

# Predicting political party affiliation from text

Felix Biessmann<sup>1</sup>

Pola Lehmann<sup>2</sup>

Daniel Kirsch

Sebastian Schelter<sup>3</sup>

July 7, 2016

---

<sup>1</sup> `felix.biessmann@gmail.com`

<sup>2</sup> `pola.lehmann@wzb.eu`

<sup>3</sup> `sebastian.schelter@tu-berlin.de`

# Disclaimers

- (For me) This is just a hobby – it has nothing to do with my job
- I did not know a lot of literature in the field
  - Some of this might sound naive (like the title)
  - I hope nobody (who has been active in the field for years) takes this personal

# Disclaimers

- (For me) This is just a hobby – it has nothing to do with my job
- I did not know a lot of literature in the field
  - Some of this might sound naive (like the title)
  - I hope nobody (who has been active in the field for years) takes this personal

# Disclaimers

- (For me) This is just a hobby – it has nothing to do with my job
- I did not know a lot of literature in the field
  - Some of this might sound naive (like the title)
  - I hope nobody (who has been active in the field for years) takes this personal

# Disclaimers

- (For me) This is just a hobby – it has nothing to do with my job
- I did not know a lot of literature in the field
  - Some of this might sound naive (like the title)
  - I hope nobody (who has been active in the field for years) takes this personal

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
    - Preprocessing
    - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction



# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out-data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out-data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out-data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out-data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out-data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out-data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction



# Overview

- Motivation
- Methods
  - Data
  - Preprocessing
  - Classification Model
- Results
  - In-domain held-out data
  - Out-of-domain held-out-data
- Challenges of automated analyses
- Tools for better interpretability of models
- Conclusion
- Web applications of political bias prediction

# Data

- In-domain data (training data domain)
  - <http://www.bundestag.de/plenarprotokolle>
- Out-of-domain data (test data domain)
  - <https://manifestoproject.wzb.eu/>
  - Texts from public Facebook pages of parties

# Preprocessing

- Basic text cleaning (regexps, stopwords)
- Stemming
- n-grams (1-5)
- Tf-idf normalisation

# Classification Model: Multinomial Logistic Regression

Party affiliation estimate is modelled as

$$p(y = k|\mathbf{x}) = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}} \text{ with } z_k = \mathbf{w}_k^\top \mathbf{x}. \quad (1)$$

With

- Labels  $y \in \{1, 2, \dots, K\}$  (true party affiliation)
- $\mathbf{w}_1, \dots, \mathbf{w}_K \in \mathbb{R}^d$  weight vectors of  $k$ th party

# Model Selection

All hyperparameters optimised with nested cross-validation.

## Results: In-domain Predictions

Table: **17th Bundestag**

	precision	recall	f1-score	N
cducsu	0.62	0.81	0.70	706
fdp	0.70	0.37	0.49	331
gruene	0.59	0.40	0.48	298
linke	0.71	0.61	0.65	338
spd	0.60	0.69	0.65	606
total	0.64	0.63	0.62	2279

# Results: Out-of-domain Predictions

Table: **Tested on manifesto quasi-sentences**

	prec.	recall	f1-score	N
cducsu	0.26	0.58	0.36	2030
fdp	0.38	0.28	0.33	2319
gruene	0.47	0.20	0.28	3747
linke	0.30	0.47	0.37	1701
spd	0.26	0.16	0.20	2278
total	0.35	0.31	0.30	12075

# Why is out-of-domain classification so bad?

- Length of texts
- Text domain differences



# Effect of Text Length

Table: (topic level) **Manifesto data predictions**

	precision	recall	f1-score	N
cducsu	0.64	1.00	0.78	7
fdp	1.00	1.00	1.00	7
gruene	1.00	0.86	0.92	7
linke	1.00	1.00	1.00	7
spd	0.80	0.50	0.62	8
total	0.88	0.86	0.86	36

# Effect of Text Length

Table: **Facebook post predictions** (text length: 1000 words).

	precision	recall	f1-score	N
cducsu	0.65	1.00	0.79	50
gruene	0.67	0.12	0.20	50
linke	0.60	0.82	0.69	50
spd	1.00	0.92	0.96	50
avg / total	0.73	0.71	0.66	200

# Effect of Text Length

- Longer texts are easier to predict
- Intuitively makes sense
- In line with previous findings, see e.g. Hirst et al. [2014]
- But still, accuracies are far from perfect

# Effect of Text Length

What – except length – decreases generalization performance?

# Effect of Text Domain

Table: Classification texts into government and opposition (long texts).

	<b>In-Domain</b>	<b>Out-of-Domain</b>	
	Parliament	Manifestos	Facebook Posts
Accuracy	0.88	0.60	0.76

- Despite less noisy, longer texts:  
**Out-of-domain accuracy on manifesto data close to chance**
- Recognized in previous work, see e.g. Yu et al. [2008]

# Effect of Text Domain

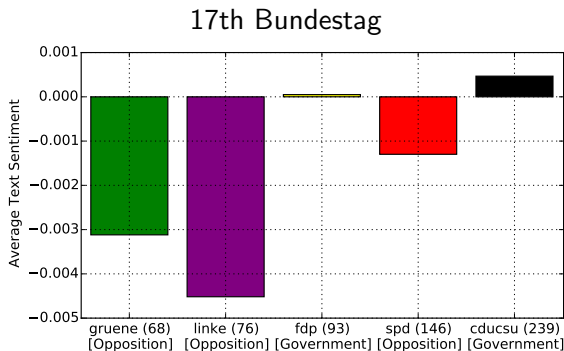
- Every ML model is biased by its training data
- Generalization from biased data is *the* central problem of ML
- Potential strategies to ensure generalization
  - Empirical risk minimization / Regularization
  - More (heterogeneous) data
  - Better models:  
Cov. shift adaptation, transfer/semi-supervised learning, ...
  - **Domain knowledge**

→ Using domain knowledge requires interpretable models

# Some ML Tools for Leveraging Domain Knowledge

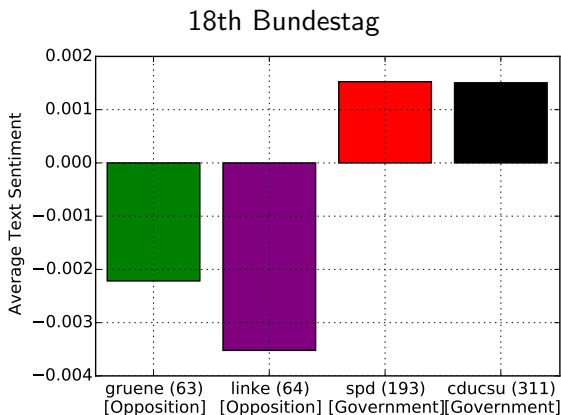
- Relation between misclassifications and party policy
- Covariation Text Features and Party labels
- Explicit tests of domain knowledge: Sentiment and Power

# Sentiment correlates with political power





# Sentiment correlates with political power

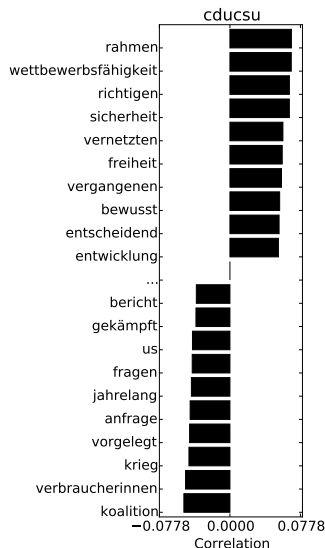


# Sentiment correlates with political power

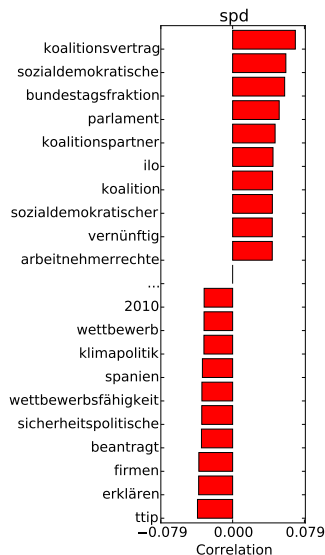
Table: Correlation coefficient between average sentiment with government membership and number of seats in the parliament.

Sentiment vs.	Gov. Member	Seats
17th Bundestag	0.84	0.70
18th Bundestag	0.98	0.89

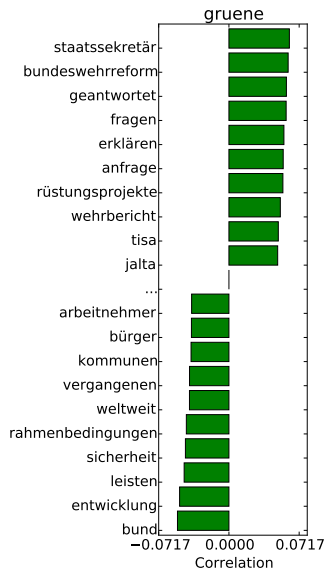
# Finding Discriminative Features



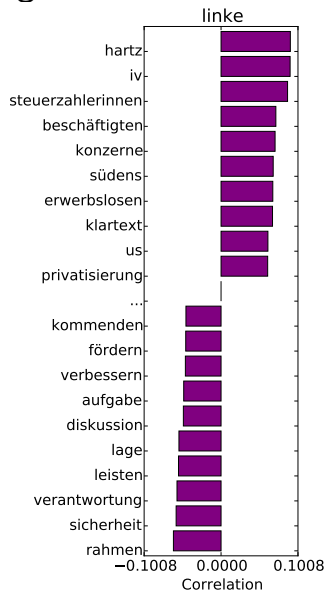
# Finding Discriminative Features



# Finding Discriminative Features



# Finding Discriminative Features



# Misclassifications and Policy Change

Confusion Matrix 17th Bundestag

		Predicted				
		cducsu	fdp	gruene	linke	spd
True	cducsu	7	0	0	0	0
	fdp	0	7	0	0	0
	gruene	0	0	6	0	1
	linke	0	0	0	7	0
	spd	4	0	0	0	4

# Conclusion

- Out-of-domain prediction of political bias possible
- Challenges
  - Text length, see also Hirst et al. [2014]
  - Domain transfer, see also Hirst et al. [2014]; Yu et al. [2008]
- Generalization should leverage domain knowledge
- Tools for leveraging domain knowledge
  - Relating misclassifications to policy changes
  - Interpreting discriminative features
  - Testing human experts' hypotheses explicitly



# Some Web Applications

The screenshot shows the 'linksrechts' website interface for 'Politische Gesinnungsanalyse'. The background features a word cloud with various political terms. The main heading 'linksrechts' is prominently displayed. Below it, the title 'Politische Gesinnungsanalyse' is centered. A paragraph explains the tool's function: 'Auf dieser Seite können Sie die politische Gesinnung von Texten und Internetseiten analysieren\*. Sie können einen Text in das erste Formular kopieren oder eine Internetseite analysieren, indem Sie eine URL in das zweite Formular kopieren. Wir analysieren auch kontinuierlich einige der großen Nachrichten-Seiten.' The interface includes two input sections: 'Analysiere einen Text' with a text area containing 'Reiche Banken und neoliberale gefährden den Sozialstaat.' and 'Analysiere eine Internetseite' with a text area for a URL. Both sections have a blue 'Analyse starten' button. At the bottom, a semi-circular gauge chart shows the political distribution of the analyzed text, with segments for 'linke' (purple), 'gruene' (green), 'spd' (red), and 'cdu' (black).

## linksrechts

### Politische Gesinnungsanalyse

Auf dieser Seite können Sie die politische Gesinnung von Texten und Internetseiten analysieren\*. Sie können einen Text in das erste Formular kopieren oder eine Internetseite analysieren, indem Sie eine URL in das zweite Formular kopieren. Wir analysieren auch kontinuierlich einige der großen Nachrichten-Seiten.

**Analysiere einen Text**

Reiche Banken und neoliberale gefährden den Sozialstaat.

**Analysiere eine Internetseite**

Hier eine URL zu einem Text reinpasten

Analyse starten

Analyse starten

linke gruene spd cdu

# Some Web Applications

ungarn flüchtlinge eu regierung für dublin orbán migranten jobbik  
polizisten



- 🇪🇺 Flüchtlinge in München: Ein freundliches, fröhliches Durcheinander
- 🇪🇺 Ungarn: Orbán droht mit Zaun an Grenze zu Kroatien
- 🇪🇺 Ungarn: Flüchtlinge treffen an der Grenze auf Rechtsradikale
- 🇪🇺 Flüchtlinge: "Deutschland hat eine mutige Entscheidung getroffen"
- 🇪🇺 Ungarns Ex-Premier nimmt Flüchtlinge auf
- 🇪🇺 Ungarische Polizei versucht Flüchtlinge in Aufnahmelager zu schleusen

pérez guatemala erlassen otto molina prääsidenten haftbefehl justiz  
zurückgetreten immunität



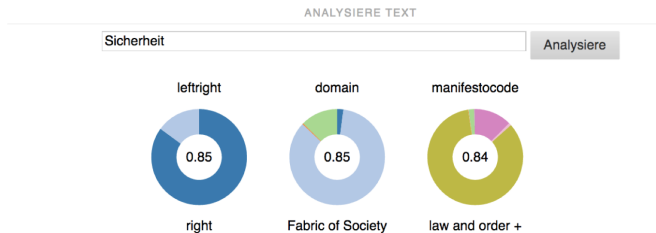
- 🇪🇺 Otto Pérez: Haftbefehl gegen Präsident von Guatemala erlassen
- 🇪🇺 Guatemala: Otto Pérez Molina tritt wegen Korruptionsaffäre zurück
- 🇪🇺 Guatemalas Präsident tritt zurück
- 🇪🇺 Lateinamerika: Guatemala braucht mehr als einen neuen Präsidenten

trump donald bush republikanischen spanisch republikaner  
unabhängiger us kandidat präsidenschaftskandidaten



- 🇪🇺 Trump über Jeb Bush: "Er sollte wirklich Englisch sprechen"
- 🇪🇺 Donald Trump erklärt Loyalität zu US-Republikanern
- 🇪🇺 Donald Trump verpflichtet sich Republikanern
- 🇪🇺 US-Präsidenschaftskandidat: Donald Trump meint es ernst

# Some Web Applications



# Some Web Applications

TOPIC 5

## hollande verfassungsänderung verfassungsreform franzosen

Anschläge von Paris: François Hollande zieht umstr ...

spiegel left (89%) Welfare and Quality of Life (58%) social justice + (56%)

Frankreich: Hollande zieht Verfassungsänderung zur ...

zeit left (50%) Welfare and Quality of Life (28%) gov-admin efficiency + (27%)

Francois Hollande begräbt Pläne für Verfassungsänd ...

faz right (53%) External Relations (60%) europe + (55%)

François Hollande: Das Ende einer politischen Schn ...

welt right (96%) Political System (86%) political authority + (81%)

Frankreichs Gewerkschaften blockieren Reformen ...

welt right (100%) Political System (97%) political authority + (96%)

# References

- G. Hirst, Y. Riabinin, J. Graham, and M. Boizot-Roche. Text to ideology or text to party status? In I. M. Bertie Kaal and A. van Elfrinkhof, editors, *From Text to Political Positions: Text analysis across disciplines*, pages 47–70, 2014.
- B. Yu, S. Kaufmann, and D. Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5 (1):33–48, 2008.