Talk of Norway

Martin G. Søyland
Emanuele Lapponi

Who?

A cross faculty collaboration between two PhD fellows

- Martin is a political scientist using empirical methods in his research
- Interested in adopting NLP techniques in his current work

- Emanuele is a computer scientist investigating the use of NLP outside of the field
- Interested in applying NLP techniques to real-world problems

What?

- 250373 Speeches from the Norwegian Parliament, 1998 to 2016
- A rich set of 83 metadata variables describing speaker, party, government (and more!) at the time the speech was uttered
- Speeches annotated with sentence and token boundaries, lemmas, parts-of-speech and morphological features

How?

- Raw speeches and metadata pulled from the Storting API (with a head start, thanks to holderdeord.no)
- Additional metadata scraped from the storting website
- More information from Søyland, 2017 (forthcoming!)
- Automatic language identification and morphological analysis done with langid.py and OBT as implemented in the Language Analysis Portal

Why?

- So you don't have to!
- Currently the core object of study in Martin and Emanuele's PhD work
- Recent years have seen increasing interest in automatic analysis of parliamentary proceedings, e.g. Høyland et al., 2014, and Bäck and Debus, 2016

What impossibly complicated format did you use?

- A .csv containing one raw speech with associated metadata per row

```
"tale000025","MAA","MAA","Marit","Arnstad","Marit Arnstad","1997-10-01","2001-09-30","Vararepresentant","Nord-Trøndelag"," 1","Sp","Senterpar
"tale000026","BEH","BEH","Bent","Hegna","Bent Hegna","1997-10-01","2001-09-30","Representant","Telemark"," 5","A","Arbeiderpartiet","Opposit
"tale000027","BYR","BYR","Bror Yngve","Rahm","Bror Yngve Rahm","1997-10-01","2001-09-30","Representant","Telemark"," 4","KrF","Kristelig Folk
"tale000028","KNA","KNA","Kjellaug","Nakkim","Kjellaug Nakkim","1997-10-01","2001-09-30","Representant","Østfold"," 5","H","Høyre","Oppositio
"tale000029","GKV","GKV","Gunnar","Kvassheim","Gunnar Kvassheim","1997-10-01","2001-09-30","Representant","Rogaland","11","V","Venstre","Cabi
"tale000030","GS","GS","Gunnar","Skaug","Gunnar Skaug","1997-10-01","2001-09-30","Representant","Østfold"," 1","A","Arbeiderpartiet","Opposit
"tale000031","OH","OH","Odd","Holten","Odd Holten","1997-10-01","2001-09-30","Representant","Østfold"," 4","KrF","Kristelig Folkeparti","Cabi
"tale000032","RF","RF","Ranveig","Frøiland","Ranveig Frøiland","1997-10-01","2001-09-30","Representant","Hordaland"," 4","A","Arbeiderpartiet
"tale000033","MAA","MAA","Marit","Arnstad","Marit Arnstad","1997-10-01","2001-09-30","Vararepresentant","Nord-Trøndelag"," 1","Sp","Senterpar
```

What impossibly complicated format did you use?

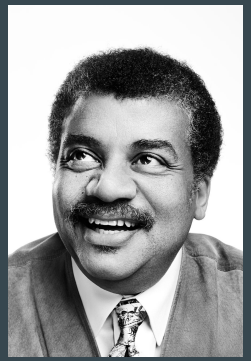
- A CoNLL-like format for linguistic annotations
- Filenames are synced with the “id” variable in the csv

```
1   Ærede   ære adj fl|<perf-part>|tr1
2   medrepresentanter   medrepresentant subst   appell|mask|ub|fl
3   !   $!  clb <<<|<utrop>|<<<

1   Tidligere   tidlig  adj komp
2   stortingsrepresentant   stortingsrepresentant   subst   appell|mask|ub|ent
3   Sjur    Sjur    subst   prop|mask|<*>
4   Lindebrække Lindebrække subst   prop|<*>
5   er  være    verb    pres|a5|pr1|pr2|<aux1/perf_part>
6   død død adj ub|m/f|ent|pos
7   –   –   symb    _
8   89  89  det fl|kvant
9   år  år  subst   appell|nøyt|ub|fl
10  gammel  gammel  adj ub|m/f|ent|pos
11  .   $.  clb <<<|<punkt>|<<<
```

Stats

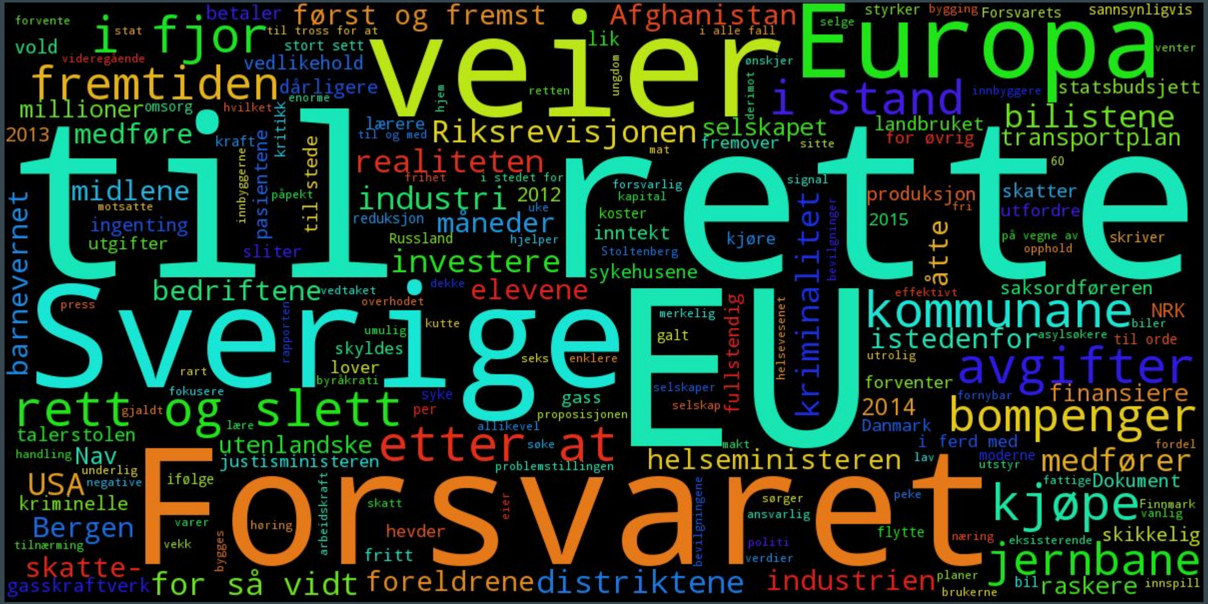
	#speeches	#tokens	%nno
President	72,646	2,525,733	0.7%
Ap	43,483	16,008,420	0.9%
H	32,945	11,481,762	0.2%
FrP	30,217	9,729,435	0.5%
SV	19,941	7,218,136	18%
KrF	19,720	6,653,088	19%
Sp	18,255	5,874,381	33%
V	11,579	3,830,095	0.8%
MDG	508	153,834	0.01%
Kp	492	128,709	0.06%
TF	409	97,001	0%
Independent	131	38,284	0%
Other	47	64,715	19%
Total	250373	63,803,593	19%



Science level:
blog post

Using the data, science level: blog post

- Using a well-informed, idf-based stop-word list
- Counting over surface forms



Using the data, science level: blog post

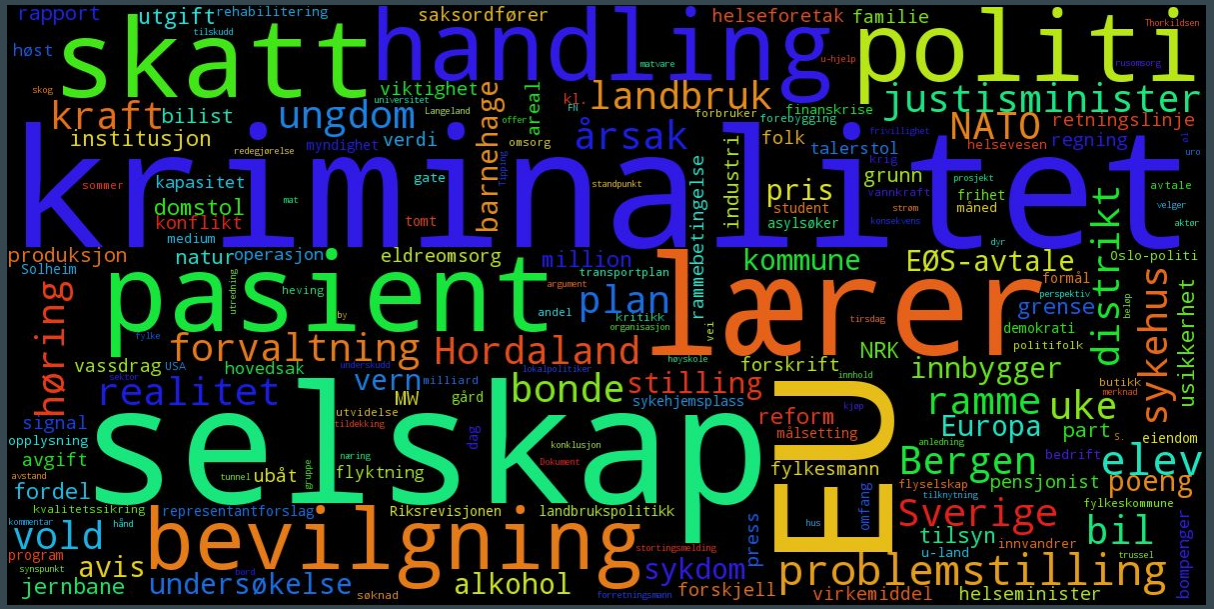
- Using a well-informed, idf-based stop-word list
- Counting over lemmas
- Filtering out everything but nouns

Using the data, science level: blog post

- Using a well-informed, idf-based stop-word list
- Counting over lemmas
- Filtering out everything but nouns

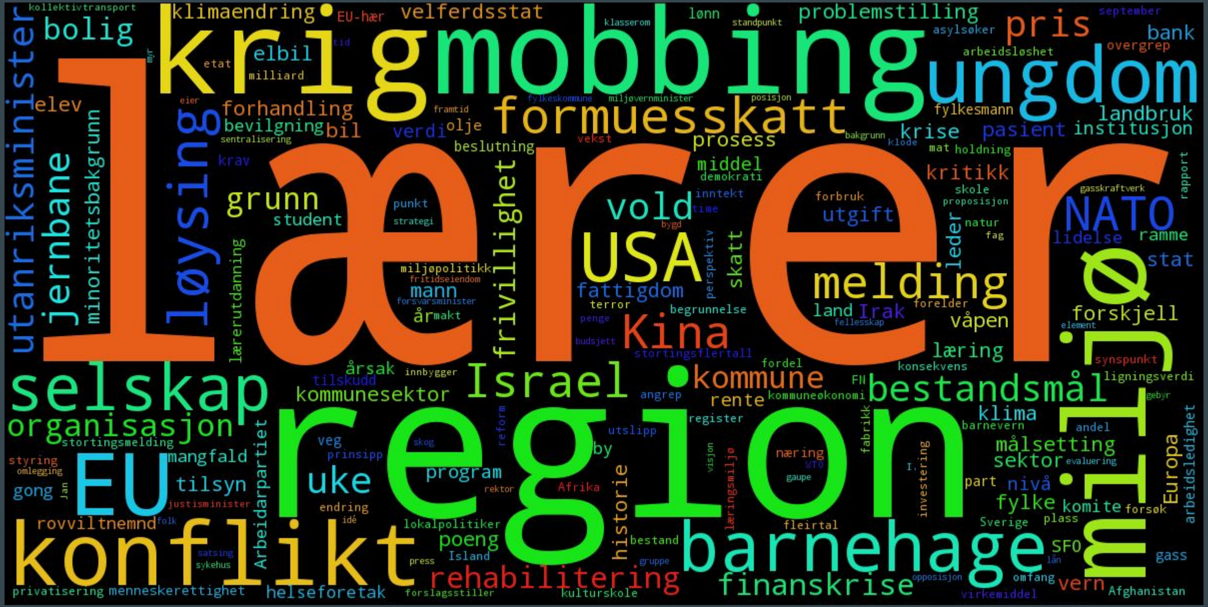
Using the data, science level: blog post

- Using a well-informed, idf-based stop-word list
- Counting over lemmas
- Filtering out everything but nouns
- Looking only at sessions led by the social committee



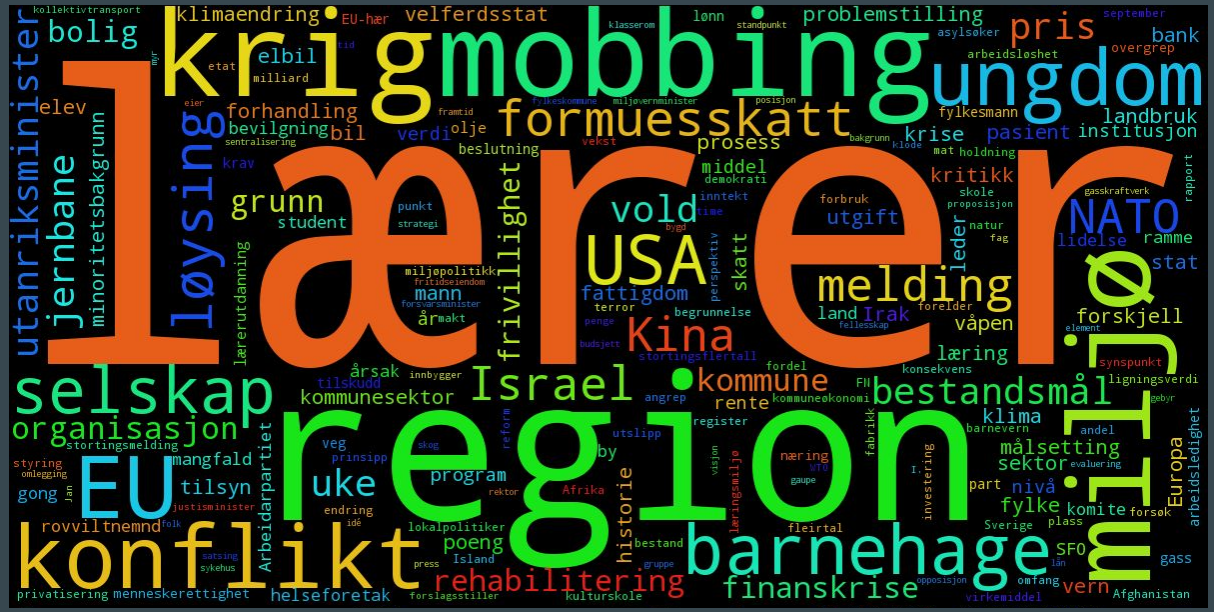
Using the data, science level: blog post

- Using a well-informed, idf-based stop-word list
- Counting over lemmas
- Filtering out everything but nouns
- Looking only at sessions led by the social committee



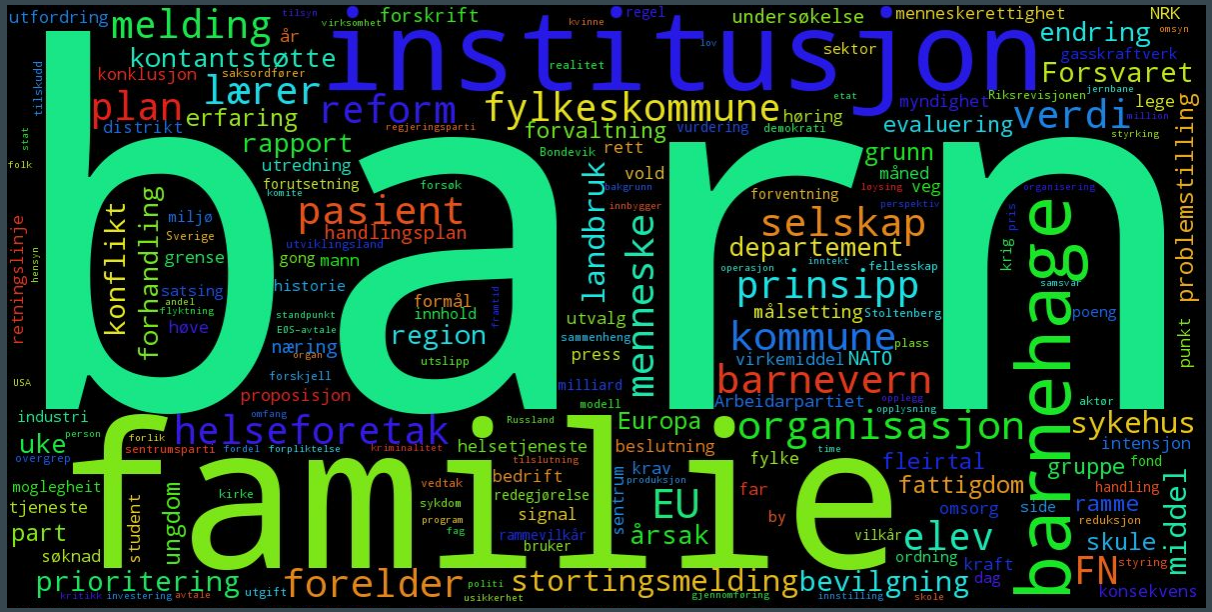
Using the data, science level: blog post

- Using a well-informed, idf-based stop-word list
- Counting over lemmas
- Filtering out everything but nouns
- Looking only at sessions led by the social committee



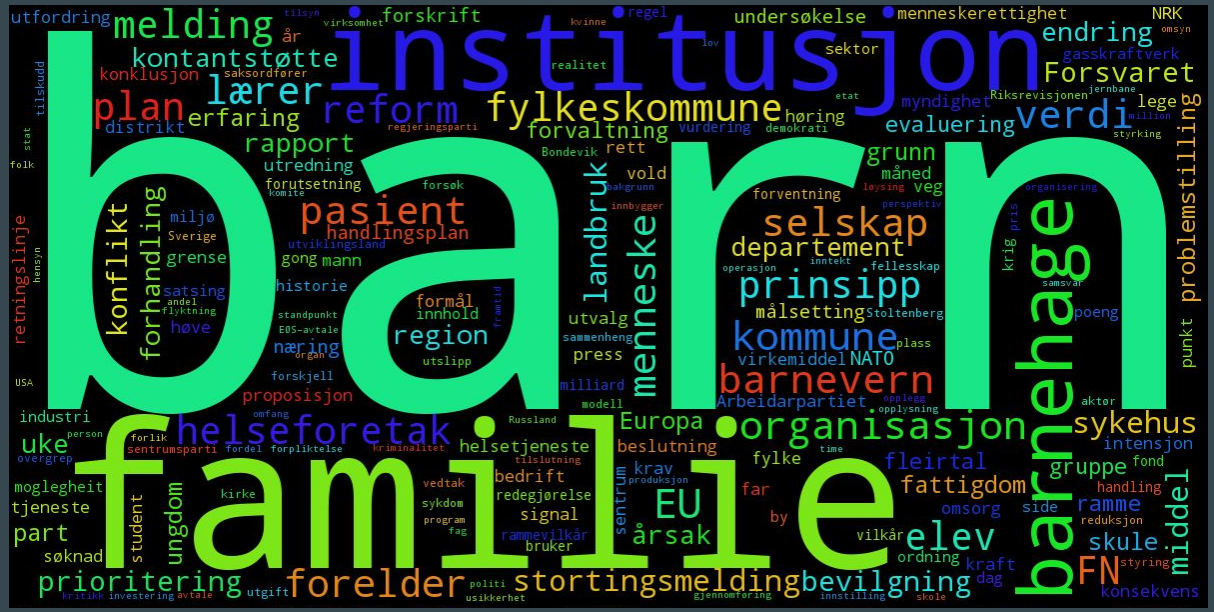
Using the data, science level: blog post

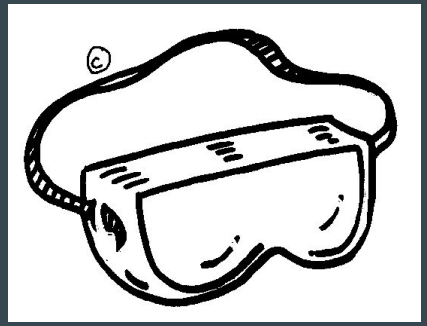
- Bonus round!



Using the data, science level: blog post

- Bonus round!

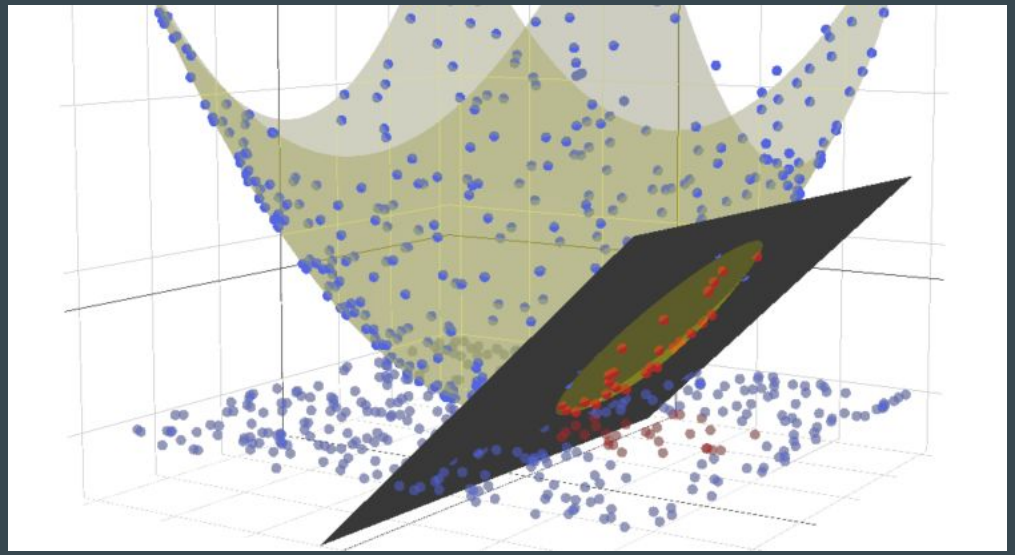




Classification Experiments

Classification experiments

- Using a subset of the data
  - Speeches with a party label
  - Parties in Storting 1998–2016
- Six fold cross–validation, each cabinet a fold
- Linear SVM (through sklearn)



Classification experiments

Classifier	feature-set	**P**	**R**	**F_1**	**Accuracy**	**Error Reduction**
	Baseline	0.035	0.142	0.056	0.248	—
1	Token	0.425	0.400	0.412	0.432	0.244
2	+ Lemma	0.432	0.405	0.418	0.437	0.009
3	+ Ngrams	0.549	0.487	0.516	0.518	**0.143**
4	+ PoS	0.551	0.491	0.520	0.523	0.009
5	+ Meta	**0.570**	**0.511**	**0.538**	**0.539**	0.035

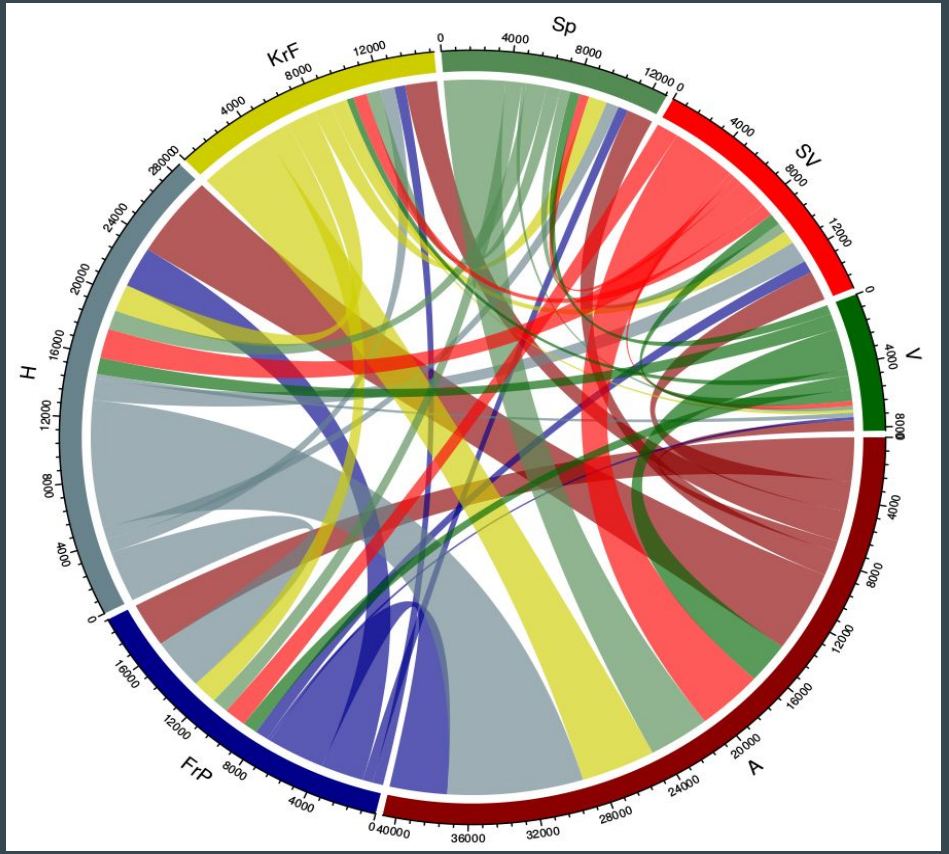
Classification experiments

- Compares favorably to multi-party classification experiments - easier to classify Norwegian parties than EU
- Evenly spread results: classifier performance not driven by class size distribution
- Except FrP much, much easier to classify

	P	**R**	**F_1**
SV	0.578	0.490	0.531
A	0.471	0.624	0.537
Sp	0.618	0.527	0.569
KrF	0.578	0.433	0.495
V	**0.637**	0.351	0.452
H	0.503	0.485	0.494
FrP	0.603	**0.665**	**0.632**
MACRO	0.570	0.511	0.538

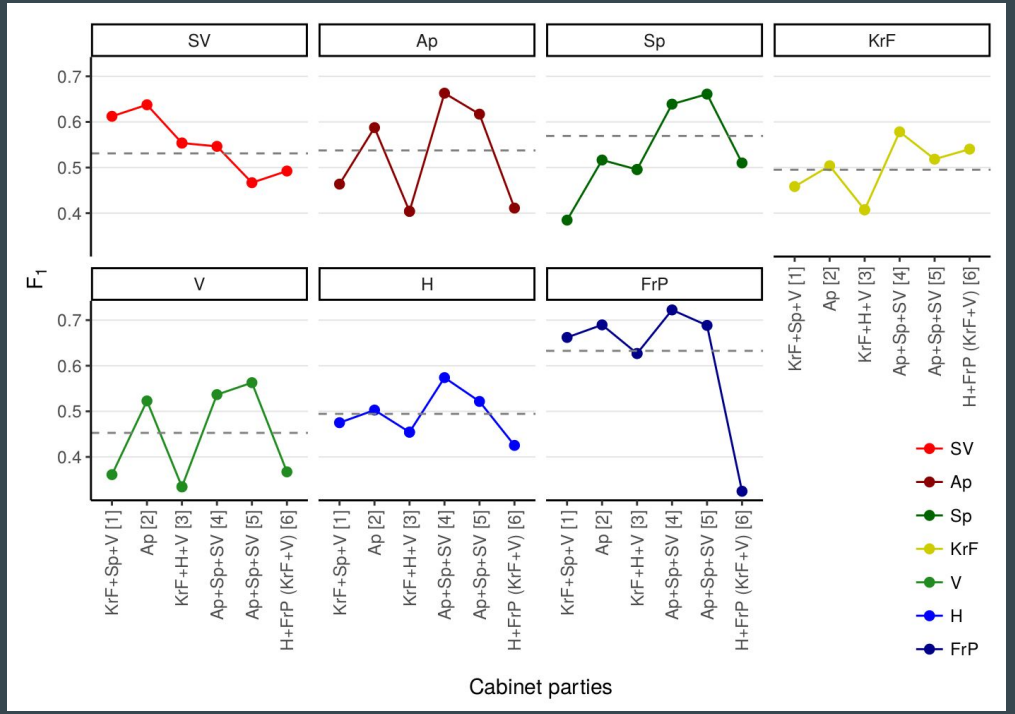
Classification experiments

- False Positives / False Negatives



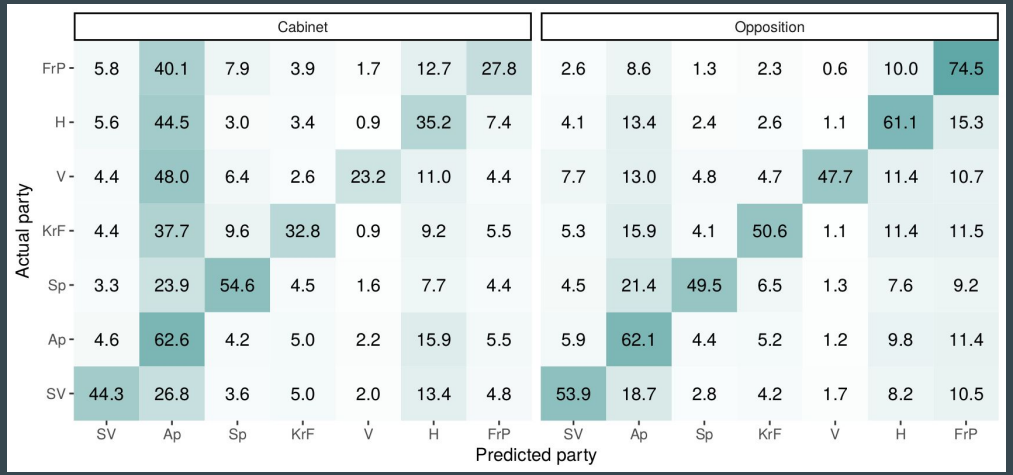
Classification experiments

- In general, easier to classify parties in opposition
- Except for Ap
- FrP classification disintegrates when in government



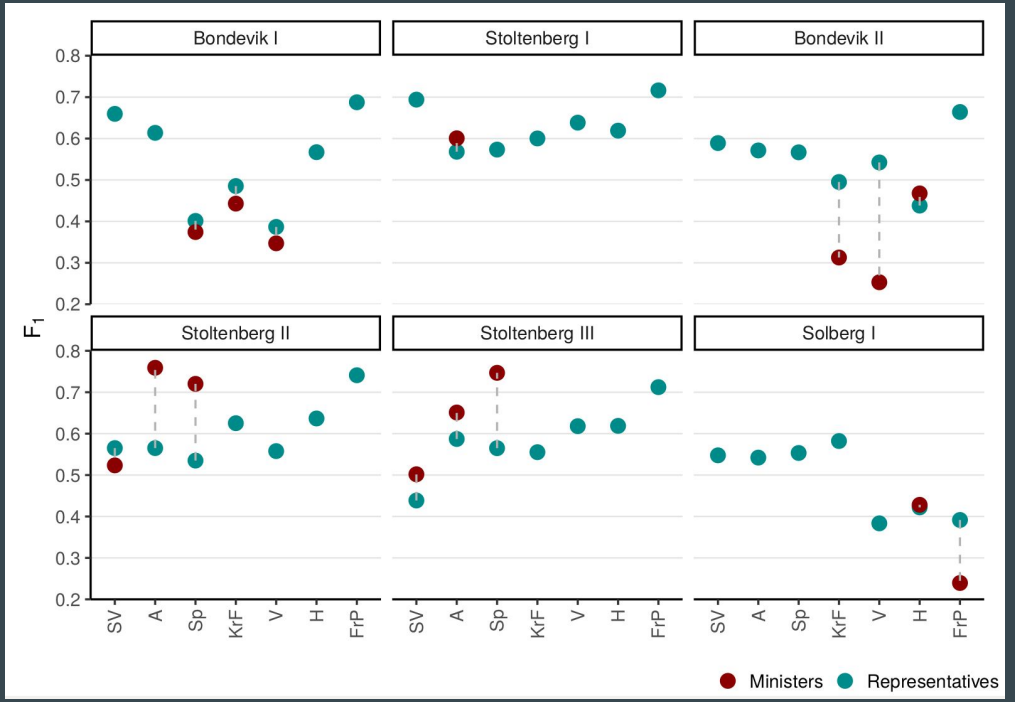
Classification experiments

- In cabinet, speakers confused with Ap and H
- Easier to maintain an ideological profile in opposition



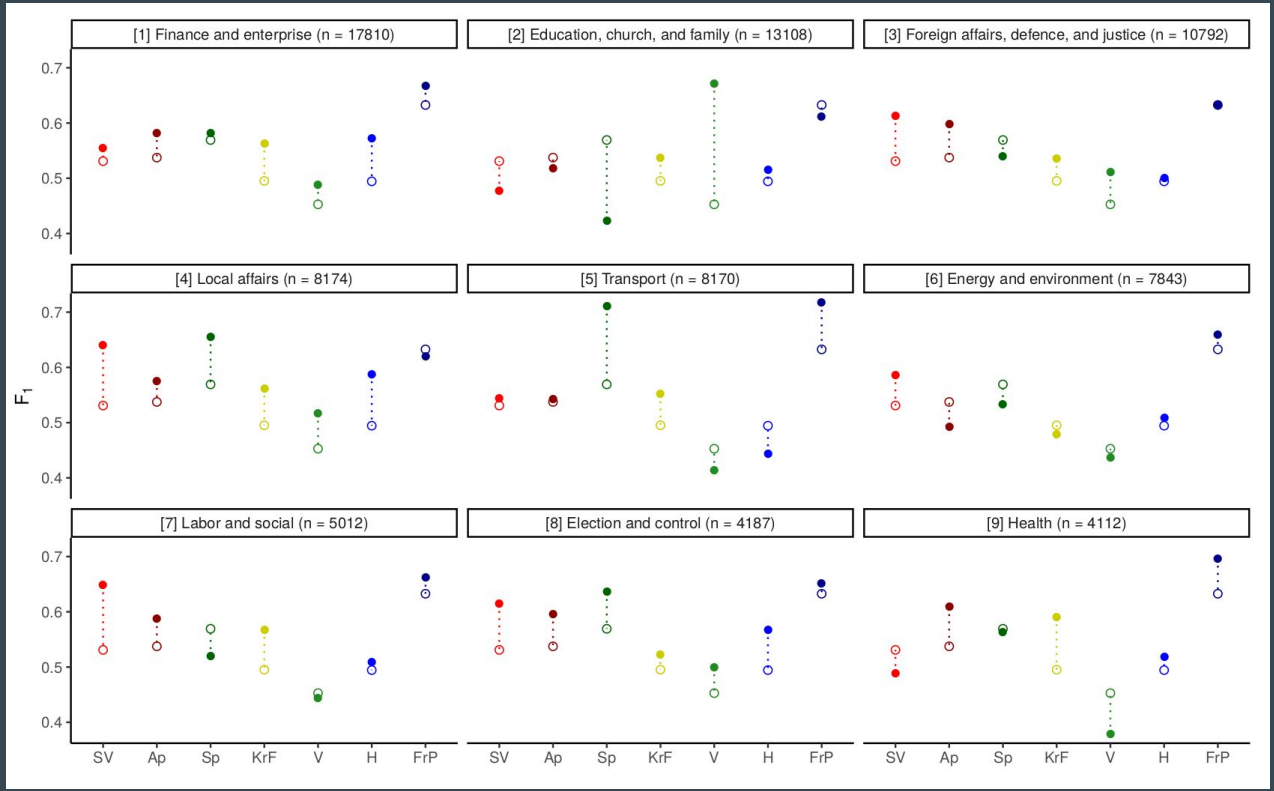
Classification experiments

- Minister classification is quite different from representative classification



Classification experiments

- Performance in topical subsets (committees)
- More often than not, drops/increases in performance indicate that position is not the only driving force in the model



Conclusions

- The ToN corpus, useful for quantitative polsci research
- First results using classification performance as a quantitative measure for political analysis on the Norwegian Parliament

https://github.com/ltgoslo/talk-of-norway

https://www.mn.uio.no/ifi/english/research/projects/ton

Thank you!