Measuring Rhetorical Similarity with Machine Learning

Nicolai Berk

University of Amsterdam

July 9, 2020

Relevance

Group similarity of interest for many areas in political science:

- Representation
  - Do certain (e.g. female/working class/PoC) MPs communicate differently?
  - How similar is legislation to interest group's demands?
  - Do MP's become more similar over time?
  - Are directly elected MPs communicating differently?
- Government formation
  - Which parties communicate more similar and are more likely to form a coalition?
- Party Competition
  - Did parties become more or less similar to each other?
  - Did parties fill political opportunity spaces opened by other parties?
  - Which MPs are likely to leave their party?
- Polarisation, ...

Current Methods

- Simple similarity measures: cosine, jaccard
  $\rightarrow$ no weighting of terms
- Scaling methods: WORDSCORE, WORDFISH (Laver, Benoit, and Garry 2003; Slapin and Proksch 2008)
  $\rightarrow$ no similarity measures, cannot measure similarity of a document to a collection of texts
- Peterson and Spirling 2018: Machine learning accuracy to measure polarisation $\rightarrow$ no estimates for single texts

$\rightarrow$ **Use ML to scale texts on 'groupness'**

A supervised ML measure of group similarity

1. Pre-process texts
2. Select well-performing classifier
3. Balance data, using SMOTE (Chawla et al. 2002)
4. BoW-classifier trained on full set
5. Predicted probabilities assigned to all texts
6. Train one classifier/time-period to 'control' for changing language

Application

- How similar are speakers and parties to the radical right?
- Parliamentary speeches from lower Chambers in AT, DE, NL (Rauh and Schwalbach 2020)
- estimates of similarity to radical right party obtained using Python's `scikit-learn`
- Logistic regression on unstemmed text, tfidf-weighted

Validation

Three approaches

- Content of RR language
- Government formations with RR participation (AT & NL)
- Speaker estimates of party exits (DE & NL)

Best predictor Words

+0.796	SPÖ
+0.688	geehrten
+0.677	einmal
+0.616	eben
+0.610	Patienten
+0.604	Freiheitliche
+0.590	Hohes
+0.550	Kollegin
+0.535	Jahren
+0.533	Bevölkerung
+0.477	Auch
+0.460	Bundesminister
+0.458	Rednerpult
+0.448	Zur
+0.447	Wien
+0.443	geehrte
+0.443	Sozialdemokratie
... 6818 more positive ...	
... 7981 more negative ...	
-0.445	Kurz
-0.448	halte
-0.456	Finanzminister
-0.517	Entschließungsantrag
-0.529	Vielen
-0.540	FPÖ
-0.552	Bürgerinnen
-0.589	Menschen
-0.619	Liebe
-0.624	Ministerin
-0.727	Bundesministerin
-0.758	Kolleginnen
-0.841	ÖVP

(a) Austria

+5.317	Bürger
+5.209	Merkel
+5.037	deutschen
+4.758	AfD
+4.564	Deutschland
+4.491	Altparteien
+4.149	hier
+4.066	Regierung
+3.879	Vielen
+3.843	bedanke
+3.575	Damen
+3.501	Danke
+3.468	Herren
+3.447	Arbeitnehmer
... 11658 more positive ...	
... 15143 more negative ...	
-3.385	Soldatinnen
-3.425	Staatssekretär
-3.448	Herzlichen
-3.463	wichtig
-3.792	Koalitionsvertrag
-3.893	letzten
-3.943	Deswegen
-4.160	Linke
-4.327	Kollegen
-4.486	Deshalb
-4.504	brauchen
-4.567	Bürgerinnen
-4.604	Demokraten
-4.618	sagen
-4.884	finde
-9.485	Kolleginnen

(b) Germany

+10.354	immigratie
+9.579	enzovoorts
+9.199	islam
+9.081	PVV
+8.541	islamitische
+7.271	Brussel
+7.216	Mogelijk
+6.936	immigratiebeleid
+6.902	illegalen
+6.882	137d
+6.562	tekenen
+6.523	dierenpolitie
+6.414	huishoudelijke
+6.357	gigantisch
+6.012	kortingen
+5.965	bene
+5.863	statushouders
+5.856	hemelsnaam
... 13823 more positive ...	
... 25807 more negative ...	
-5.966	Kops
-6.149	afweging
-6.164	Tegelijkertijd
-6.209	afspraken
-6.311	zorgvuldig
-6.387	ontraad
-6.962	heer
-6.977	kijken
-7.071	Agema
-7.877	helder
-7.888	overwegende
-8.770	Baudet

(c) Netherlands

Best predictor Words I

What distinguishes the radical right in **Austria**?

- talking about **opposition**
- talking about Vienna (SD-governed)
- talking about health reform (RR-led ministry)
- less likely to address house as colleagues
- less likely to use gender-inclusive language

Best predictor Words II

What distinguishes the radical right in **Germany**?

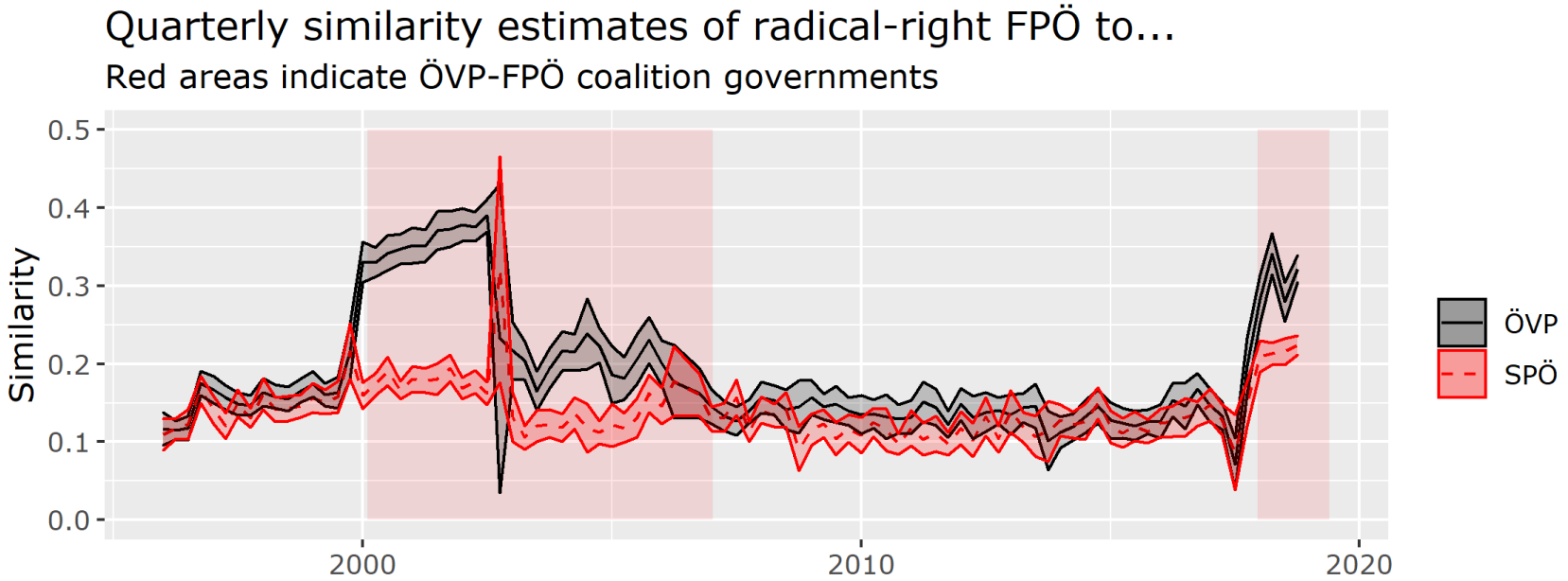
- talking about **government**
- talking about Germany
- less likely to address house as colleagues
- less likely to use gender-inclusive language
- less causal language

Best predictor Words III

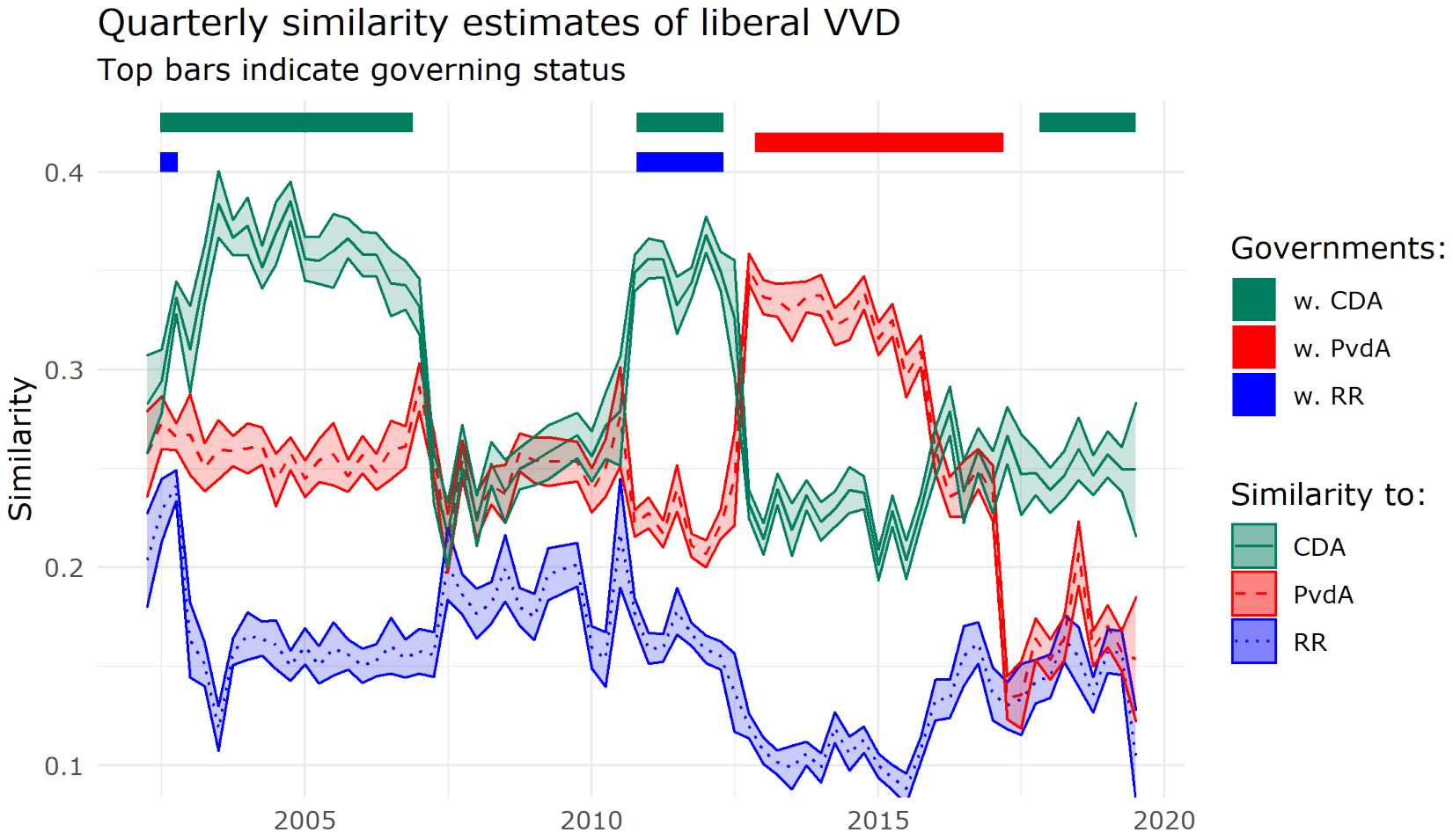
What distinguishes the radical right in **the Netherlands**?

- strong immigration focus
- more informal language
- less nuancing language

Radical-right government participation I



Radical-right government participation II



Speaker distribution I

Leaving the AfD

- One day after the national election in 2017, Frauke Petry and Mario Mieruch left the party
- They became independent members of the Bundestag

Leaving the VVD

- Geert Wilders left the VVD in September 2004
- this followed a rift with his party over the immigration issue
- as well as a series of escalations (position paper in July, vote on Turkey end of August)

Speaker distribution II

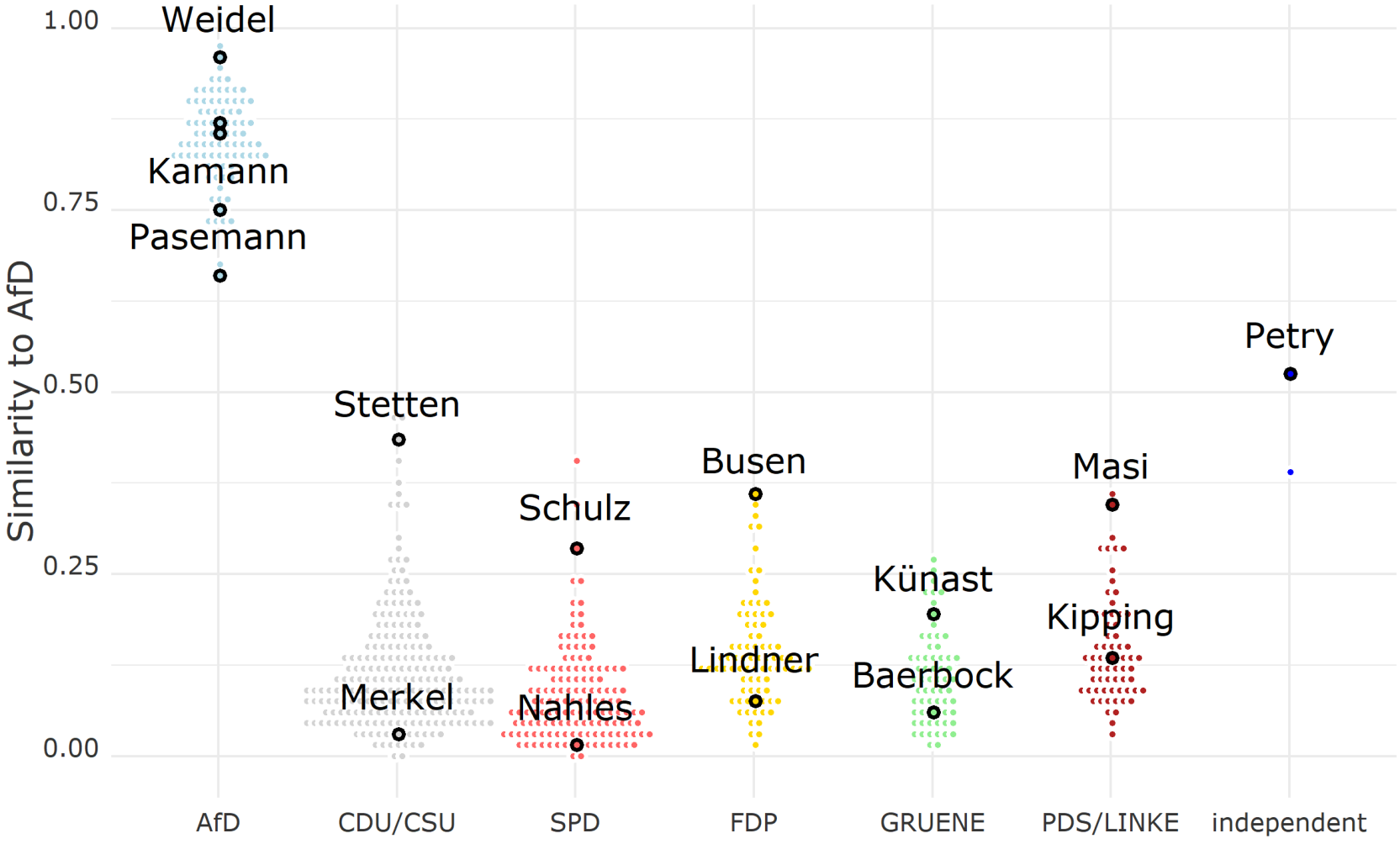
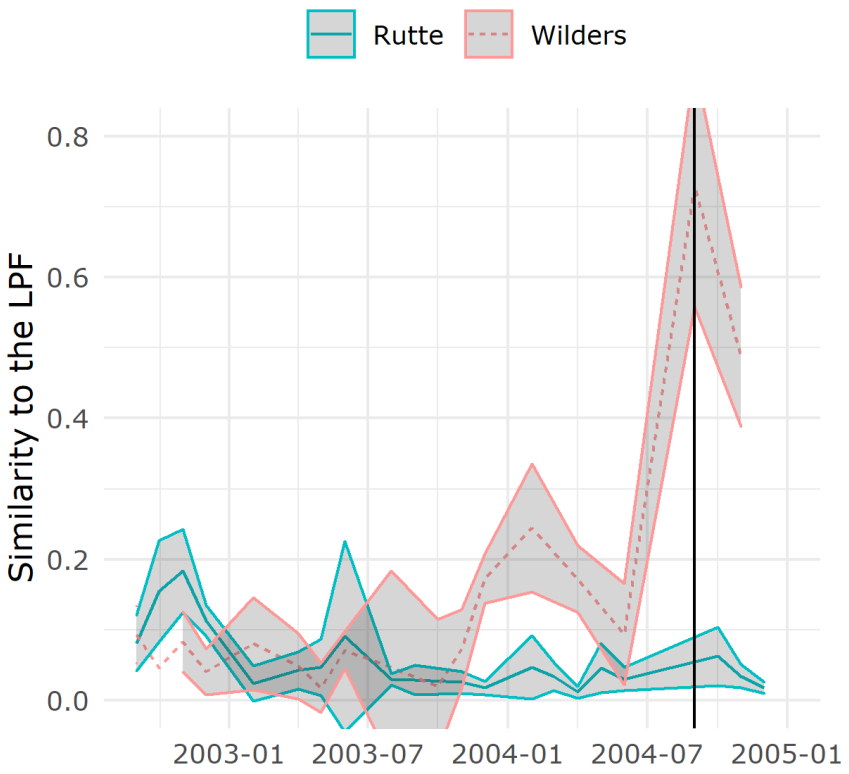
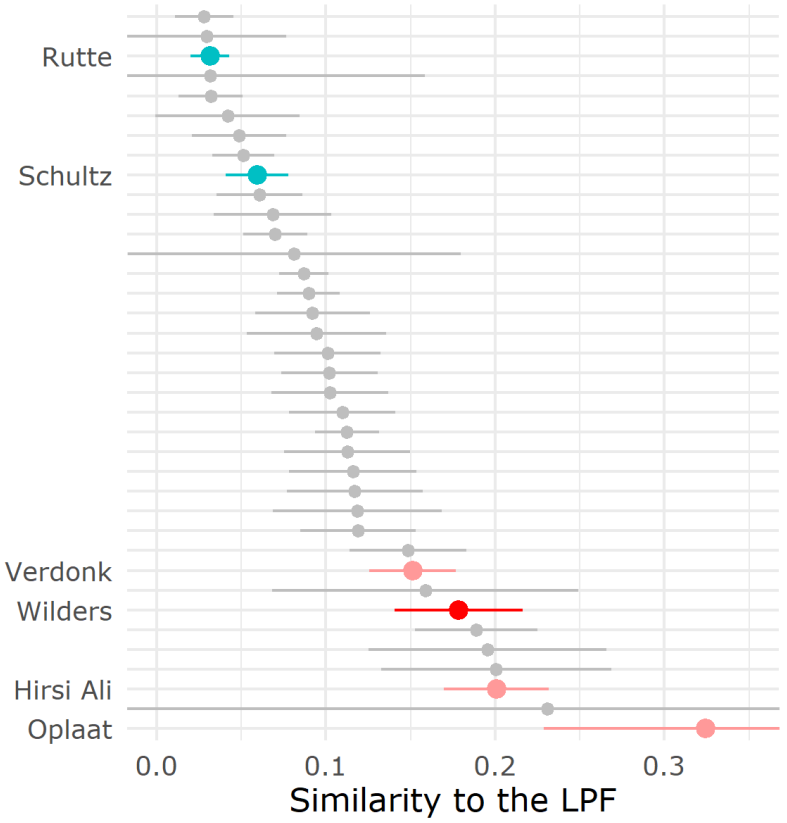


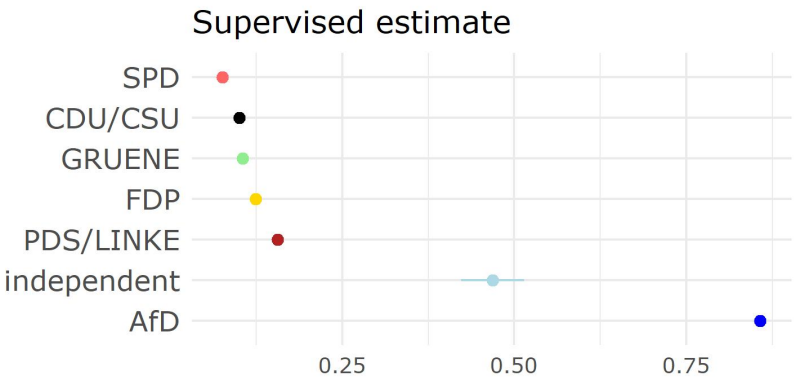
Figure: Speakers in the current German Bundestag.

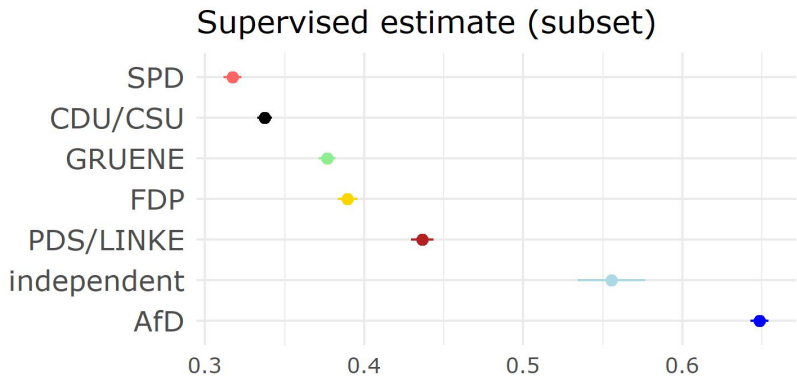
Speaker distribution III

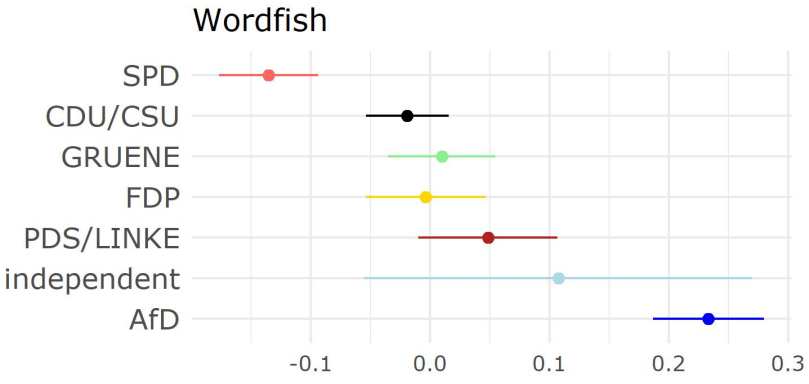


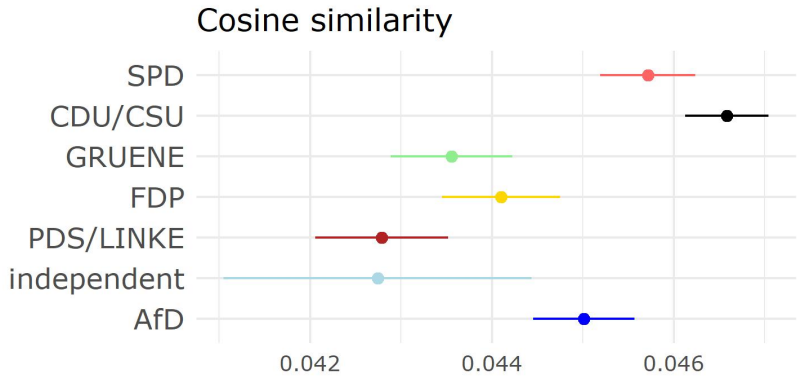


Comparable Measures









Summary

- SML estimate gives precise measure of rhetorical similarity
- the meaning of these differences can be assessed in detail
- outperforms other methods like WORDFISH and cosine similarity

Reviewer comments

- Scaling speeches on 'partyness' $\rightarrow$ not a novel method
- Doesn't build on Peterson and Spirling 2018 (scaling, not aggregate accuracy)
- Cosine similarity not a measure of similarity (document to corpus)

Possible routes forward

Substantive piece

- Testing classical theories of RR accommodation
- Improving predictors of coalition formation
- Historical: Showing accommodation of NSDAP by conservatives (Levitsky & Ziblatt)

Methods contribution

- Arguing for extended use of SML for better content validity
- Showing a possible application

Thank you!

Questions?

Resources I

- Chawla, Nitesh V. et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357. ISSN: 10769757. DOI: 10.1613/jair.953. arXiv: 1106.1813. URL: https://arxiv.org/pdf/1106.1813.pdf%7B%5C%%7D0Ahttp://www.snopes.com/horrors/insects/telamonia.asp.
- Laver, Michael, Kenneth Benoit, and John Garry (2003). "Extracting policy positions from political texts using words as data". In: *American Political Science Review* 97.2, pp. 311–331. ISSN: 00030554. DOI: 10.1017/S0003055403000698.
- Peterson, Andrew and Arthur Spirling (2018). "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems". In: *Political Analysis* 26.1, pp. 120–128. ISSN: 14764989. DOI: 10.1017/pan.2017.39.

Resources II

- Rauh, Christian and Jan Schwalbach (2020). "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies". In: *Harvard Dataverse, V1*, pp. 1–14. DOI: https://doi.org/10.7910/DVN/L4OAKN,. URL: https://dataverse.harvard.edu/dataverse/ParlSpeech.
- Slapin, Jonathan B. and Sven Oliver Proksch (2008). "A scaling model for estimating time-series party positions from Texts". In: *American Journal of Political Science* 52.3, pp. 705–722. ISSN: 00925853. DOI: 10.1111/j.1540-5907.2008.00338.x.