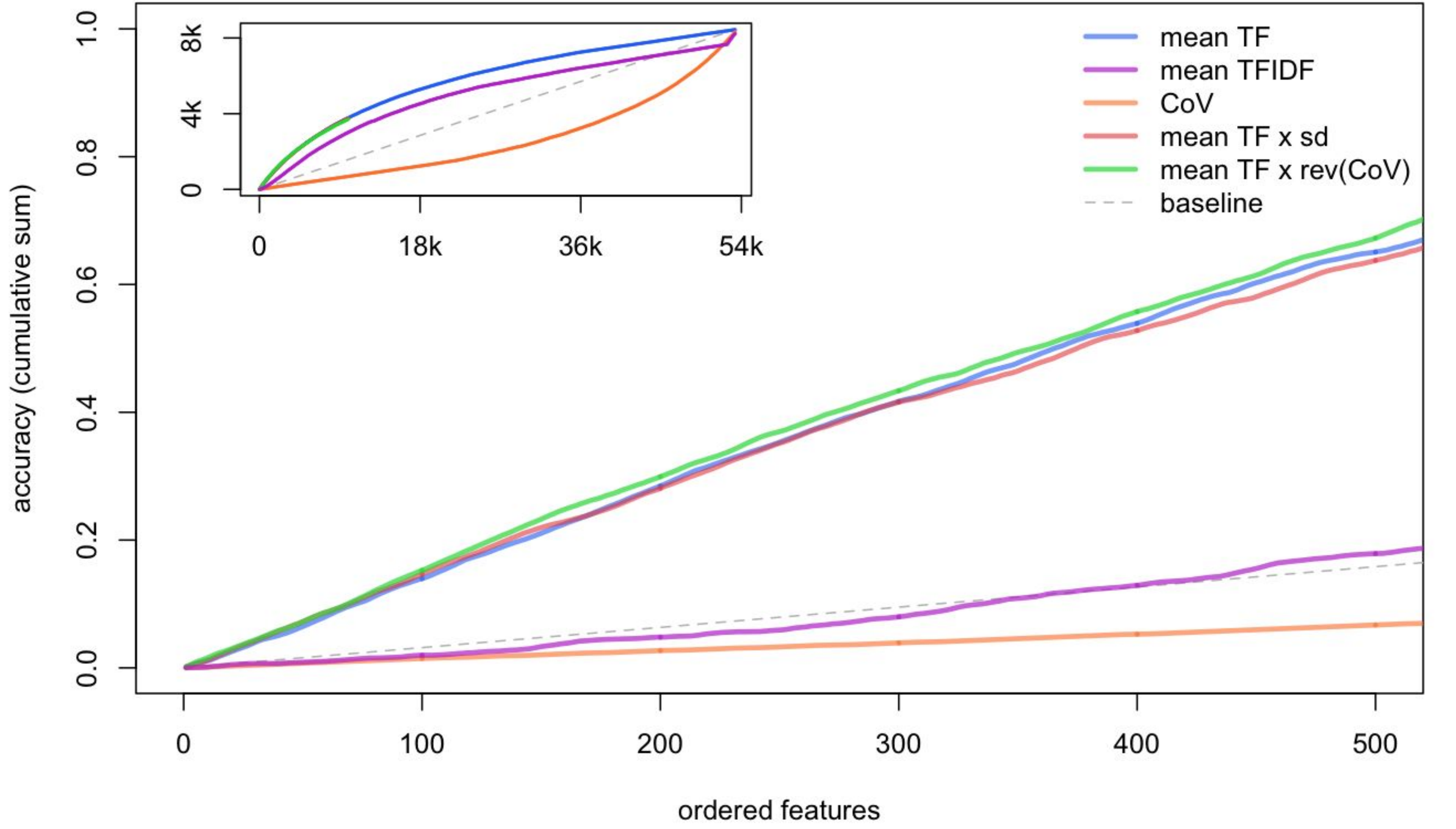

mean TF vs. rev(CoV)
mean TF
rev(CoV)
CoV
accuracy (cross-validated)
0.0
0.2
0.4
0.6
0.8
0
10000
20000
30000
40000
50000
featured ordered according to TF, CoV, and rev(CoV)

Combining TF and rev(CoV)?

$$\omega_i = TF_i \times rev(CoV)_i$$

Which can be represented as:

$$\omega_i = \mu_i \times rev(\sigma_i / \mu_i)$$



mean TF
mean TFIDF
CoV
mean TF x sd
mean TF x rev(CoV)
baseline
accuracy (cumulative sum)
ordered features
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
0
4k
8k
18k
36k
54k

Conclusions

- TF aka MFW – confirmed as a generally good way of ordering of features
- TF-IDF – useful for detection of what to cull
- TFxCOV – a promising update of traditional approach

Acknowledgements

We are grateful for the financial support we received:

JB was partially funded for the research by Poland's National Science Centre (grant number 2017/26/HS2/01019).

ME was partially funded by Poland’s National Science Centre (grant number 2014/12/W/ST5/00592).

National Science Centre Poland

PAN IJP Institute of Polish Language Polish Academy of Sciences

10 Computational 01 01 Stylistics 0101000 11 Group 011010110

Thank you!

Presentation (and future place for code):
https://github.com/JoannaBy/Feature-Selection-in-Authorship-Attribution

@MaciejEder
@jbyszuk

What words in TFxCoV?

- largely overlaps with word frequencies (so basically - term frequencies),
- with some of the words taking primary position over personal pronouns:
    - e.g. the discriminative power of prepositions, such as "to", "as", "with", "for", "at", "from", "before.