Text Mining and Search

FINAL PROJECT REPORT

Single-Label, Multi-Class text Classification

- The 20Newsgroups dataset - Overview
- The 20Newsgroups dataset - Content Masking
- Before Supervised Learning - Preprocessing
- Implementation - Filtering & Preprocessing
- Implementation - Representation
- Experimenting - Running the System
- Results - Observations

The 20Newsgroups dataset

- Industry-standard dataset
- 'bydate' version already split
- 20 different classes
- Each message belongs to just one

Newsgroup	Messages	Total
alt.atheism	480	480
comp.graphics	584	2936
comp.os.ms-windows.misc	591	
comp.sys.ibm.pc.hardware	590	
comp.sys.mac.hardware	578	
comp.windows.x	593	
misc.forsale	585	585
rec.autos	594	2389
rec.motorcycles	598	
rec.sport.baseball	597	
rec.sport.hockey	600	
sci.crypt	595	2373
sci.electronics	591	
sci.med	594	
sci.space	593	
soc.religion.christian	599	599
talk.politics.guns	546	1952
talk.politics.mideast	564	
talk.politics.misc	465	
talk.religion.misc	377	

The 20Newsgroups dataset – Content Filtering

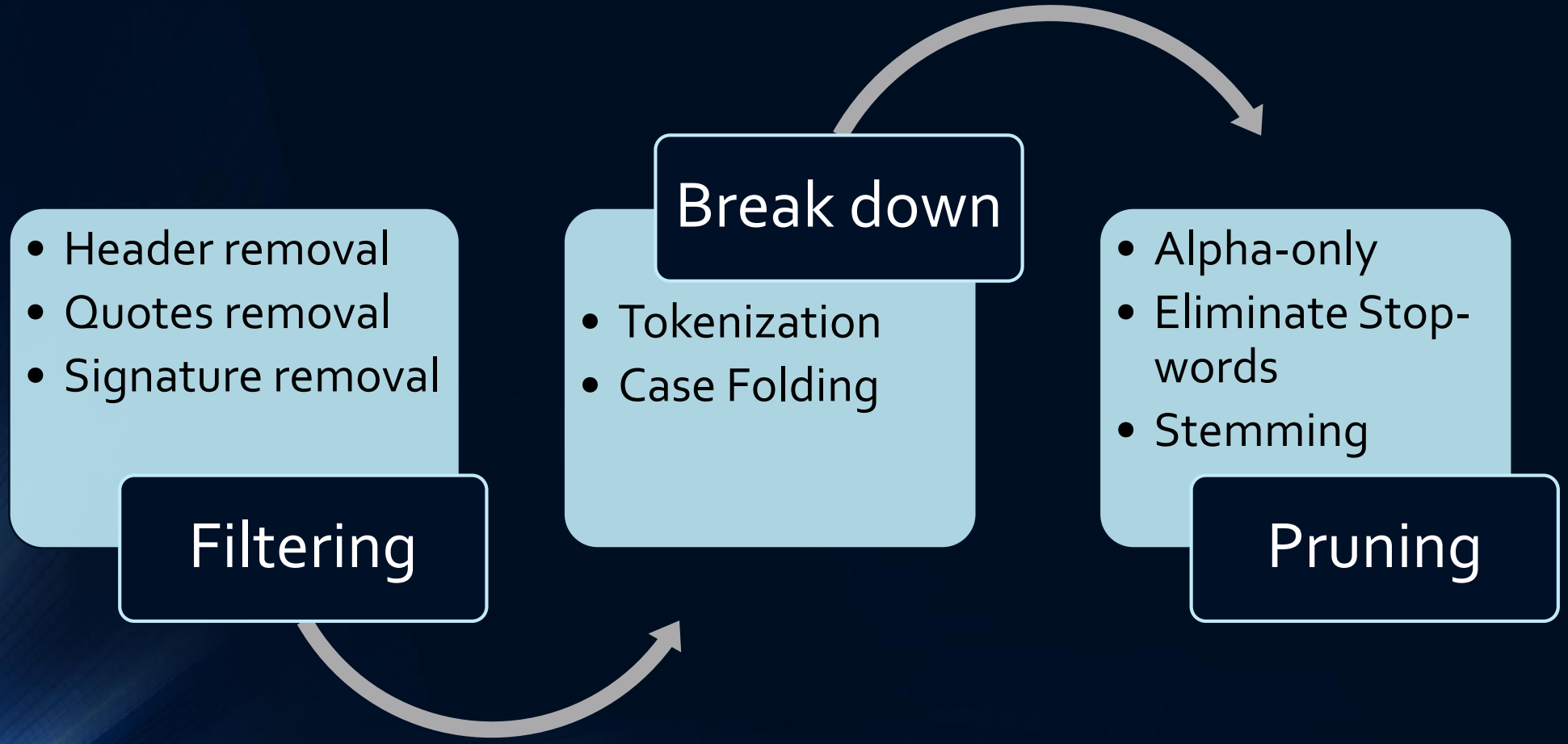
From: dmuntz@quip.eecs.umich.edu (Dan Muntz)
Subject: Re: new encryption
Organization: University of Michigan EECS Dept., Ann Arbor
Lines: 13

In article <strnlghtC5wC3z.Erw@netcom.com> strnlght@netcom.com (David Sternlight) writes:
>psionic@wam.umd.edu, whose parenthesized name is either an unfortunate
>coincidence or casts serious doubt on his bona fides, posts a message in
>which he seems willing to take the word of a private firm about which he
>knows little that their new encryption algorithm is secure and contains no
>trapdoors, while seemingly distrusting that of the government about clipper.

Will someone please post the David Sternlight FAQ to alt.privacy.clipper before
someone unfamiliar with him takes him seriously and starts yet another
flame fest?

-Dan

Before Supervised Learning - Preprocessing



The 20Newsgroups dataset – Content Filtering

From: dmuntz@quip.eecs.umich.edu (Dan Muntz)
Subject: Re: new encryption
Organization: University of Michigan EECS Dept., Ann Arbor
Lines: 13

In article <strnlghtC5wC3z.Erw@netcom.com> strnlght@netcom.com (David Sternlight) writes:
>psionic@wam.umd.edu, whose parenthesized name is either an unfortunate
>coincidence or casts serious doubt on his bona fides, posts a message in
>which he seems willing to take the word of a private firm about which he
>knows little that their new encryption algorithm is secure and contains no
>trapdoors, while seemingly distrusting that of the government about clipper.

Will someone please post the David Sternlight FAQ to alt.privacy.clipper before
someone unfamiliar with him takes him seriously and starts yet another
flame fest?

-Dan

The 20Newsgroups dataset – Content Filtering

From: dmuntz@quip.eecs.umich.edu (Dan Muntz)
Subject: Re: new encryption
Organization: University of Michigan EECS Dept., Ann Arbor
Lines: 13

In article <strnlghtC5wC3z.Erw@netcom.com> strnlght@netcom.com (David Sternlight) writes:
>psionic@wam.umd.edu, whose parenthesized name is either an unfortunate
>coincidence or casts serious doubt on his bona fides, posts a message in
>which he seems willing to take the word of a private firm about which he
>knows little that their new encryption algorithm is secure and contains no
>trapdoors, while seemingly distrusting that of the government about clipper.

Will someone please post the David Sternlight FAQ to alt.privacy.clipper before
someone unfamiliar with him takes him seriously and starts yet another
flame fest?

-Dan

The 20Newsgroups dataset – Content Filtering

From: dmuntz@quip.eecs.umich.edu (Dan Muntz)
Subject: Re: new encryption
Organization: University of Michigan EECS Dept., Ann Arbor
Lines: 13

In article <strnlghtC5wC3z.Erw@netcom.com> strnlght@netcom.com (David Sternlight) writes:
>psionic@wam.umd.edu, whose parenthesized name is either an unfortunate
>coincidence or casts serious doubt on his bona fides, posts a message in
>which he seems willing to take the word of a private firm about which he
>knows little that their new encryption algorithm is secure and contains no
>trapdoors, while seemingly distrusting that of the government about clipper.

Will someone please post the David Sternlight FAQ to alt.privacy.clipper before
someone unfamiliar with him takes him seriously and starts yet another
flame fest?

-Dan

The 20Newsgroups dataset – Preprocessing

will, someone, please, post, the, david, sternlight, faq, to, alt.privacy.clipper, before, someone, unfamiliar, with, him, takes, him, seriously, and, starts, yet, another, flame, fest

The 20Newsgroups dataset – Preprocessing

~~will~~, someone, ~~please~~, post, ~~the~~, david, sternlight, faq, ~~to~~, alt.privacy.clipper, ~~before~~, ~~someone~~, unfamiliar, ~~with~~, ~~him~~, ~~take~~, ~~him~~, seriously, ~~and~~, starts, ~~yet~~, another, flame, fest

Implementation

Filtering and Preprocessing

- Preprocessing implemented
  - in Python 3
  - Using the NLTK library
  - Masking using custom functions
  - Tokenization using Treebank and Punkt algorithms
  - Stemming using the Porter algorithm
  - Stop-word filtering using NLTK's English list
- Filtering functions
  - Custom-written regular expressions for each section
  - Selective (on-off) filtering

Implementation

Representation

- From a list of terms to document representation
  - Chosen representation: Document-Term matrix with Tf-Idf weights
  - Minimize sparsity by imposing a minimum term frequency
  - Minimize noise from overrepresented terms by imposing a maximum term frequency
- Choosing the right thresholds through parameter sweep
  - Lower limit has strong impact on sparsity and freq.
  - Upper limit is negligible as no overly frequent terms

Implementation

Classification models

- Four different classification models used
  - Multinomial Naïve Bayes
  - Support Vector Machine
  - K Nearest Neighbours
  - MultiLayer Perceptron
- Classification model implementations
  - Implemented by the **`scikit-learn`** library
  - Standard interfaces, consistent for all models
- Evaluating the Classification Models
  - Micro-average accuracy over all classes
  - Training time for the specific model

Experimenting

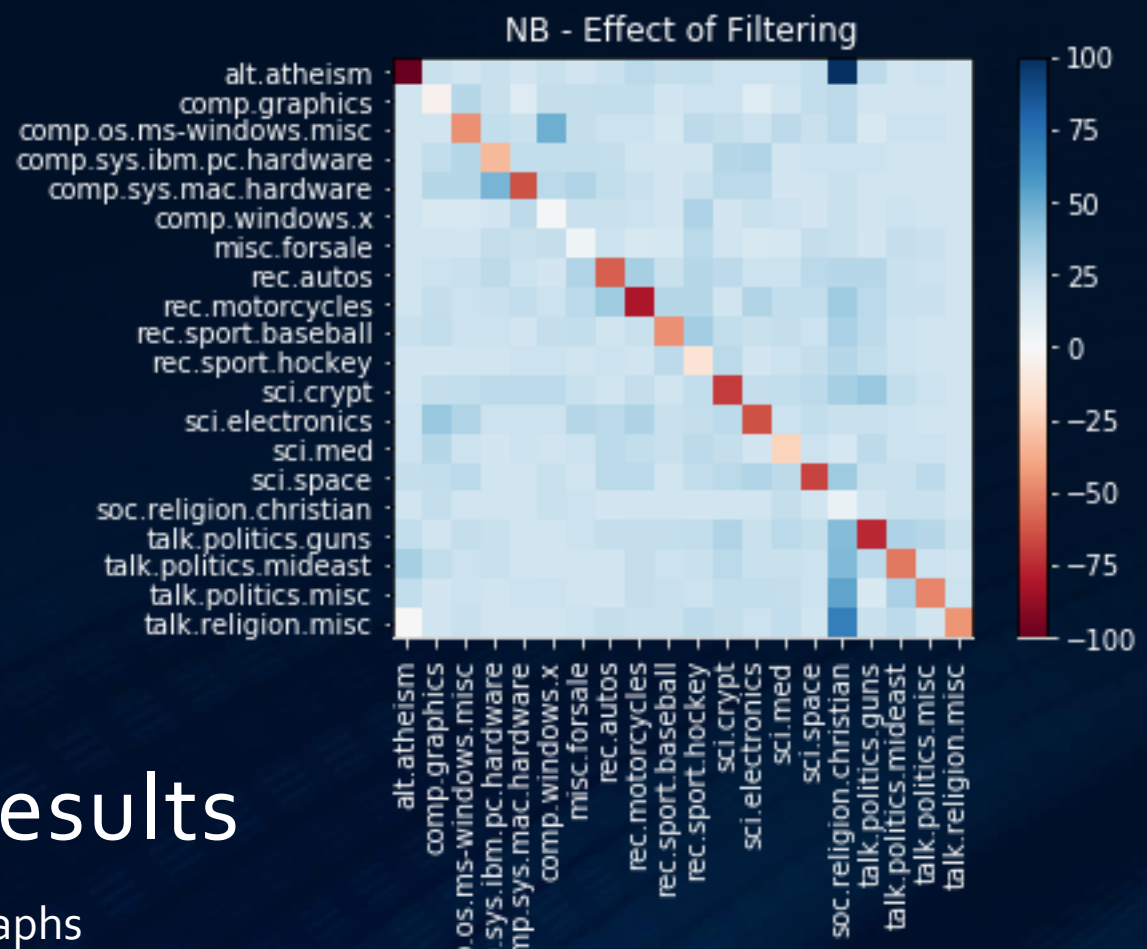
Running the system

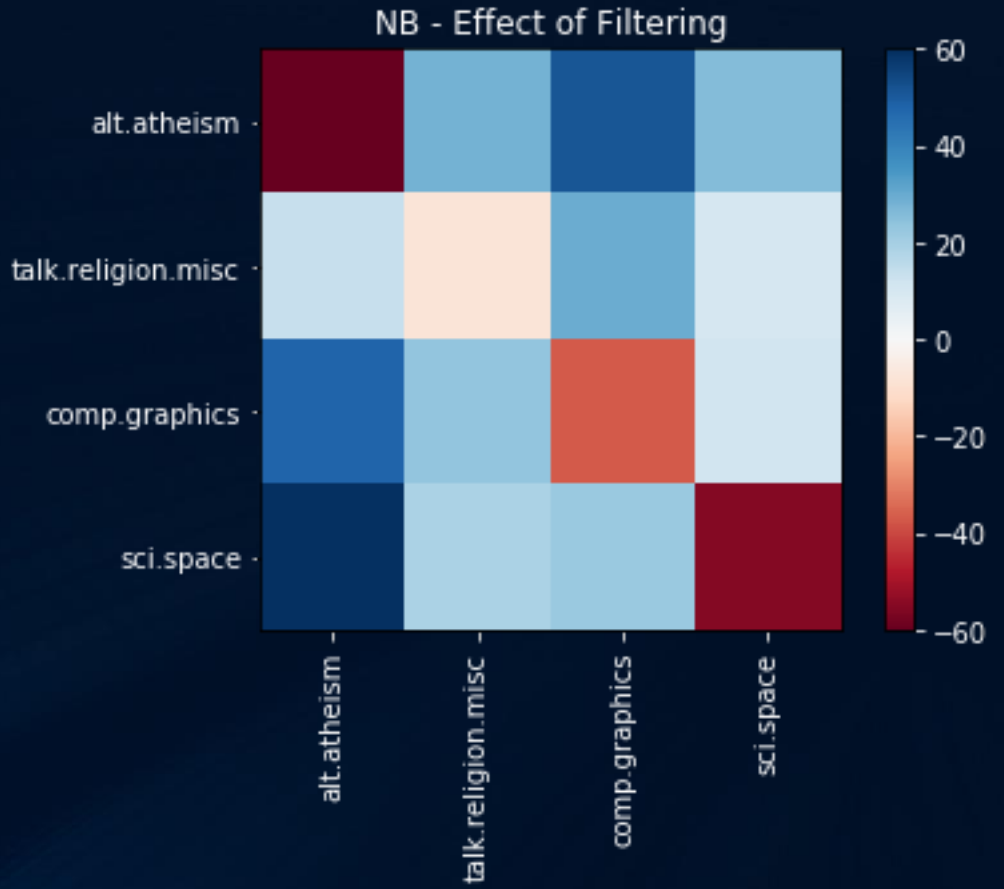
- Two separate experiments:
  - Classify messages over all 20 classes
  - Classify messages over 4 select classes
    - alt.atheism
    - talk.religion.misc
    - comp.graphics
    - sci.space
- For each combination of masking:
  - Run full pre-processing pipeline
  - Execute each classification model
  - Evaluate results for each model
  - Store evaluation results to file

Results

Observations

- Filtering has a negative effect on accuracy
  - Consistent with literature and expectations
  - Headers contain strongly identifying components
  - author, e-mail address, etc.
  - Citations can mention original poster's name and mail
  - Signatures tend to be almost unique for each user





Results

Graphs

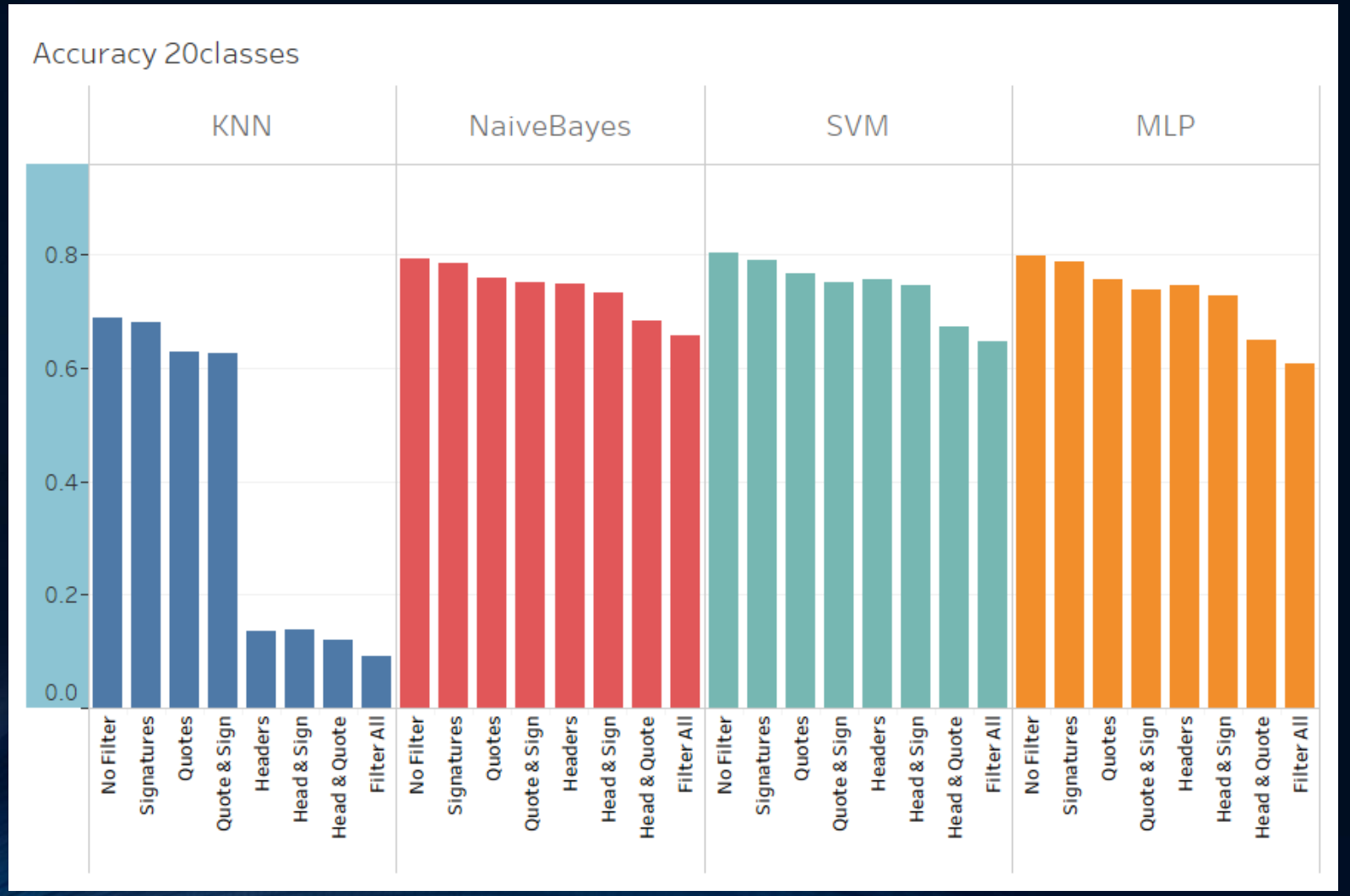
- Matrices showing the difference between the confusion matrix for «no filter» classification and «full-filter» classification for both 20-class and 4-class scenarios.

Finally

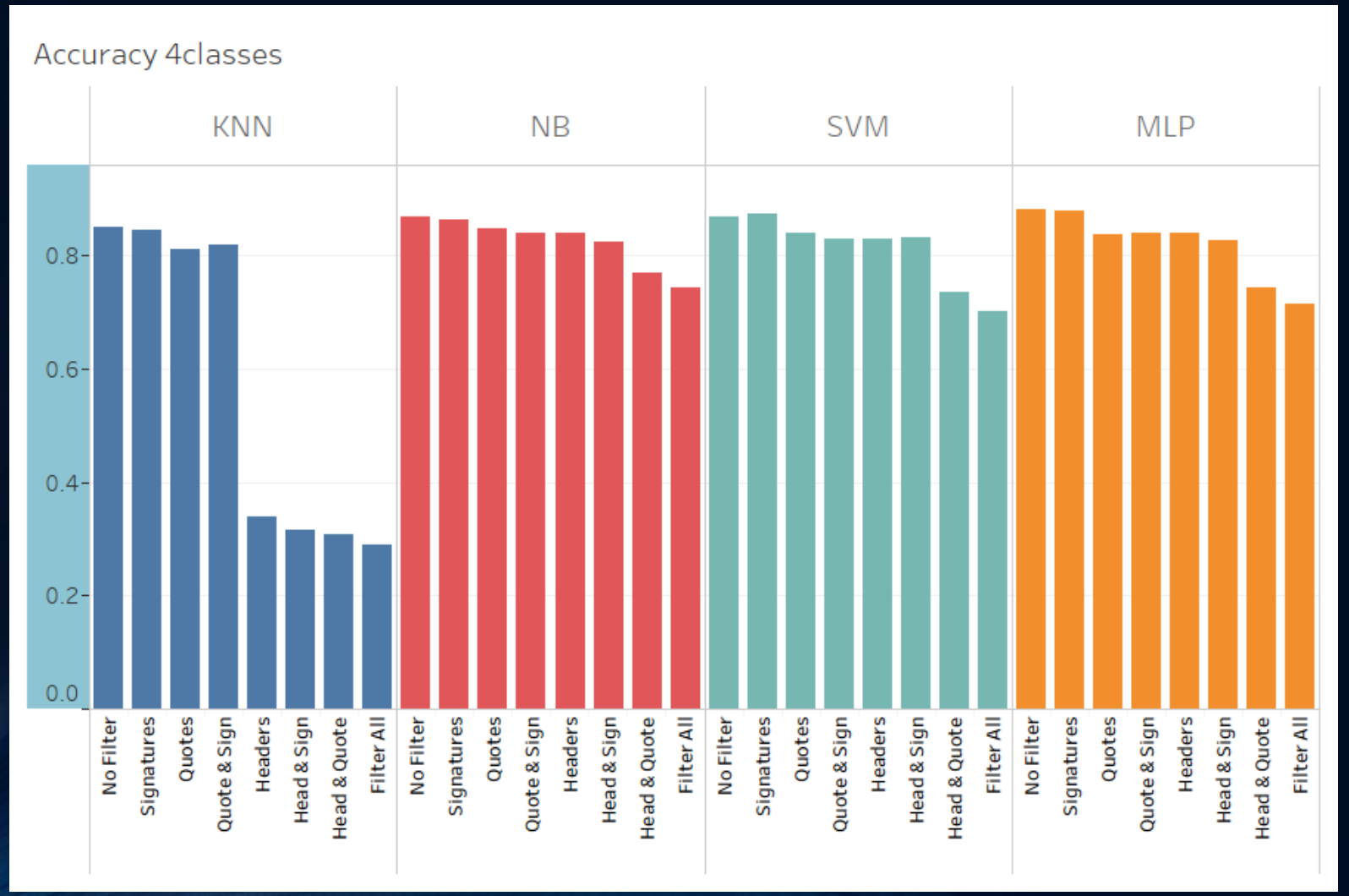
Observations and Conclusion

- Comparing out-of-the-box models: Naïve Bayes algorithm retains first place over our dataset and experiment space.
- Further analysis can be conducted by tuning hyperparameter configurations
- Filtering parts of the document has a strong impact on classification tasks for this dataset and should be taken into account when developing online systems that may rely on metadata.
- Reducing scope from 20 to 4 classes
  - improves performance for all classifiers
  - improves accuracy for all classifiers
  - still shows NB having the overall advantage

Results - Accuracy over all 20 classes



Results - Accuracy over 4 select classes



Thank you

Paolo Nicoli 833311

Andrea Sassanelli 835119