# Comparing ML models for single-label text classification

Ana Cardoso Cachopo
Jan 2017

# Outline

- Problem statement, goals and success criteria
- Explain data and run statistical analysis
- Apply ML models to data
- Summary and future work

# Problem statement, goals and success criteria

- How well do different Machine Learning models perform on single-label text classification tasks?

- Goals: reproduce part of my PhD work using state-of-the-art libraries in Python, assess how this area evolved in 10 years.

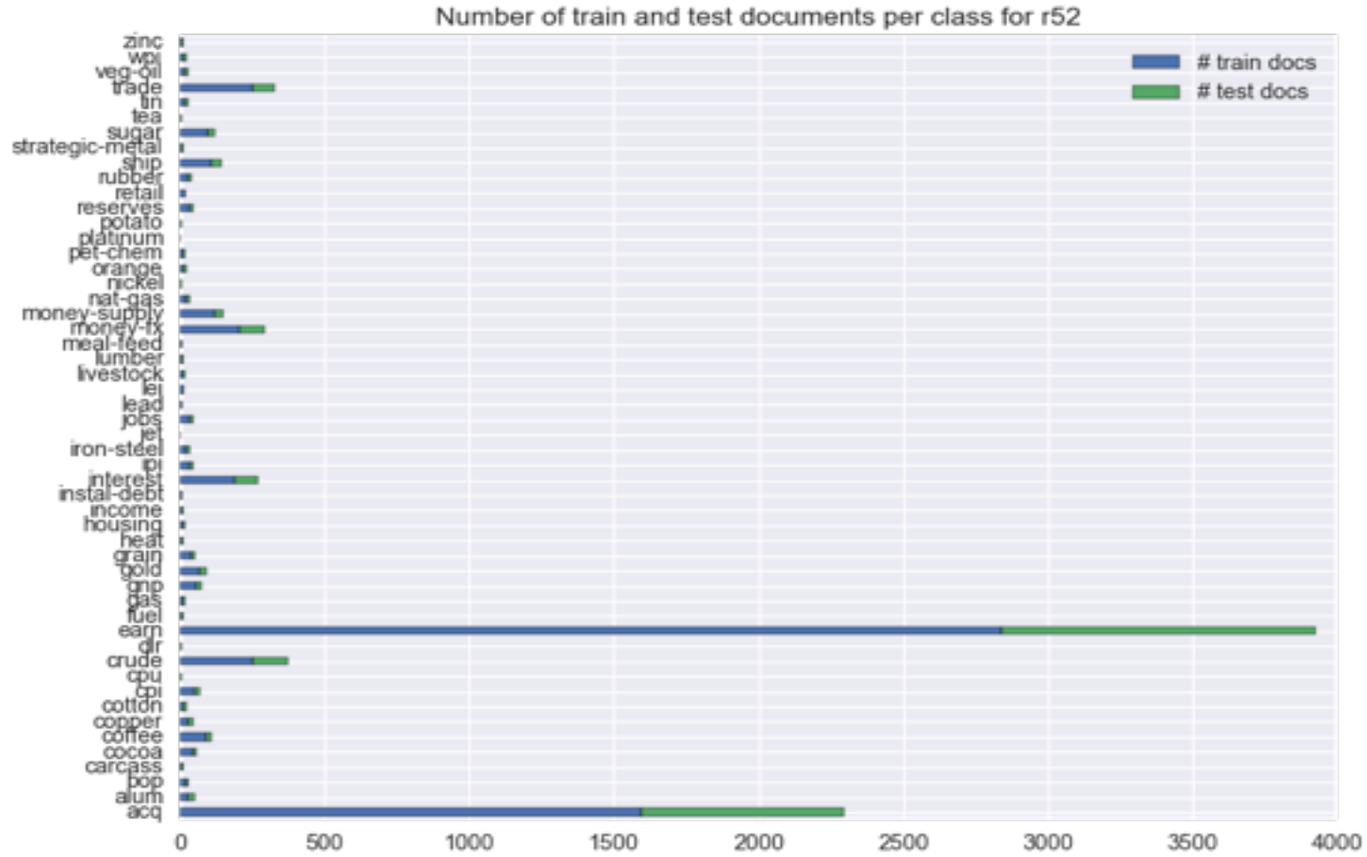
- Successful if: reproduce the initial "related work" from my thesis.

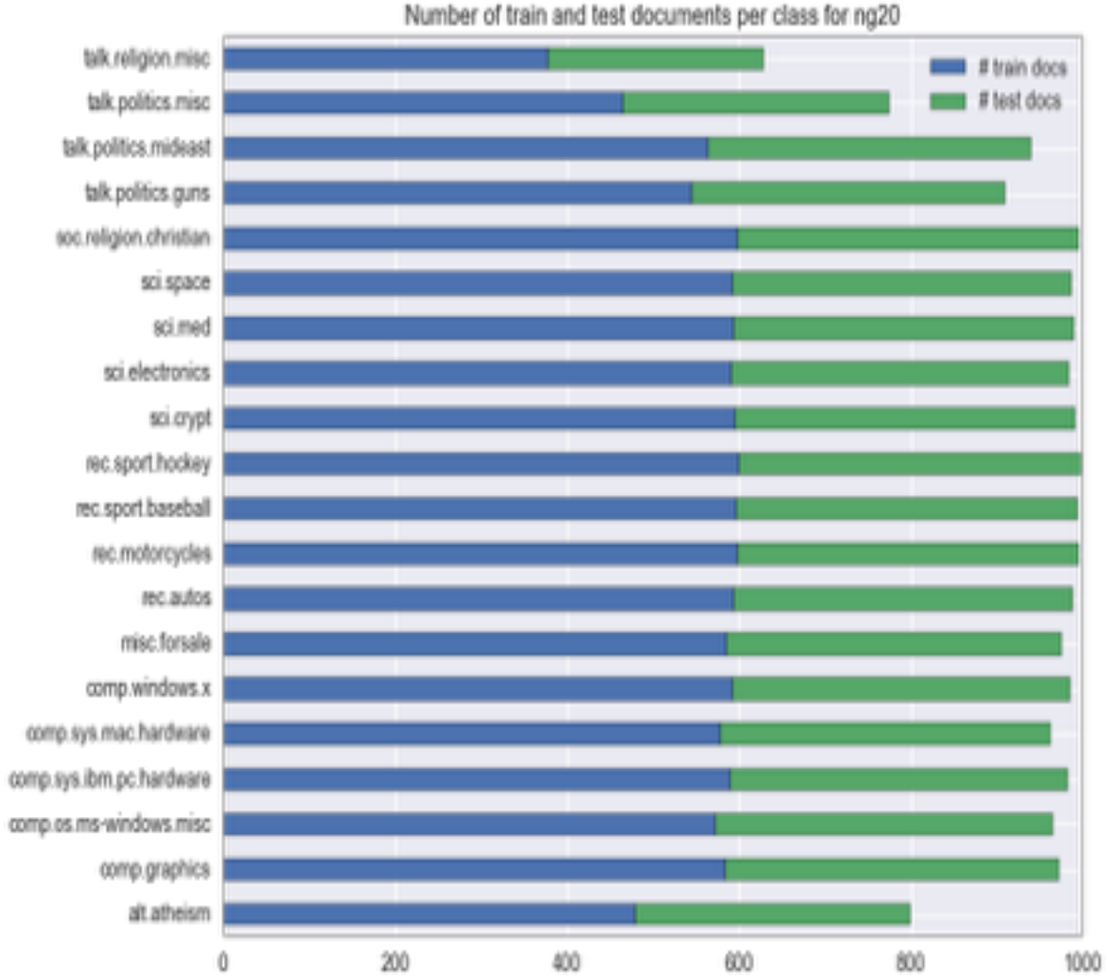
# Explain data and run statistical analysis

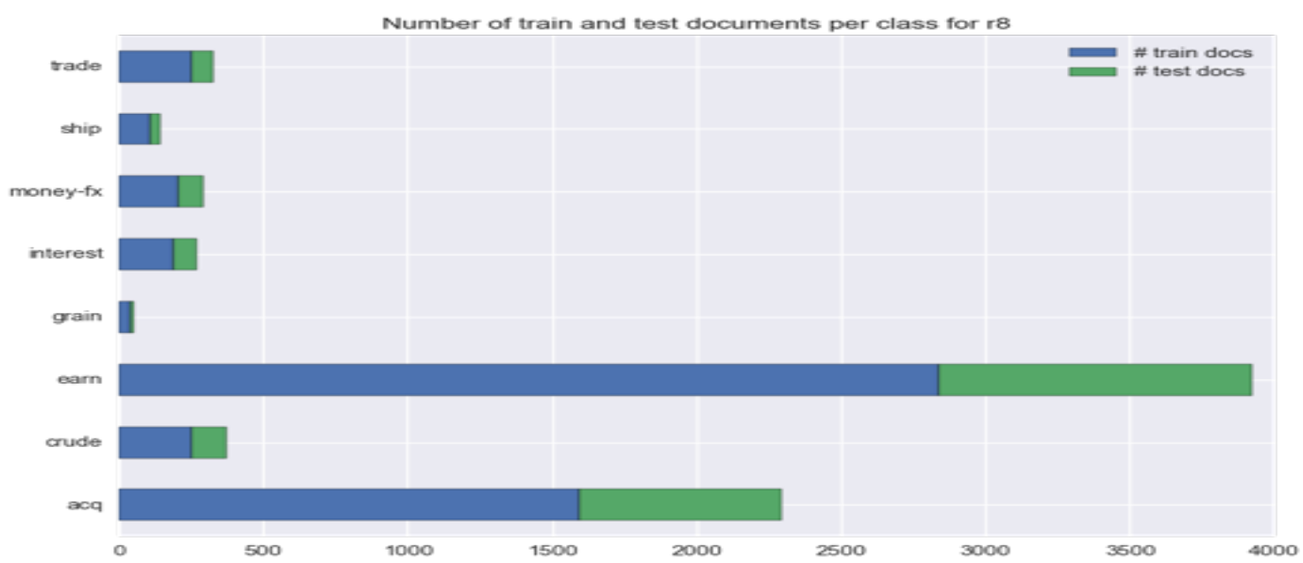
- Datasets: 20 Newsgroups, Reuters-21578
- Numbers of documents
- Numbers of features
- Word clouds

# Numbers of documents

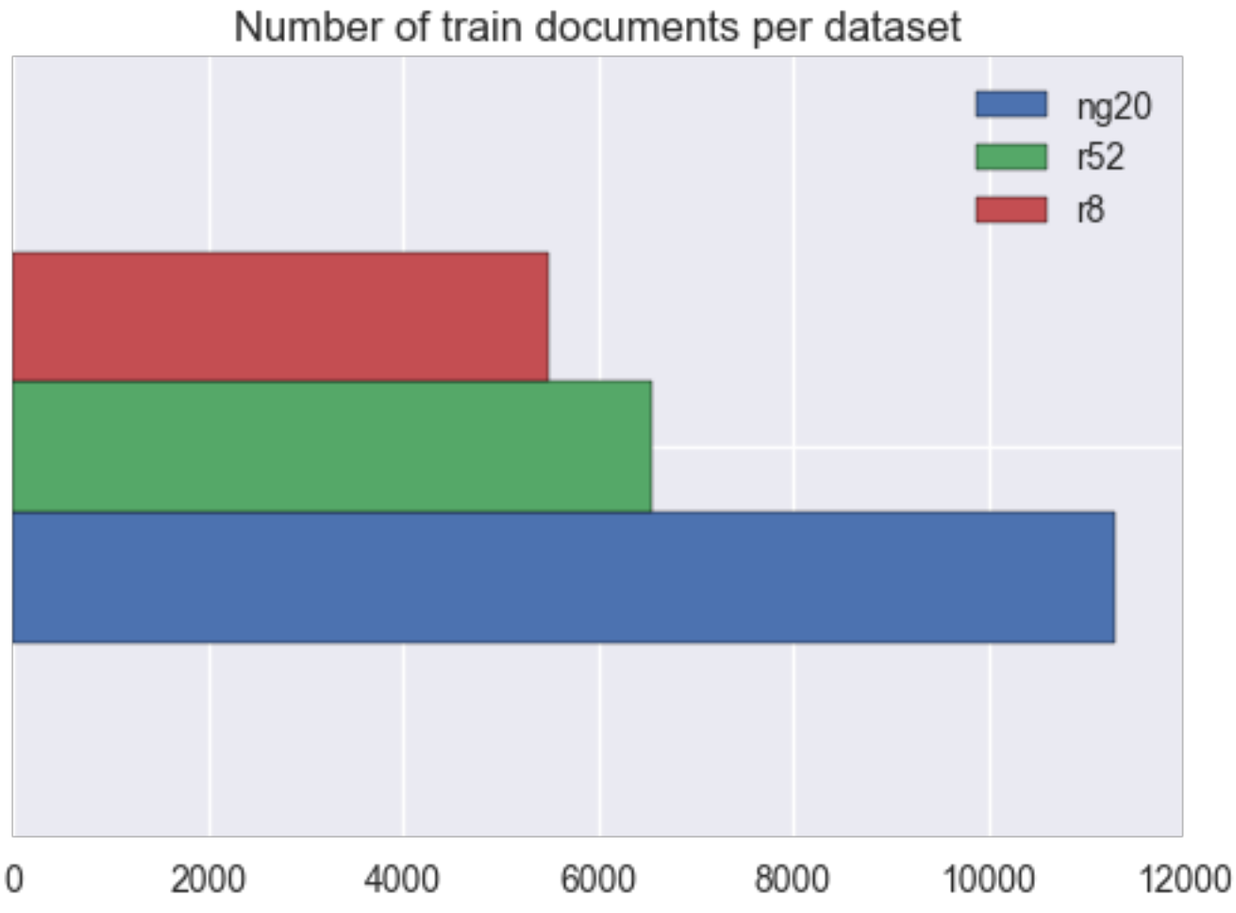
| | # train docs | # test docs | total # docs |
|---|---|---|---|
| **r8** | 5485 | 2189 | 7674 |
| **r52** | 6532 | 2568 | 9100 |
| **ng20** | 11293 | 7528 | 18821 |

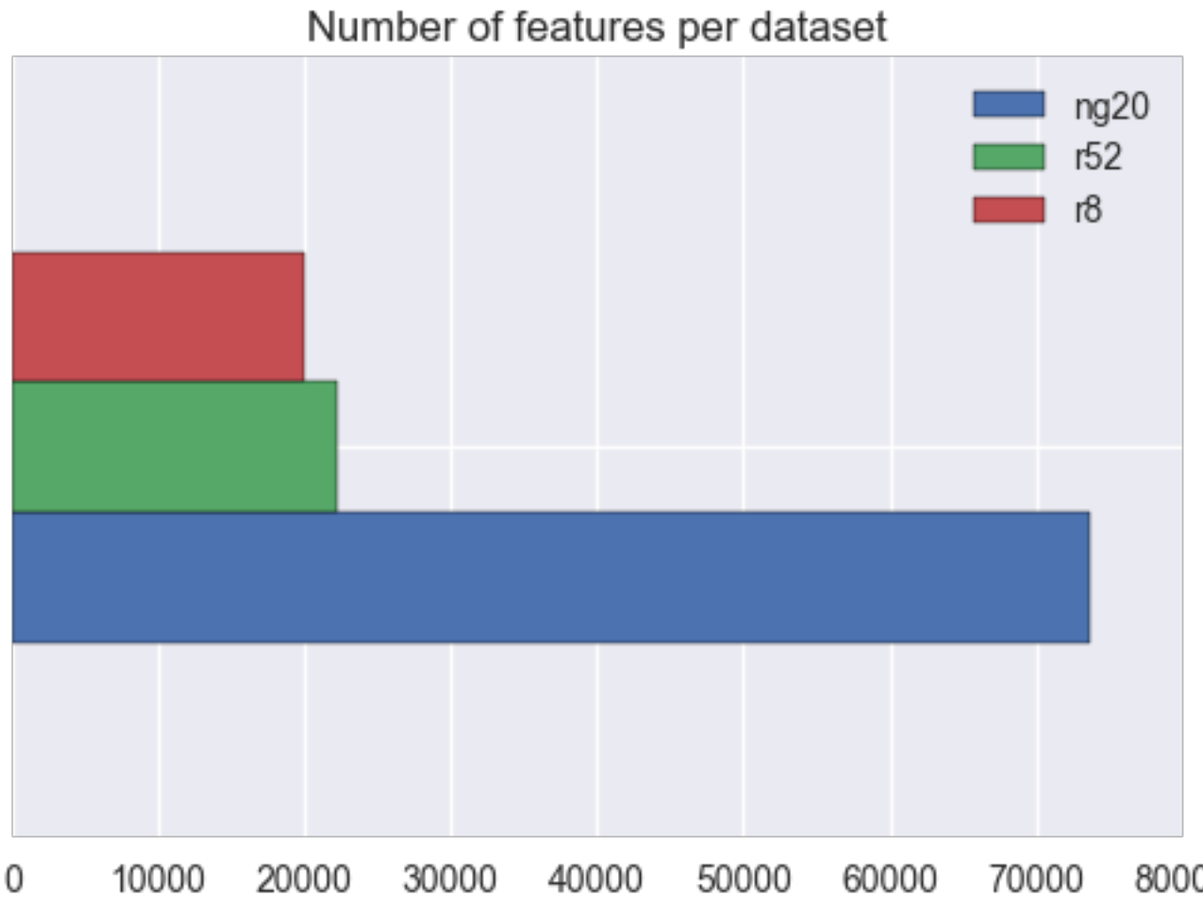


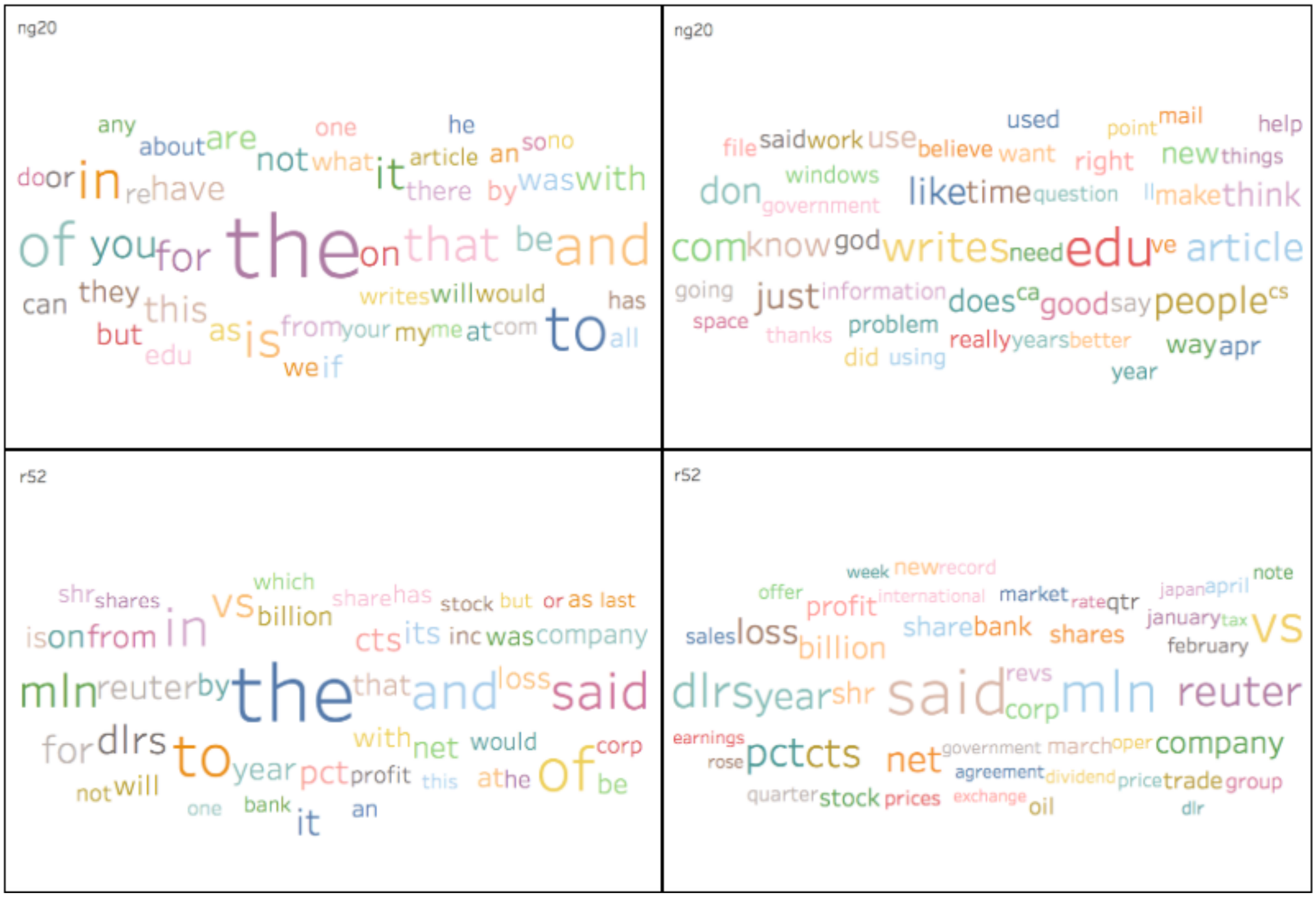



# Numbers of features



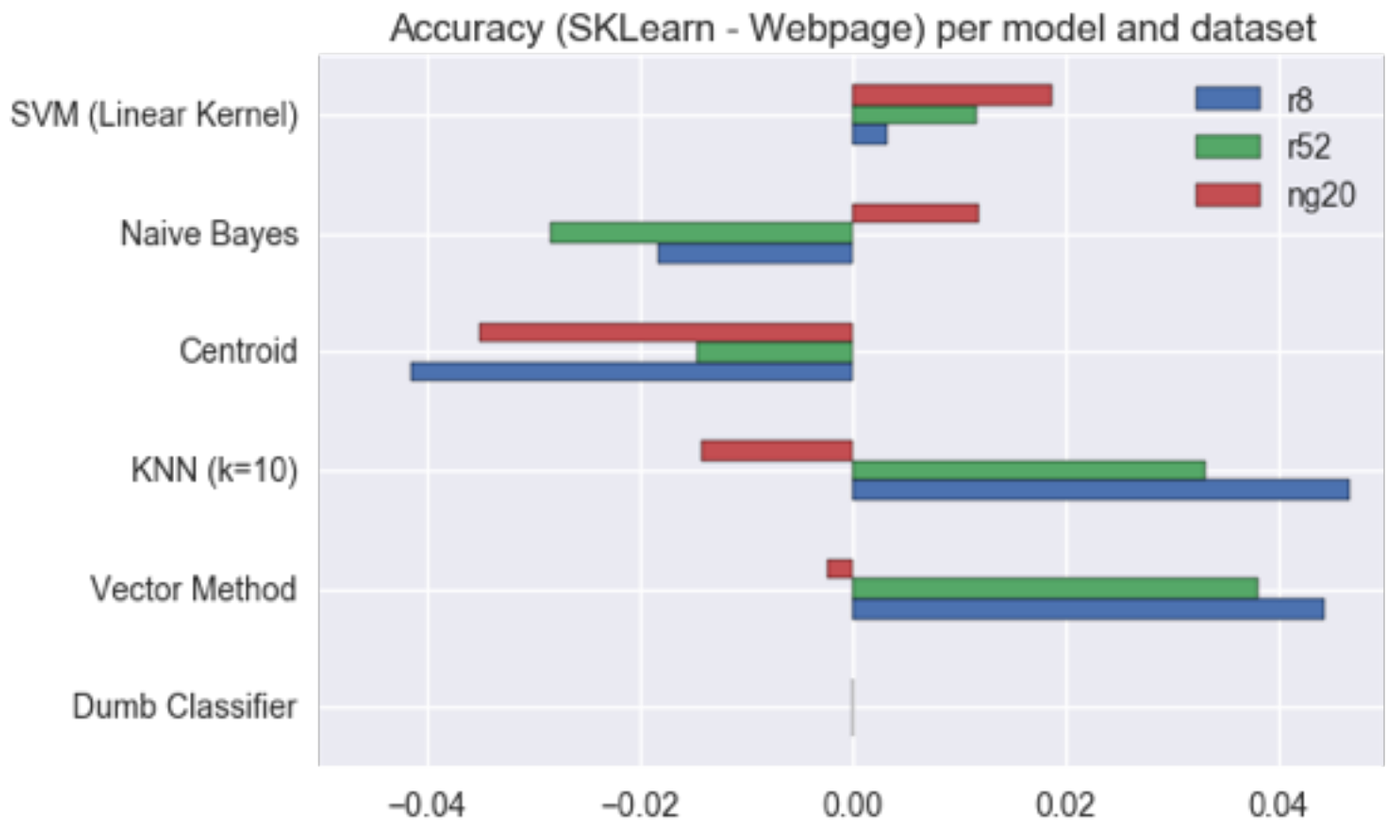
# Word clouds


ng20
any about are one he
door in re have not what it article an so no
there by was with
of you for the on that be and
can they this writes will would has
but edu as is from your my me at com to all
we if
ng20
used point mail help
file said work use believe want right new things
windows
don government like time question ll make think
com know god writes need edu ve article
going just information does ca good say people cs
space thanks problem really years better way apr
did using
year
r52
which
shr shares vs billion share has stock but or as last
is on from in
cts its inc was company
mln reuter by the that and loss said
for dlrs to year pct profit this at the of be corp
with net would
not will
one bank it an
r52
week new record
offer profit international market rate qtr japan april note
sales loss billion share bank shares january tax vs
february
dlrs year shr said revs corp mln reuter
earnings rose pct cts net government march oper company
agreement dividend price trade group
quarter stock prices exchange oil
dlr

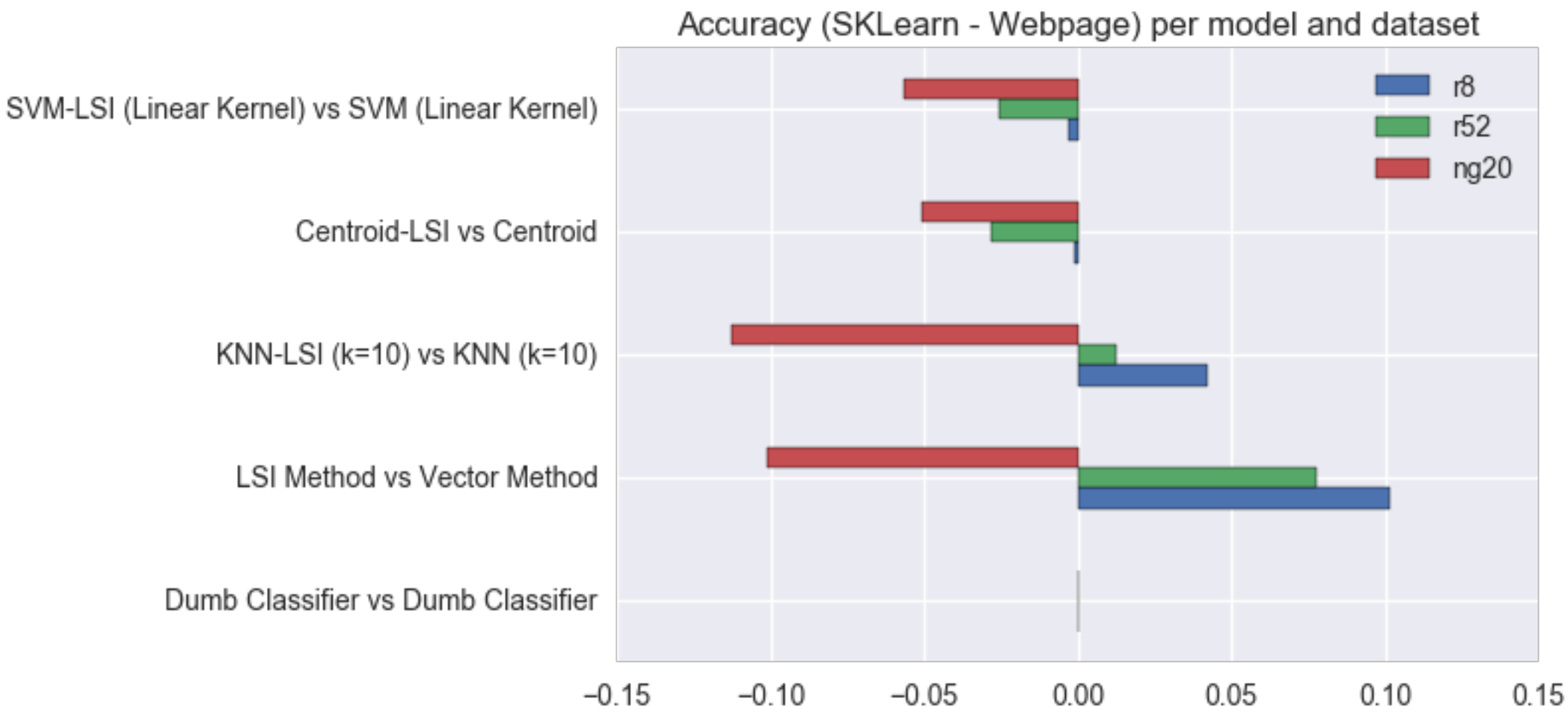
# Word clouds (cont.)

# Apply ML models
## (reproduce previous results)

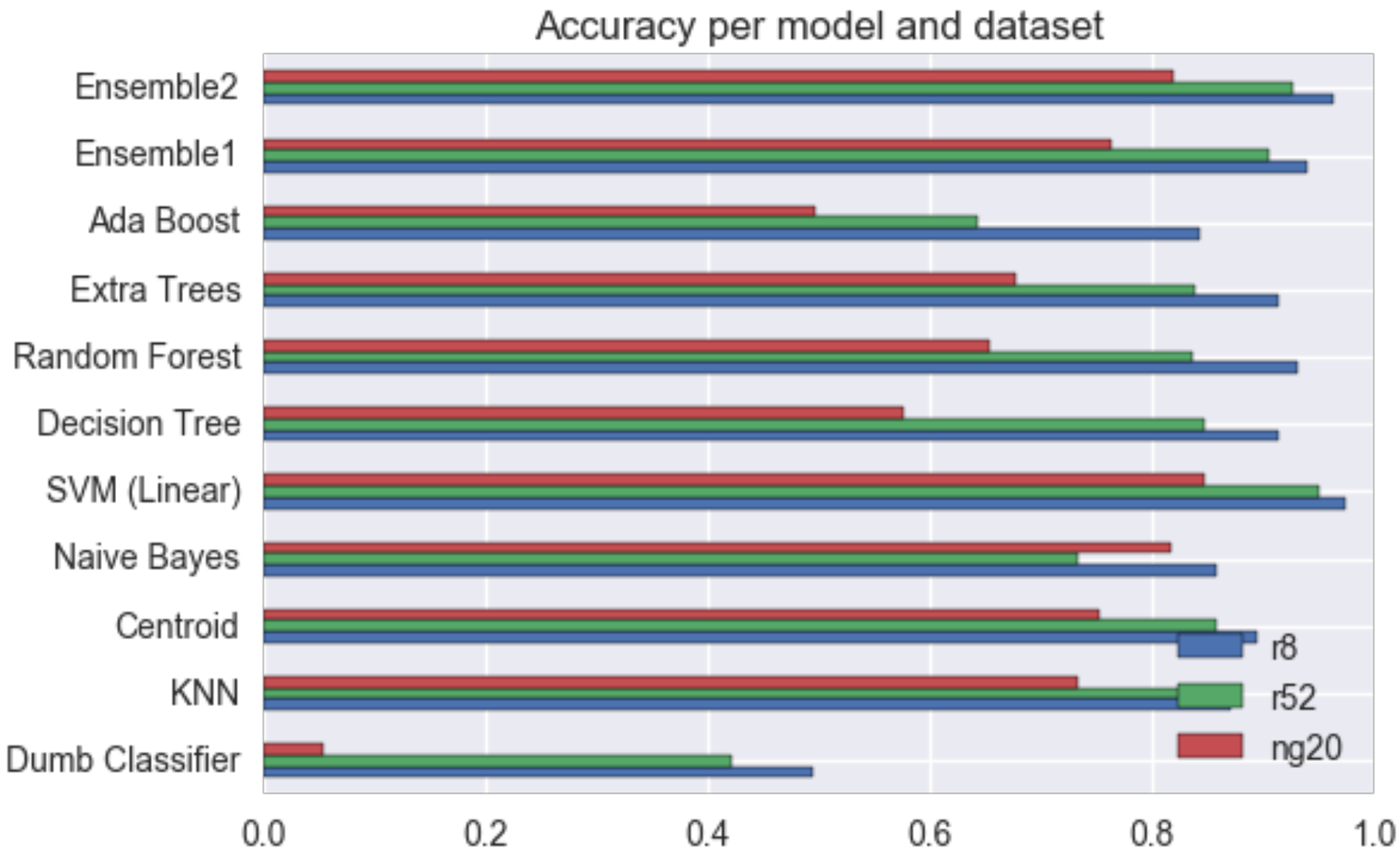
| Accuracy Values | | | | | |
|---|---|---|---|---|---|
| **Classification Method** | **R8** | **R52** | **20Ng** | **Cade12** | **WebKb** |
| Dumb classifier | 0.4947 | 0.4217 | 0.0530 | 0.2083 | 0.3897 |
| Vector Method | 0.7889 | 0.7687 | 0.7240 | 0.4142 | 0.6447 |
| kNN (k = 10) | 0.8524 | 0.8322 | 0.7593 | 0.5120 | 0.7256 |
| Centroid (Normalized Sum) | 0.9356 | 0.8717 | 0.7885 | 0.5148 | 0.8266 |
| Naive Bayes | 0.9607 | 0.8692 | 0.8103 | 0.5727 | 0.8352 |
| SVM (Linear Kernel) | 0.9698 | 0.9377 | 0.8284 | 0.5284 | 0.8582 |



# Apply ML models
## (use LSI/LSA/SVD to improve results)



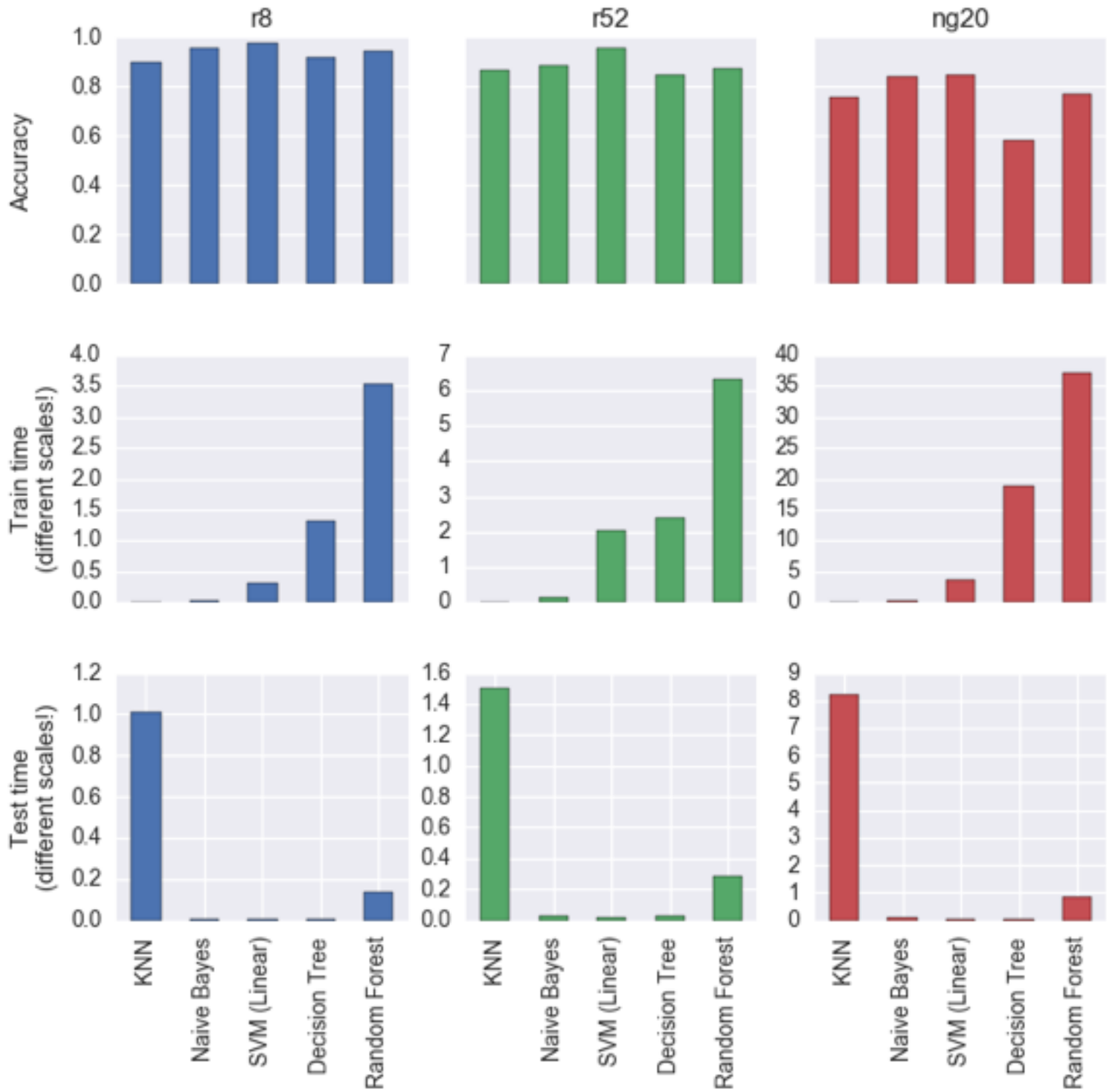
# Apply ML models
## (use more RT/Boost/Ensemble)



# Apply ML models
## (optimise parameters — GridSearchCV)

# Apply ML models

## (compare results and time)



# Summary

- **SVMs (still) rock!**
- Future work:
  - More ML models
  - More (and bigger) datasets
  - Different feature selection
  - Other success measures
  - AWS to parallelise tasks

# Thank you!

- Questions?