

# Predicting political party affiliation from text

Felix Biessmann<sup>1</sup>

Pola Lehmann<sup>2</sup>

Daniel Kirsch

Sebastian Schelter<sup>3</sup>

July 13, 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)

# 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)

# 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)

# 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?

1. Length of texts
2. Text domain differences

# Effect of Text Length

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

|         | precision | recall | f1-score | N  |
|---------|-----------|--------|----------|----|
| cducusu | 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   |
|-------------|-----------|--------|----------|-----|
| cducusu     | 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. ?
- 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:  
**Accuracy on manifesto data close to chance**



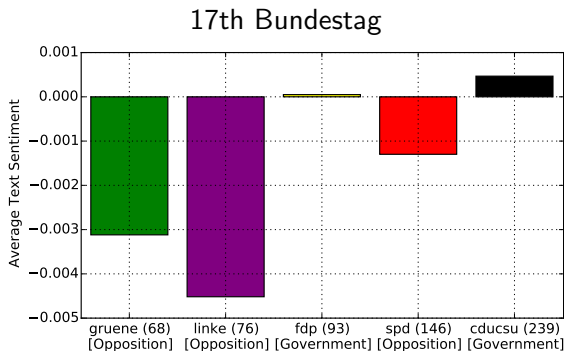
# Effect of Text Domain

- Recognized in previous work, see e.g. ?
  - Every ML model is biased by its training data
  - Generalization from biased data is *the* central problem of ML
  - Strategies to improve generalization
    - Empirical risk minimization / Regularization
    - More (heterogeneous) data
    - Better models:  
Cov. shift adaptation, transfer/semi-supervised learning, ...
    - **Domain knowledge**
- How can political scientists leverage domain knowledge for automatic text analysis models?

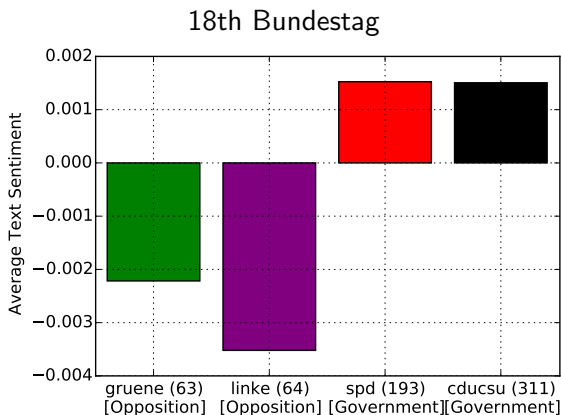
# Some ML Tools for Leveraging Domain Knowledge

- Relation between misclassifications and party policy
- Covariation Text Features and Party labels  
(**not model coefficients!**) ?
- Explicit tests of domain knowledge: Sentiment and Power

# Sentiment correlates with political power



# Sentiment correlates with political power

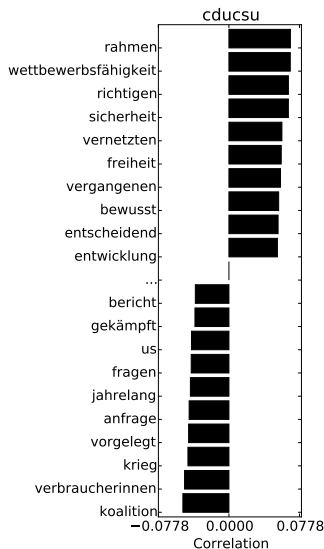


# Sentiment correlates with political power

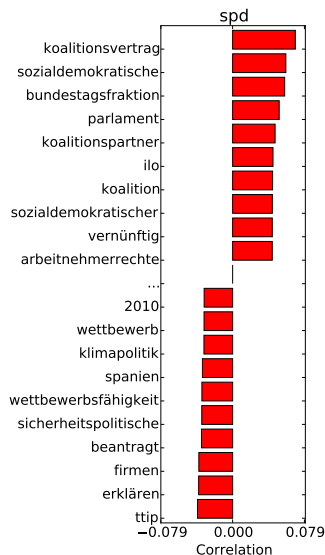
Table: Correlation coefficient between average sentiment with government membership and number of seats in the parliament.

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

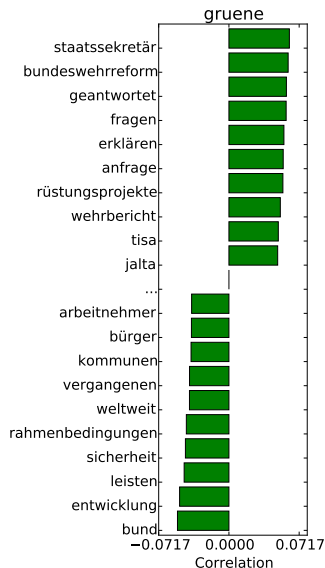
# Finding Discriminative Features



# Finding Discriminative Features

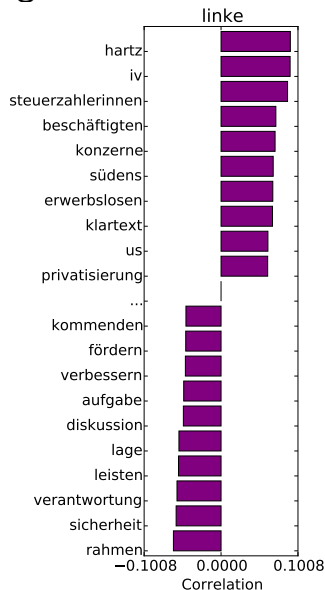


# Finding Discriminative Features





# Finding Discriminative Features



# Misclassifications and Policy Change

Confusion Matrix 17th Bundestag

|             |        | <b>Predicted</b> |     |        |       |     |
|-------------|--------|------------------|-----|--------|-------|-----|
|             |        | cducsu           | fdp | gruene | linke | spd |
| <b>True</b> | 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 ?
  - Domain transfer, see also ??
- 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, which is dedicated to 'Politische Gesinnungsanalyse' (Political Sentiment Analysis). The page features a large title 'linksrechts' at the top, followed by the subtitle 'Politische Gesinnungsanalyse'. Below this, there is a paragraph explaining the tool's purpose: '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.' (On this page, you can analyze the political sentiment of texts and websites\*. You can copy a text into the first form or analyze a website by pasting a URL into the second form. We also continuously analyze some of the major news sites.)

There are two main input sections:

- Analysiere einen Text**: A text input field containing the example text 'Reiche Banken und neoliberale gefährden den Sozialstaat.' and a blue 'Analyse starten' button.
- Analysiere eine Internetseite**: A text input field with the placeholder 'Hier eine URL zu einem Text reinpastein' and a blue 'Analyse starten' button.

At the bottom of the page, there is a semi-circular gauge chart showing the distribution of political sentiment. The chart is divided into five segments, each representing a different political party or ideology:

- linke** (Left): A large purple segment.
- gruene** (Green): A small green segment.
- spd** (Social Democratic Party): A red segment.
- cdu** (Christian Democratic Union): A black segment.
- rechts** (Right): A small red segment.

# 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



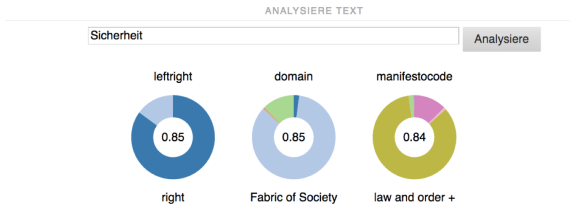
- Otto Pérez: Haftbefehl gegen Präsident von Guatemala erlassen
- Guatemala: Otto Pérez Molina tritt wegen Korruptionsaffäre zurück
- Guatemalas Präsident tritt zurück
- Lateinamerika: Guatemala braucht mehr als einen neuen Präsidenten

## trump donald bush republikanischen spanisch republikaner unabhängiger us kandidat präsidentchaftskandidaten



- Trump über Jeb Bush: "Er sollte wirklich Englisch sprechen"
- Donald Trump erklärt Loyalität zu US-Republikanern
- Donald Trump verpflichtet sich Republikanern
- US-Präsidentchaftskandidat: 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 (56%) 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 (87%) political authority + (96%)

# PyData Hackathon 2016 Berlin

What? Follow-up event of PyData Berlin 2016

Inviting Data Scientists, Social Scientists, UX Designers, ...



Data Ambassadors for

1. Manifesto Data
2. Parliament Data
3. Social Network Data

When? First weekend of October 2016 (1.-2.)

Where? Berlin



# References

- S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Biessmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.
- 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.