Classifying Headlines

Harry Chirayil

Christopher Keim

Stefan Schmutz

Can we predict the source of a News Headline?

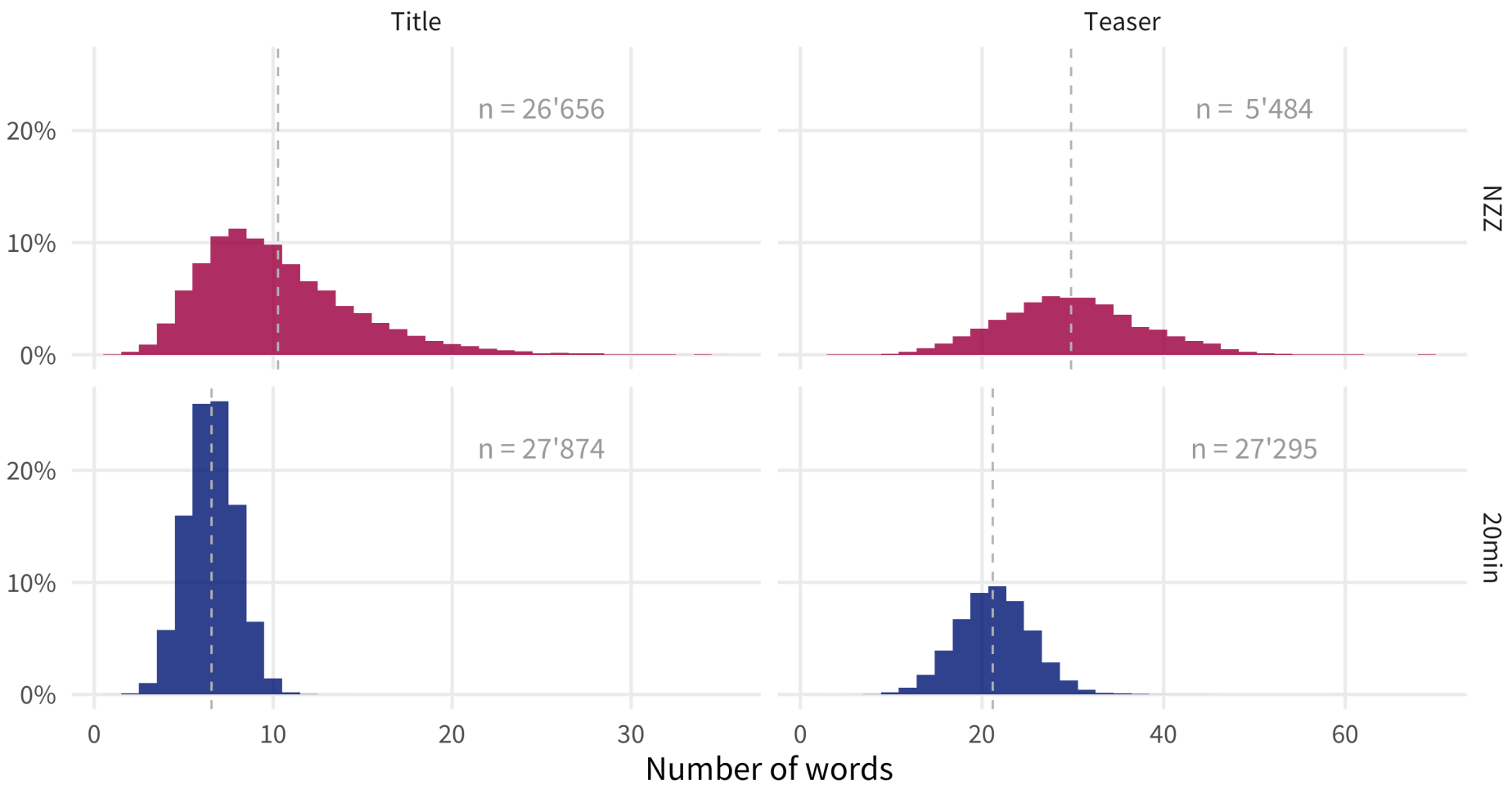
Dramatische Rettung bei der härtesten Segelregatta der Welt
???

Künstliche Intelligenz sagt erfolgreich voraus, zu welcher Struktur sich Proteine falten
???

Neue Zürcher Zeitung

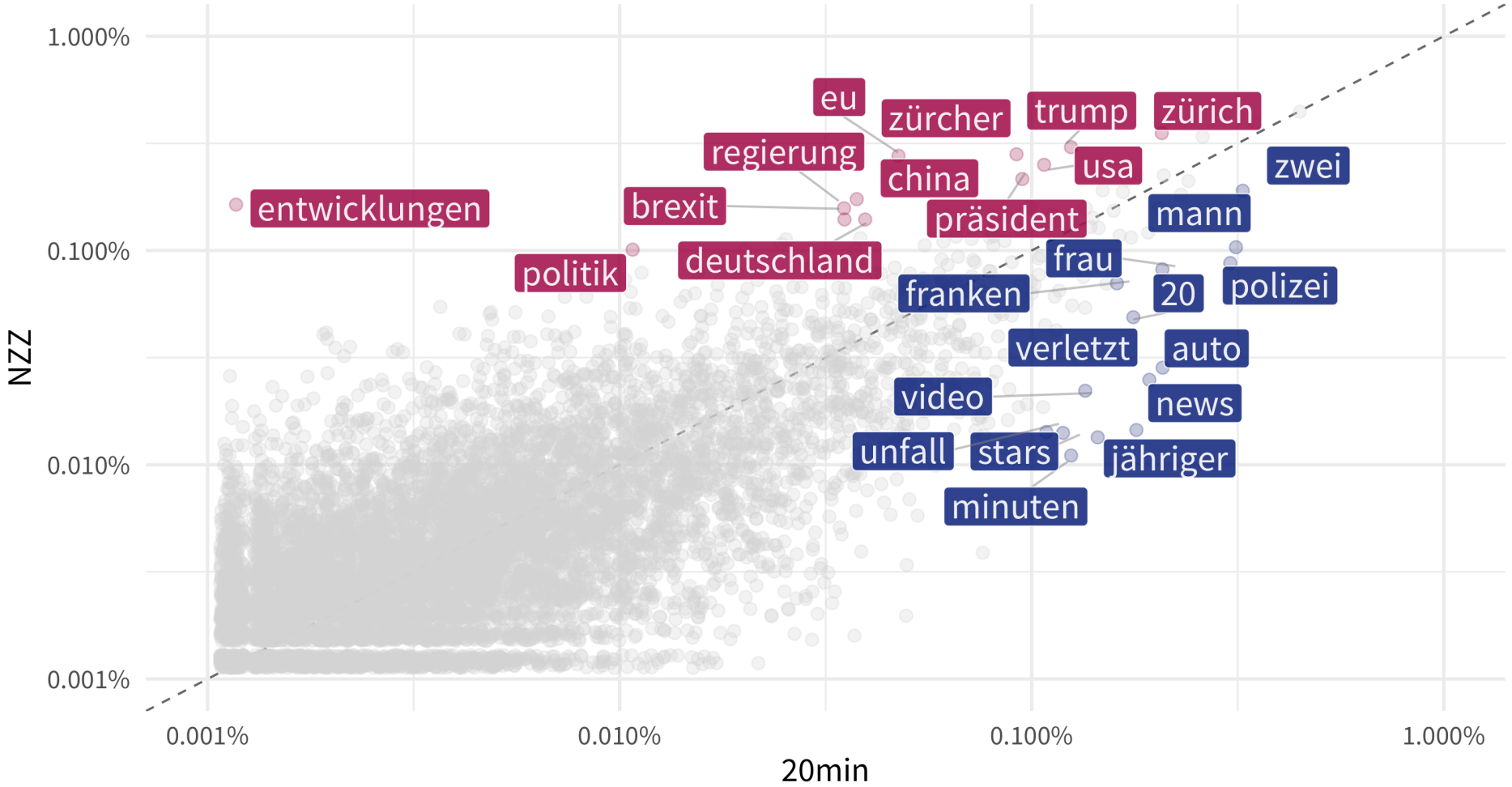
Headline length distribution

amount of words in **NZZ** and **20min** headlines



Proportion differences

reveal words which are more prevalent in **NZZ** or **20min** headlines



Multinomial Naïve Bayes

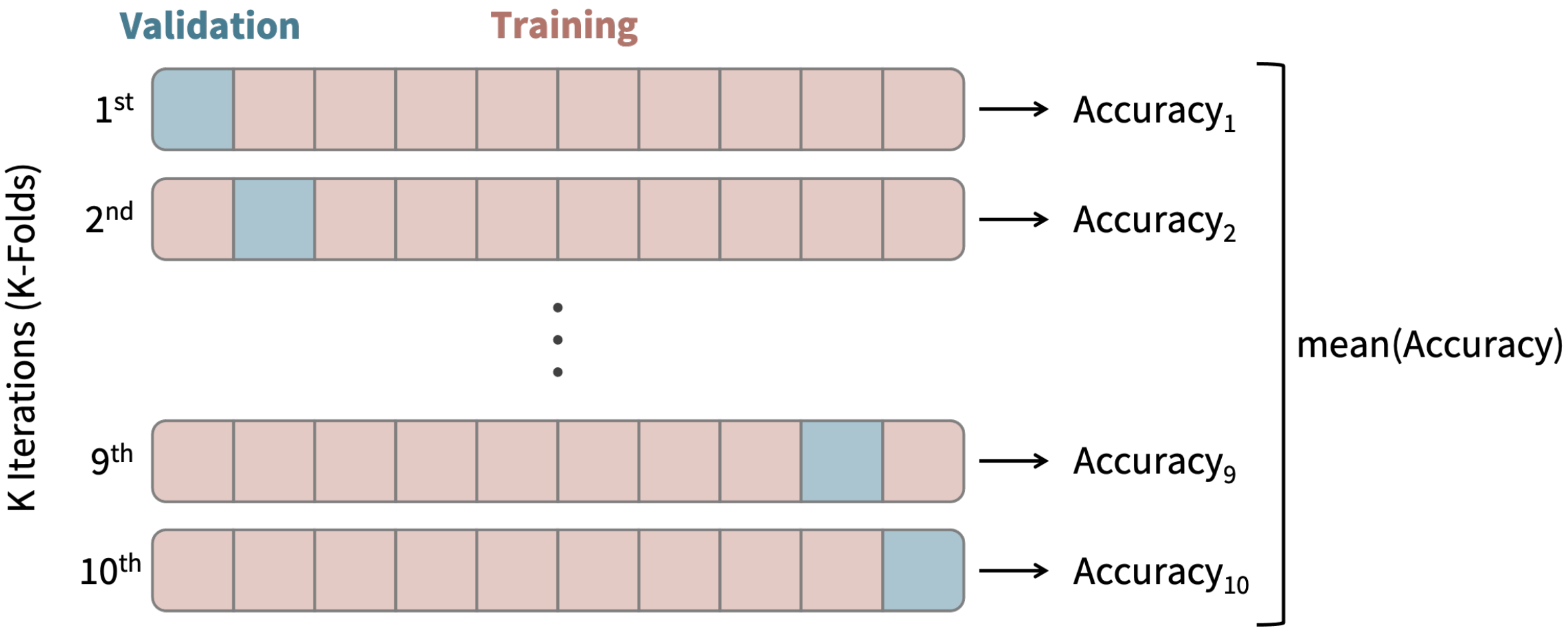
Document Term Matrix

Documents

Terms

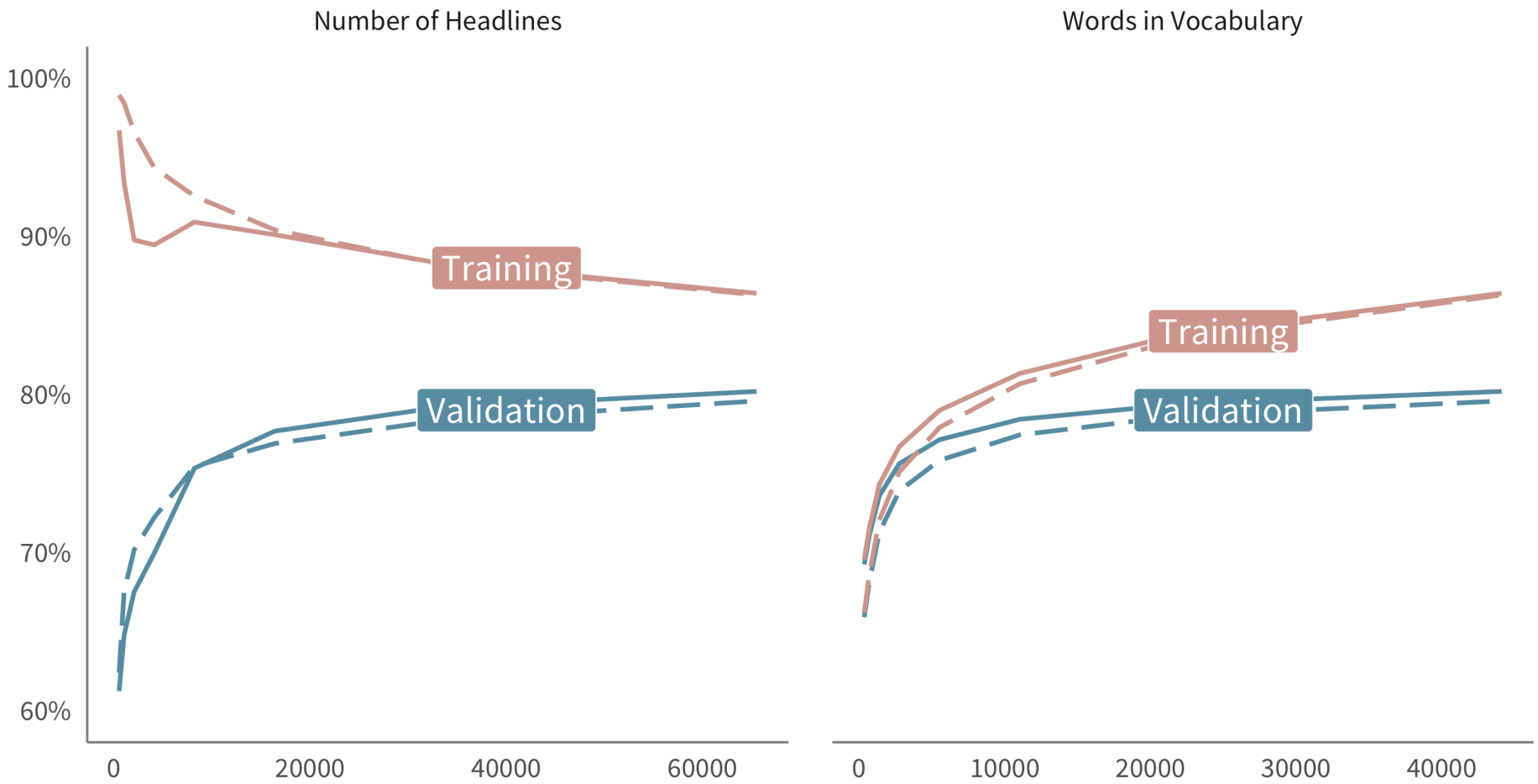
	Word 1	Word 2	Word 3	...	Word n
Headline 1	0	1	1	...	0
Headline 2	1	0	0	...	0
Headline 3	2	0	3	...	1
...	...	...	...	...	...
Headline m	1	1	0	...	0

Counts

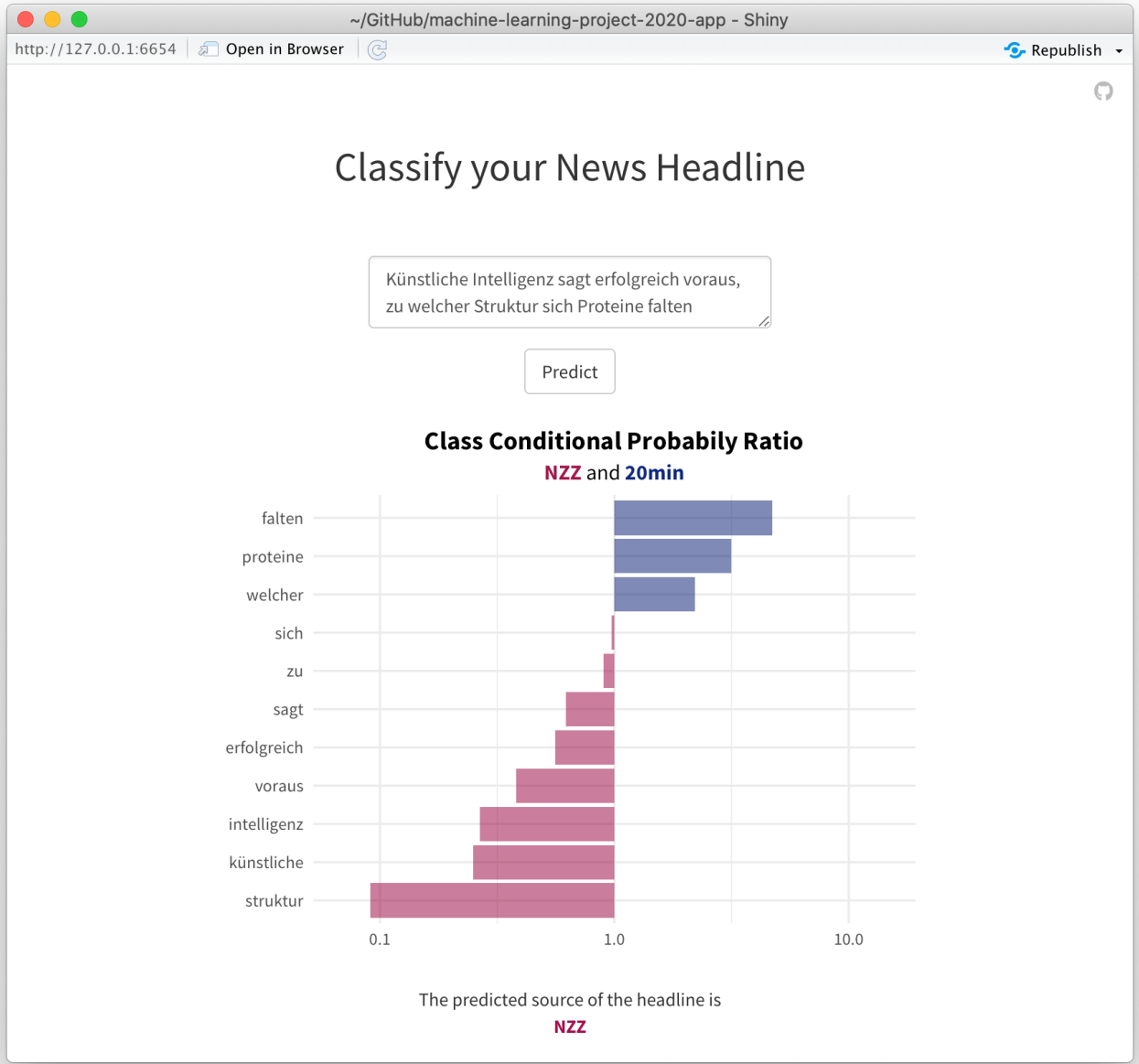

Validation
Training
K Iterations (K-Folds)
1st
2nd
9th
10th
Accuracy1
Accuracy2
Accuracy9
Accuracy10
mean(Accuracy)

10-Fold Cross Validation Accuracy

with **stop words removed** (dashed line) and **included** (solid line)



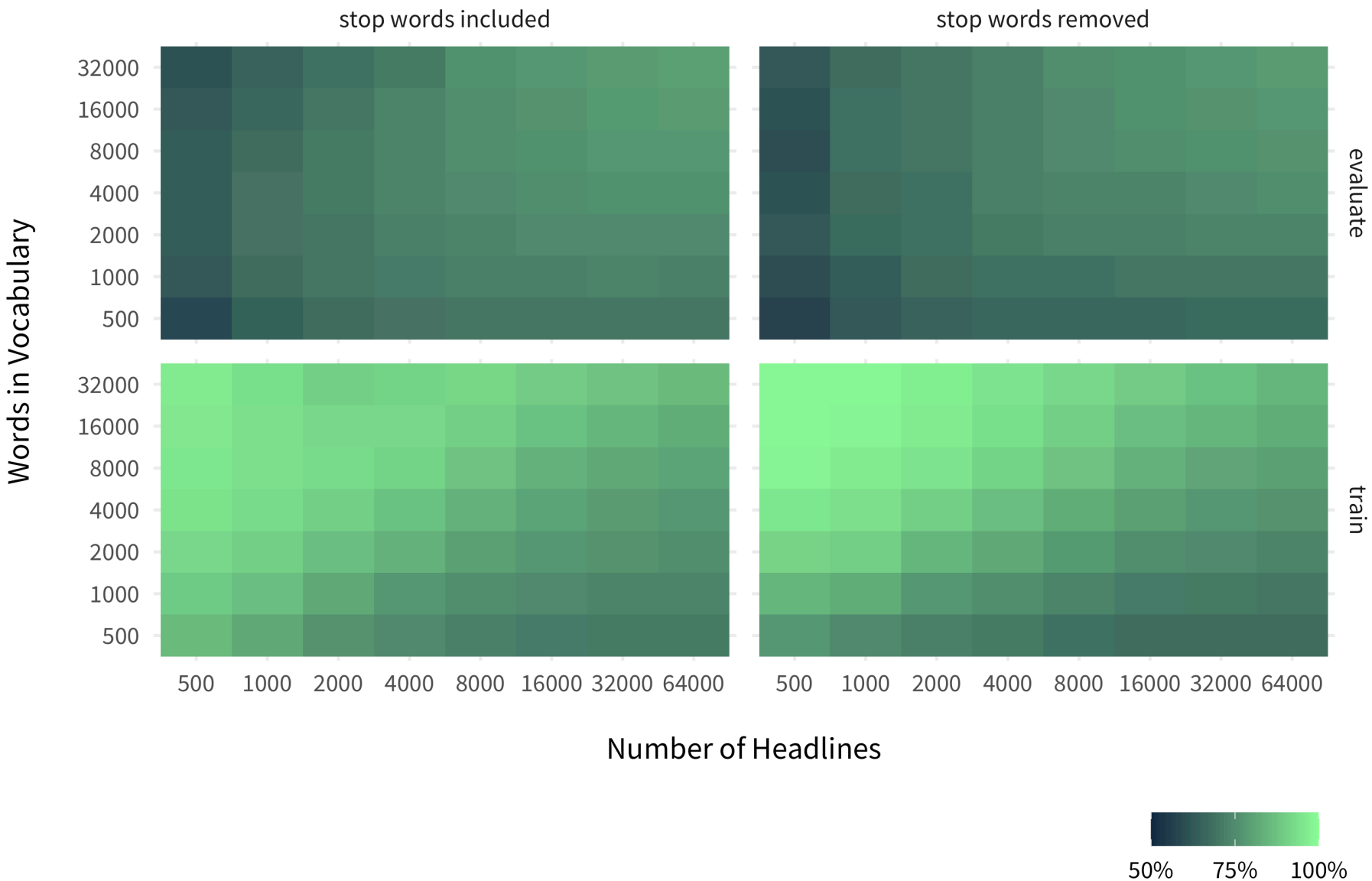
TODO: Insert performance numbers of the test set (confusion matrix and accuracy)

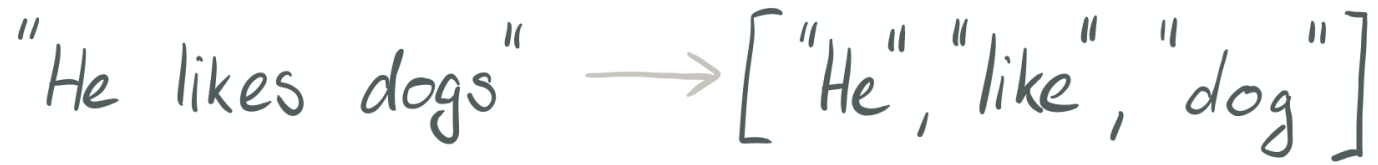


bit.ly/classify-headlines

Appendix

10-Fold Cross Validation Accuracy



Just Tokenisation

Tokenisation and Lemmatisation