Text Mining and Validation of Results

What is good enough?

Janneke van der Zwaan

netherlands eScience center

by SURF & NWO

Validation study

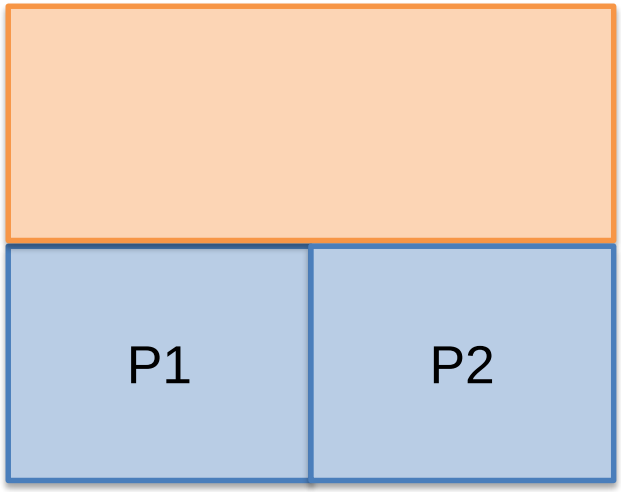
1. How do opinions of political parties change over time?
2. Can we validate topics and associated opinions?

"Validating Cross-Perspective Topic Modeling for Extracting Political Parties' Positions from Parliamentary Proceedings" by Van der Zwaan, Marx, and Kamps, 2016

Example Topic and Opinions

Topic 33		Opinion VVD		Opinion GroenLinks	
asylum seeker	0.0538	illegal	0.0566	careful	0.0655
foreigner	0.0342	criminal	0.0350	illegal	0.0561
shelter	0.0263	discretionary	0.0255	strict	0.0306
return	0.0239	incorrect	0.0249	punishable	0.0213
procedure	0.0238	non-Western	0.0243	underprivileged	0.0193
residence permit	0.0218	strict	0.0222	righteous	0.0188
pardon	0.0203	unlawful	0.0218	safe	0.0184
origin	0.0203	minor	0.0213	enormously	0.0169
stay	0.0189	vulnerable	0.0209	false	0.0164
illegal	0.0188	punishable	0.0206	fine	0.0152

Cross-Perspective Topic Modeling



Nouns → topics

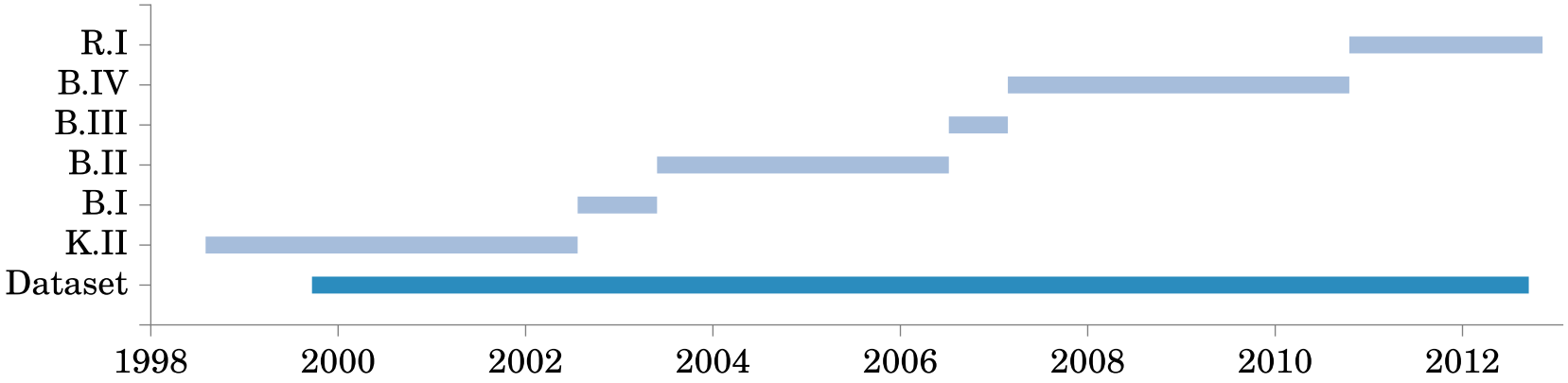
Adjectives → opinions

Perspectives
(political parties)

"Mining Contrastive Opinions on Political Text using Cross-Perspective Topic Model" by Yang, Si, Somasundaram, and Yu, 2012

Data

- **Dutch parliamentary proceedings**
  - **Kok II – Rutte I**
    - **September 21, 1999 - September 11, 2012**
  - **House of representatives + senate**
  - **20,594 documents**



Topic models

- **Perspectives**
  - **parties (11)**
  - **parties through time (59)**

- **100 topics**
- **Software: https://github.com/nlesc/cptm**

Example Topic and Opinions

Topic 33		Opinion VVD		Opinion GroenLinks	
asylum seeker	0.0538	illegal	0.0566	careful	0.0655
foreigner	0.0342	criminal	0.0350	illegal	0.0561
shelter	0.0263	discretionary	0.0255	strict	0.0306
return	0.0239	incorrect	0.0249	punishable	0.0213
procedure	0.0238	non-Western	0.0243	underprivileged	0.0193
residence permit	0.0218	strict	0.0222	righteous	0.0188
pardon	0.0203	unlawful	0.0218	safe	0.0184
origin	0.0203	minor	0.0213	enormously	0.0169
stay	0.0189	vulnerable	0.0209	false	0.0164
illegal	0.0188	punishable	0.0206	fine	0.0152

Types of validity

Type	Description
Face	The extent to which results appear to be valid.
Content	The extent to which a method for measuring a latent construct represents all of its facets.
Criterion	Correlation between a measure and other measures that reflect the same concept.
Construct	The extent to which a measure behaves as expected in a theoretical context.

Table 1: Types of validity (adapted from [9]).

[9] E. G. Carmines and R. A. Zeller, *Reliability and validity assessment*, Sage publications, 1979.

Validity

- **Topics**
  - **Are all relevant political subjects covered?**
  - **Can we map topics to these political subjects?**
- **Opinions**
  - **Are opinions representative of party manifestos?**
  - **Can use the opinions to rank parties from left to right?**

Opinion Validity

- **Given a party manifesto, whose opinion is expressed?**

$$\underset{i \in C}{\operatorname{argmax}}\, p(d|o^i)$$

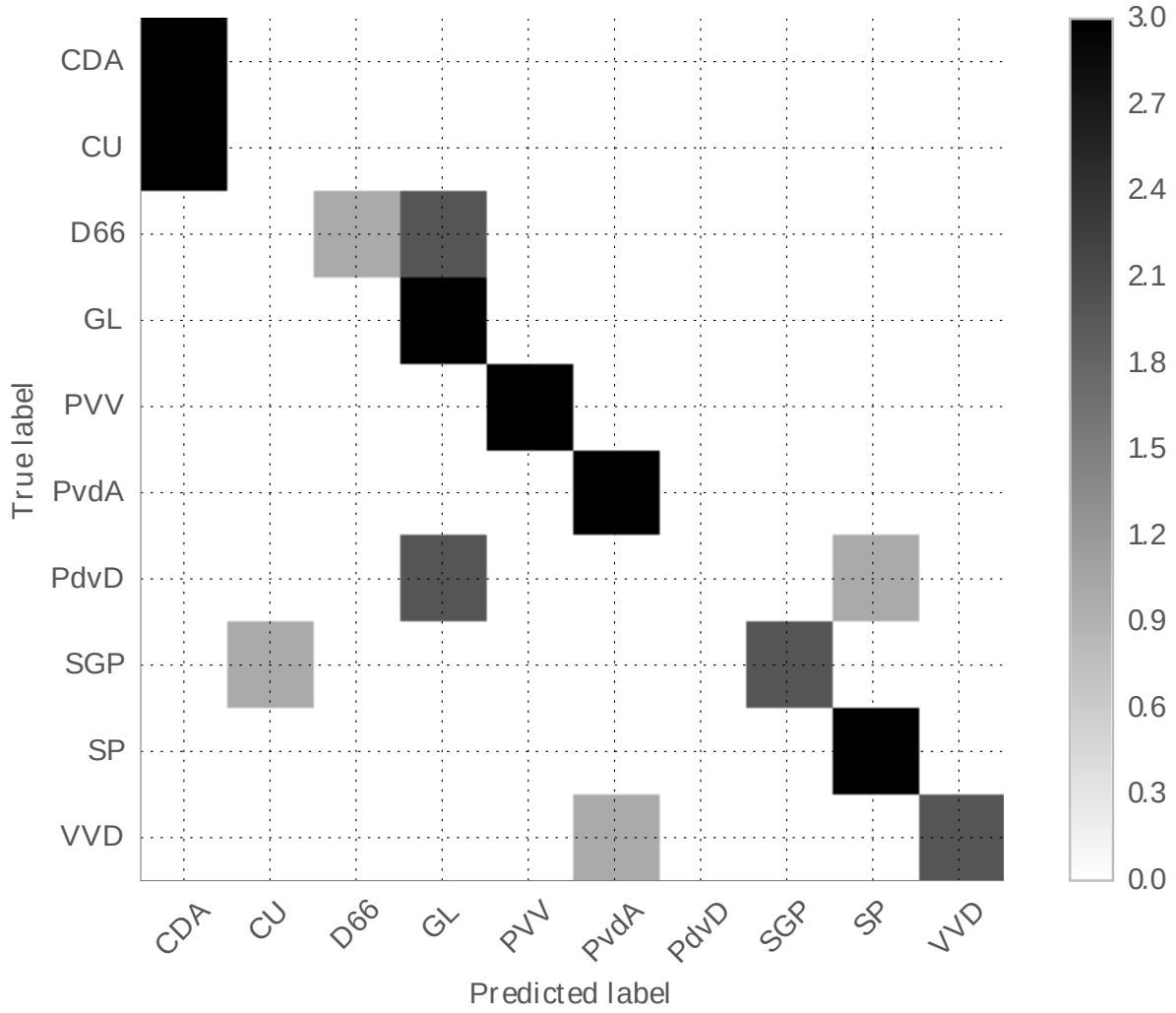
$$perplexity(d) = exp - \frac{log(p(\mathbf{o}))}{N_o}$$

$$p(\mathbf{o}) = \prod_{i=1}^{N_o} \sum_{k=1}^{K} p(o_i|z_i = k)p(z_i = k|d)$$

Party Manifestos

- **10 parties (all parties except LPF)**
- **Manifestos from 2006, 2010, and 2012**

Results



Accuracy: 0.667
(Random: 0.333)

Conclusion

- **Cross-perspective topic modeling on Dutch parliamentary proceedings**
  - **Opinions are representative of party manifestos**

- **We need more validation studies!**
- **Validation is doable!**