Automating Political Bias Prediction

Abstract

Every day media generate large amounts of text. An unbiased view on media reports requires an understanding of the political bias of media content. Assistive technology for estimating the political bias of texts can be helpful in this context. This study proposes a simple statistical learning approach to predict political bias from text. Standard text features extracted from speeches and manifestos of political parties are used to predict political bias in terms of political party affiliation and in terms of political views. Results indicate that political bias can be predicted with above chance accuracy. Mistakes of the model can be interpreted with respect to changes of policies of political actors. Two approaches are presented to make the results more interpretable: a) discriminative text features are related to the political orientation of a party and b) sentiment features of texts are correlated with a measure of political power. Political power appears to be strongly correlated with positive sentiment of a text. To highlight some potential use cases a web application shows how the model can be used for texts for which the political bias is not clear such as news articles.

1. Introduction

Analysis and classifications of political text is and has been a very important tool to generate political science data (Benoit et al. 2015). Traditionally such classifications are done by experts, who read and label the text of interest.[1] This is, however, a very time consuming task and thus sets various limits to the possible amount of data that a few experts can analyze. The growing field of automated text analysis, that allows the analysis of much more text in less time, is therefore of great interest to political scientists. Additionally automated text analyses allow for a more objective and replicable analysis of political text then human coders could achieve (Benoit et al. 2009)

One area which produces large amounts of text which are of high interest for political scientists is modern media. Standard newspapers, but also new forms as twitter, facebook or blogs produce an ever growing amount of text influencing political debate. These texts can be of interest to political scientist in various analyses. The question we will focus on in this analysis is the potential political bias that can be detected in these sources. In many cases it is obvious which political bias an author has. In other cases some expertise is required to judge the political bias of a text. Keeping an unbiased view on what media report on requires to understand the political bias of texts. Different newspaper outlets are, for example, often ascribed as favoring different positions on the political left-right axis. Data from the European Media Systems Survey show that experts clearly associate different media outlets with specific parties (http://www.mediasystemsineurope.org). However such expert surveys on the political bias of newspapers can normally only give us a very general tendency of a newspaper, they do not distinguish between different journalists or different sections. Assistive technology can help in this context to try and obtain a more unbiased sample of information. It gives us the possibility to get a more nuanced view of the spectrum of political positions favored in important news outlets in a country. In this paper we will test whether it is possible to automatically detect political bias in text and if such a model can be used on different kind of texts. The aim of this study is to provide some empirical evidence indicating that leveraging open data sources automated political bias prediction is possible with above chance accuracy. We will test our model on both in-domain and out-of-domain data. Prediction on out-of-domain data is less precise as on in-domain data, but still generating satisfying results.

In order to validate and explain the predictions of the models three strategies that allow for better interpretations of the models are proposed. First the model misclassifications are related to changes in party policies. Second univariate measures of correlation between text features and party affiliation allow to relate the predictions to the kind of information that political experts use for interpreting texts. Third sentiment analysis is used to investigate whether this

[1] See for example the Manifesto Project (Budge et al. 2001; Klingemann et al. 2006), the Comparative Agendas Project or Poltext (Ptry/Duval 2015).

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

aspect of language has discriminatory power.

In the following section 2 gives an overview of the data acquisition and preprocessing methods, section 3 presents the model, training and evaluation procedures; in section 4 the results are discussed and section 5 concludes with some interpretations of the results and future research directions.

2. Data Sets and Feature Extraction

All experiments were run on publicly available data sets of German political texts and standard libraries for processing the text. The following sections describe the details of data acquisition and feature extraction.

2.1. Data

Annotated political text data was obtained from two sources: a) the plenary debates held in the German parliament (*Bundestag*) and b) all manifesto texts of parties winning seats in the election to the German parliament in the current 18th and the last, 17th, legislative period.

Parliament discussion data Parliament texts are annotated with the respective party label, which we take here as a proxy for political bias. The protocols of plenary debates are available through the website of the German Bundestag[2]; an open source API was used to query the data in a cleaned and structured format[3]. In total 22784 speeches were extracted for the 17th legislative period and 11317 speeches for the 18th period, queried until March 2016.

Party manifesto data The party manifesto text was taken from the Manifesto Corpus (add citation). The Corpus is accessible through an publicly available API published by the *Manifesto Project* (**?**). The data released in this project mainly comprises the complete manifestos of all parties that have won seats at a national election. Each quasi-sentence[4] is annotated with one of 56 policy issue categories. Examples for the policy categories are *welfare state expansion, welfare state limitation, democracy, equality*; for a complete list and detailed explanations on how the annotators were instructed see (**?**). In total the 9 manifestos from the 17th and 18th legislative period have been divided into 29451 quasi-sentences. Each observation has been added with two labels: the party affiliation and the manually assigned policy issue aimed at in each specific statement.

The length of each annotated statement in the party manifestos is rather short. The median length is 95 characters or 12 words.[5] This can be considered as a very valuable property of the data set, because it allows a fine grained resolution of party manifestos and is a good test for media data, as some data like tweets, facebook texts etc. are often very short, too. However for a classifier (as well as for humans) such short sentences can be rather difficult to classify. Benoit et al. (2015), for example, found in an experiment where experts where differentiating whether specific sentences were dealing with economic or social policy only about 35% agreement between all expert coders. In the Manifesto Project the human coders therefore do not code these statements separately but within context. In order to obtain less ’noisy’ data points from each party – for the party affiliation task only – all statements were aggregated into eight policy domains using the 56policy categories from the Manifesto Project. Within the Manifesto Project category scheme each of these belongs to one of seven policy domains. These were used to aggregate the data into the following topics: *External Relations, Freedom and Democracy, Political System, Economy, Welfare and Quality of Life, Fabric of Society, Social Groups*. An eight topic is generated for all statements not belonging to any one the 56 policy issues, for a complete list see also (**?**). Most party manifestos covered all eight topics, some party manifestos in the 17th Bundestag only covered seven.

2.2. Bag-of-Words Vectorization

First each data set was segmented into semantic units; in the case of the party manifesto data the semantic units were the quasi-sentences associated with one of the 56 policy issues, in the case of parliament discussions this were the speeches. Parliament speeches were often interrupted; in this case each uninterrupted part of a speech was considered a semantic unit. Strings of each semantic unit were tokenised and transformed into bag-of-word vectors as implemented in scikit-learn (**?**). The general idea of bag-of-words vectors is to simply count occurrences of words (or word sequences, also called *n-grams*) for each data point. A data point is usually a document, here it is the semantic units of parliament speeches and manifesto sentences, respectively. The text of each semantic unit is transformed into a vector $\mathbf{x} \in \mathbb{R}^d$ where d is the size of the dictionary; the wth entry of $\mathbf{x}$ contains the (normalized) count of the wth word (or sequence of words) in our dictionary. Several options for vectorizing the speeches were tried, including term-frequency-inverse-document-frequency normalisation, n-gram patterns up to size $n = 3$ and several cutoffs for discarding too frequent and too infrequent words. All of these hyperparameters were subjected to hyperparameter

[2] https://www.bundestag.de/protokolle

[3] https://github.com/bundestag

[4] A quasi-sentence has the length of an argument. It is never longer than one sentence.

[5] The longest statement is 522 characters long, the 25%/50%/75% percentiles are 63/95/135 characters. Measured in words the longest data point is 65 words and the 25%/50%/75% percentiles are 8/12/17 words, respectively.

optimization as explained in subsection 3.1.

3. Classification Model and Training Procedure

Bag-of-words feature vectors were used to train a multinomial logistic regression model. Let $y \in \{1, 2, \ldots, K\}$ be the true label, where K is the total number of labels and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{R}^{d \times K}$ is the concatenation of the weight vectors $\mathbf{w}_k$ associated with the kth party then

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

We estimated $\mathbf{W}$ using quasi-newton gradient descent. The optimization function was obtained by adding a penalization term to the negative log-likelihood of the multinomial logistic regression objective and the optimization hence found the $\mathbf{W}$ that minimized

$$L(\mathbf{W}, \mathbf{x}, \gamma) = -\log \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} + \gamma \|\mathbf{W}\|_F \tag{2}$$

Where $\| \ \|_F$ denotes the Frobenius Norm and γ is a regularization parameter controlling the complexity of the model. The regularization parameter was optimized on a log-scaled grid from $10^{-4,\ldots,4}$. The performance of the model was optimized using the classification accuracy, but we also report all other standard measures, precision $(TP/(FP + TP)$, recall $(TP/(TP + FN))$ and f1-score $(2 \times (Prec. \times Rec)/(Prec + Rec.))$.

Three different classification problems were considered:

1. **Classification of party affiliation** (five class / four class problem)
2. **Classification of government membership** (binary problem)
3. **Classification of policy issues** (56 class problem)

Party affiliation is a five class problem for the 17th legislation period, and a four class problem for the 18th legislation period. The policy issue classification is based on the categories of the Manifesto Project, see section 2 and (?). For both the party affiliation and the government membership prediction, classifiers were trained on the parliament speeches. For the third problem classifiers were trained only on the manifesto data for which policy issue labels were available.

3.1. Optimisation of Model Parameters

The model pipeline contained a number of hyperparameters that were optimised using cross-validation. We first split the training data into a training data set that was used for optimisation of hyperparameters and an held-out test data set for evaluating how well the model performs on in-domain data; wherever possible the generalisation performance of the models was also evaluated on out-of domain data. Hyperparameters were optimised using grid search and 3-fold cross-validation within the training set only: A cross-validation split was made to obtain train/test data for the grid search and for each setting of hyperparameters the entire pipeline was trained and evaluated – no data from the in-domain evaluation data or the out-of-domain evaluation data were used for hyperparameter optimisation. For the best setting of all hyperparameters the pipeline was trained again on all training data and evaluated on the evaluation data sets. For party affiliation prediction and government membership prediction the training and test set were 90% and 10%, respectively, of all data in a given legislative period. Out-of-domain evaluation data were the texts from party manifestos. For the policy issue prediction setting there was no out-of-domain evaluation data, so all labeled manifesto sentences in both legislative periods were split into a training and evaluation set of 90% (train) and 10% (evaluation).

3.2. Sentiment analysis

A publicly available key word list was used to extract sentiments (?). A sentiment vector $\mathbf{s} \in \mathbb{R}^d$ was constructed from the sentiment polarity values in the sentiment dictionary. The sentiment index used for attributing positive or negative sentiment to a text was computed as the cosine similarity between BOW vectors $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{s}$

$$\frac{\mathbf{s}^\top \mathbf{x}}{\|\mathbf{s}\| \|\mathbf{x}\|} \tag{3}$$

3.3. Analysis of bag-of-words features

While interpretability of linear models is often propagated as one of their main advantages, doing so naively without modelling the noise covariances can lead to wrong conclusions, see e.g. (**??**); interpreting coefficients of linear models (independent of the regularizer used) implicitly assumes uncorrelated features; this assumption is violated by the text data used in this study. Thus direct interpretation of the model coefficients $\mathbf{W}$ is problematic. In order to allow for better interpretation of the predictions and to assess which features are discriminative correlation coefficients between each word and the party affiliation label were computed. The words corresponding to the top positive and negative correlations are shown in subsection 4.3.

4. Results

The following section gives an overview of the results for all political bias prediction tasks. Some interpretations of the results are highlighted and a web application of the models is presented at the end of the section. Because of space restrictions, we will only report results fromt he 17th legislative period here. Results are very similar for the 18th legislative period and can be accessed via github.

4.1. Predicting political party affiliation

The evaluation results for the political party affiliation prediction on in-domain data (held-out parliamentary speech text) for the 17th and 18th Bundestag are listed in Table 1.

When predicting party affiliation on text data from the same domain that was used for training the model, average precision and recall values of above 0.6 are obtained. These results are comparable to those of (**?**) who report a classification accuracy of 0.61 on a five class problem predicting party affiliation in the European parliament; the accuracy for the 17th Bundestag is 0.63, results of the 18th Bundestag are difficult to compare as the number of parties is four and the legislation period is not finished yet.

But as the main purpose of this paper is to generate a model that can be used to predict political bias on new data, like media texts, we are interested in the question if the model also produces satisfying results on out-of-domain data. We are therefore also predicting party affiliation on party manifesto text. For out-of domain data the models yield significantly lower precision and recall values between 0.3 and 0.4 Table 2. This drop in out of domain prediction accuracy is in line with previous findings (**?**). A main factor that made the prediction on the out-of-domain prediction task particularly difficult is the short length of the strings to be classified, see also section 2. In order to investigate whether this low out-of-domain prediction performance was due to the domain difference (parliament speech vs manifesto data) or due to the short length of the data points, the manifesto data was aggregated into the previously described party domains, see section 2. The topic level results are shown in Table 3 and demonstrate that when the texts to be classified are sufficiently long and the word count statistics are sufficiently dense the classification performance on out-of-domain data can - at least in some cases - achieve reliable precision and recall values close to 1.0. This increase is in line with previous findings on the influence of text length on political bias prediction accuracy (**?**).

4.2. Validation

Differentiability of policy issues One reason why some statements are difficult to classify could be that some issues are more difficult to differentiate than others. To test this a separate suite of experiments was run to train and test the prediction performance of the text classifiers models described in section 3. As there was no out-of-domain evaluation set available in this setting only the evaluation error on in-domain data is reported. Note however that in this experiment too the evaluation data was never seen by any model during training time. In Table 8 results for the best and worst classes, in terms of predictability, are listed along with the average performance metrics on all classes. Precision and recall values of close to 0.5 on average can be considered rather high considering the large number of labels. Results are not surprising, but are equal to those by human coders.

Table 1. Classification performance on the party affiliation prediction problem for data from the 17th legislative period on in-domain data. N denotes number of data points in the evaluation set.

	17th legislative period				**18th legislative**		
	precision	recall	f1-score	N	precision	recall	f1
cducsu	0.62	0.81	0.70	706	0.66	0.82	
fdp	0.70	0.37	0.49	331			
gruene	0.59	0.40	0.48	298	0.68	0.54	
linke	0.71	0.61	0.65	338	0.77	0.58	
spd	0.60	0.69	0.65	606	0.60	0.54	
avg / total	0.64	0.63	0.62	2279	0.66	0.66	

Table 2. Classification performance on the party affiliation prediction problem for data from the 17th legislative period on out-of-domain data. Predictions are done on the **sentence level**.

	17th legislative period				**18th legislati**	
	precision	recall	f1-score	N	precision	recall
cducsu	0.26	0.58	0.36	2030	0.32	0.64
fdp	0.38	0.28	0.33	2319		
gruene	0.47	0.20	0.28	3747	0.59	0.15
linke	0.30	0.47	0.37	1701	0.36	0.48
spd	0.26	0.16	0.20	2278	0.26	0.31
avg / total	0.35	0.31	0.30	12075	0.42	0.34

Policy change Another explanation for missclassification could lie in the fact that parties change their policy positions. In order to investigate this confusion matrices were extracted for the predictions on the out-of-domain evaluation data for sentence level predictions (see Table 4) as well as topic level predictions (see Table 5). On the topic level results for the 17th legislative period are pretty good except for the SPD. In the 18th legislative period predictions are very bad for the Green party. One explanation for this missclassification could lie in policy changes. For example, the Green party has devoted large parts of its manifesto to renewable and against nuclear energy. After Fukushima the other parties adopted these positions.

Table 3. **Topic level classification performance** on the party affiliation prediction problem for data from the evaluation set (manifesto texts) of the 17th legislative period. In contrast to single sentence level predictions (see **??**, **??**, Table 4 for results and section 2 for topic definitions) the predictions made on topic level are reliable in many cases. Note that all manifesto topics of the green party in the 18th Bundestag are predicted to be from the parties of the governing coalition, CDU/CSU or SPD.

Party Manifestos
17th Bundestag

	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
avg / total	0.88	0.86	0.86	36

Party Manifestos
18th Bundestag

	precision	recall	f1-score	N
cducsu	0.50	1.00	0.67	8
gruene	0.00	0.00	0.00	8
linke	1.00	0.88	0.93	8
spd	0.56	0.62	0.59	8
avg / total	0.51	0.62	0.55	32

4.3. Correlations between words and parties

We now know that party affiliation can be predicted using our model. But which words are most important for the classification performed here? To get a more precise idea of what differentiates the parties from each other the 10 highest and lowest correlations between individual words and the party affiliation label are shown for each party in Figure 1. Correlations were computed on the data from the current, 18th, legislative period. Some unspecific stopwords are excluded. The following paragraphs highlight some examples of words that appear to be preferentially used or avoided by each respective party. Even though interpretations of these results are problematic in that they neglect the context in which these words were mentioned some interesting patterns can be found and related to the actual policies the parties are promoting.

Table 4. Classification performance of 56 political views, see section 2.

code	meaning	precision	recall	f1-score	
501	environmentalism +	0.62	0.61	0.61	16
202	democracy +	0.58	0.55	0.57	12
701	labour +	0.57	0.54	0.56	12
201	freedom/human rights +	0.58	0.54	0.56	15
106	peace +	0.52	0.57	0.55	2
. . .					
302	centralism +	0.25	0.20	0.22	1
401	free enterprise +	0.20	0.19	0.20	5
505	welfare -	0.13	0.14	0.14	1
409	keynesian demand +	0.14	0.12	0.13	
0	undefined	0.09	0.12	0.10	1
avg / total		0.47	0.46	0.46	294

Left party (linke) The Left party talks mainly about issues related to work (*Beschäftigten*) and unemployment (*Hartz IV, Erwerbslosen*).

Green party (gruene) The Green party talks about the formal procedures available to opposition parties to question government policy (*fragen, anfragen*) and, interestingly, about the reform of the Germany army (*Rüstungsprojekte, Wehrbericht*).

Social Democratic Party (SPD) The SPD often uses words related to rights of the working class (*International Labour Organisation, Arbeitnehmerrechte*) and the governing coalition (*bundestagsfraktion, koalitionspartner*).

Christian Democratic Union/Christian Social Union (CDU/CSU) The CDU/CSU party often uses words related to a pro-economy attitude, such as competitiveness or (economic) development (*Wettbewerbsfähigkeit, Entwicklung*) and words related to security (*Sicherheit*).

4.4. Predicting government status

Next to the party affiliation labels also government membership labels were used to train models that predict whether or not a text is from a party that belonged to a governing coalition of the Bundestag. In Table 6 the results are shown for the 17th legislative period. While the in-domain evaluation precision and recall values reach values close to 0.9, the out-of-domain evaluation drops again to values between 0.6 and 0.7. This is in line with the results on binary classification of political bias in the Canadian parliament (**?**). The authors report classification accuracies between 0.8 and 0.87, the accuracy in the 17th Bundestag was 0.85. While topic-level predictions were

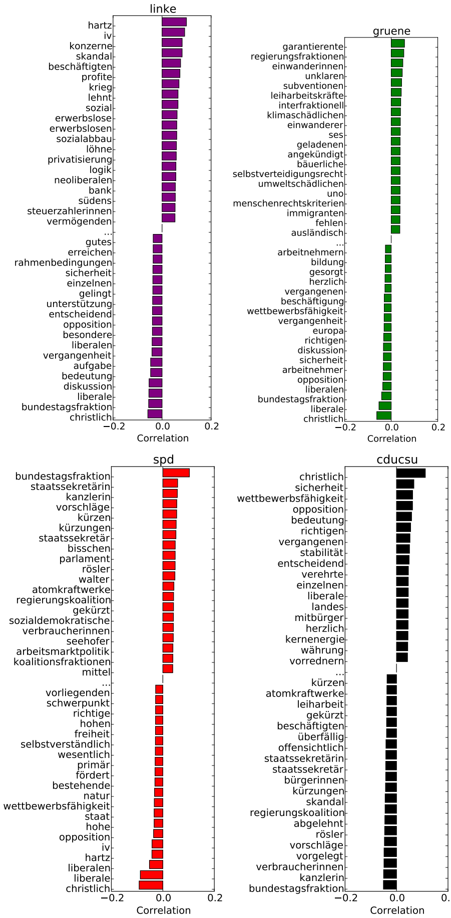

Figure 1. Correlations between words and party affiliation label for parliament speeches can help interpreting the features used by a predictive model. Shown are the top 10 positively and negatively correlated text features for the current Bundestag. For interpretations see subsection 4.3.

Table 5. **Confusion matrices (sentence level)** for predictions on evaluation data (party manifestos); classifiers were trained on parliament speeches for the 17th legislative period (left) and 18th legislative period (right); the most prominent effect is the high likelihood for a party to be taken as the strongest, governing party, cdu/csu. This can be interpreted as a change in policies of the conservative party cdu/csu towards the policies of the green party.

Party Manifestos
17th Bundestag

		Predicted				
		cducsu	fdp	gruene	linke	spd
True	cducsu	1186	289	178	198	179
	fdp	882	658	236	329	214
	gruene	1174	404	764	941	464
	linke	388	92	214	806	201
	spd	999	268	240	398	373

Party Manifestos
18th Bundestag

		Predicted			
		cducsu	gruene	linke	spd
True	cducsu	1912	156	331	584
	gruene	2092	827	1311	1444
	linke	596	186	1216	557
	spd	1284	226	563	916

not performed in this binary setting, the party affiliation results in Table 3 suggest that a similar increase in out-of-domain prediction accuracy could be achieved when aggregating texts to longer segments.

Speech sentiment correlates with political power But what are the underlying features that give rise to the classifiers performance? In order to investigate this the bag-of-words features were analysed with respect to their sentiment. The average sentiment of each political party is shown in Figure 2. High values indicate more pronounced usage of positive words, whereas negative values indicate more pronounced usage of words associated with negative emotional content.

The results show an interesting relationship between political power and sentiment. Political power was evaluated in terms of membership of the government. Correlating these indicators of political power with the mean sentiment of a party shows a strong positive correlation between speech sentiment and political power. This pattern is evident from the data in Figure 2 and in Table 9: In the current Bundestag, government membership correlates with positive sentiment with a correlation coefficient of 0.98 and the

Table 6. **Confusion matrices (topic level)** for predictions on evaluation data (party manifestos) for classifiers trained on parliament speeches for the 17th legislative period (left) and 18th legislative period (right).

Party Manifestos
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

Party Manifestos
18th Bundestag

			Predicted		
		cducsu	gruene	linke	spd
True	cducsu	8	0	0	0
	gruene	4	0	0	4
	linke	1	0	7	0
	spd	3	0	0	5

number of seats correlates with 0.89.

Note that there is one party, the social democrats (SPD), which has many seats and switched from opposition to government with the 18th Bundestag: With its participation in the government the average sentiment of this party switched sign from negative to positive, suggesting that positive sentiment is a strong indicator of government membership.

4.5. An example web application

To show an example use case of the above models a web application was implemented that downloads regularly all articles from some major German news paper websites[6] and applies some simple topic modelling to them. For each news article topic, headlines of articles are plotted along with the predictions of the policy issue of an article and two labels derived deterministically from the 56 class output, a left right index and the political domain of a text, see (?). Within each topic it is then possible to get an ordered (from left to right) overview of the articles on that topic. An example of one topic that emerged on March 31st is shown in Figure 3. A preliminary demo is live at (?) and the code is available on github(?).

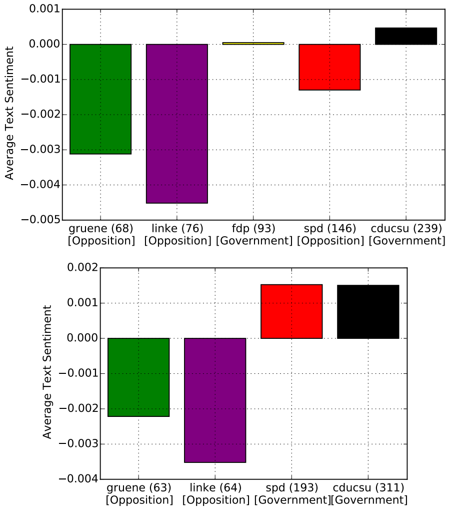

Figure 2. Speech sentiments computed for speeches of each party; parties are ordered according to the number of seats in the parliament. There is a trend for more positive speech content with more political power. Note that the SPD (red) switched from opposition to government in the 18th Bundestag: their seats in the parliament increased and the average sentiment of their speeches switched sign from negative to overall positive sentiment.

Figure 3. A screen shot of an example web application using the political view prediction combined with topic modelling to provide a heterogeneous overview of a topic.

5. Conclusions, Limitations and Outlook

This study presents a simple approach for automated political bias prediction. The results of these experiments show that automated political bias prediction is possible with above chance accuracy in some cases. It is worth noting that even if the accuracies are not perfect, they are above chance and comparable with results of comparable studies (**??**). While these results do not allow for usage in production systems for classification, it is well possible to use such a system as assistive technology for human annotators in an active learning setting.

One of the main limiting factors of an automated political bias prediction system is the availability of training data. Most training data sets that are publicly available have an inherent bias as they are sampled from a different domain. This study tried to quantify the impact of this effect. For the cases in which evaluation data from two domains was available there was a pronounced drop in prediction accuracy between the in domain evaluation set and the out of domain evaluation set. This effect was reported previously

[6] http://www.spiegel.de/politik, http://www.faz.net/aktuell/politik, http://www.welt.de/politik, http://www.sueddeutsche.de/politik, http://www.zeit.de/politik

Table 7. Classification performance on the binary prediction problem in the 17th legislative period, categorizing speeches into government (FDP/CDU/CSU) and opposition (Linke, Grüne, SPD).

	Held-out parliament speeches				**Party Manifestos**			
	precision	recall	f1-score	N	precision	recall	f1-score	N
government	0.83	0.84	0.84	1037	0.49	0.59	0.54	4349
opposition	0.86	0.86	0.86	1242	0.74	0.66	0.70	7726
avg / total	0.85	0.85	0.85	2279	0.65	0.63	0.64	12075

Table 8. Classification performance on the binary prediction problem in the 18th legislative period, categorizing speeches into government (SDP/CDU/CSU) and opposition (Linke, Grüne).

	Held-out parliament speeches				**Party Manifestos**			
	precision	recall	f1-score	N	precision	recall	f1-score	N
government	0.88	0.95	0.92	786	0.52	0.66	0.58	5972
opposition	0.86	0.71	0.78	346	0.69	0.56	0.62	8229
avg / total	0.88	0.88	0.87	1132	0.62	0.60	0.60	14201

Table 9. Correlation coefficient between average sentiment of political speeches of a party in the german Bundestag with two indicators of political power, a) membership in the government and b) the number of seats a party occupies in the parliament.

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

for similar data, see e.g. (**?**). Also the finding that shorter texts are more difficult to classify than longer texts is in line with previous studies (**?**). When considering texts of sufficient length (for instance by aggregating all texts of a given political topic) classification performance improved and in some cases reliable predictions could be obtained even beyond the training text domain.

Some aspects of these analyses could be interesting for social science researchers; three of these are highlighted here. First the misclassifications of a model can be related to the changes in policy of a party. Such analyses could be helpful to quantitatively investigate a change in policy. Second analysing the word-party correlations shows that some discriminative words can be related to the political views of a party; this allows for validation of the models by human experts. Third when correlating the sentiment of a speech with measures of political power there is a strong positive correlation between political power and positive sentiment. While such an insight in itself might seem not very surprising this quantifiable link between power and sentiment could be useful nonetheless: Sentiment analysis is a rather domain independent measure, it can be easily automated and scaled up to massive amounts of text data. Combining sentiment features with other measures of political bias could potentially help to alleviate some of the domain-adaptation problems encountered when applying models trained on parliament data to data from other domains.

All data sets used in this study were publicly available, all code for experiments and the live web application can be found online (**??**).

Acknowledgements

We would like to thank Friedrich Lindenberg for factoring out the https://github.com/bundestag/plpr-scraper from his Bundestag project. Some backend configurations for the web application were taken from an earlier collaboration with Daniel Kirsch. Michael Gaebler provided helpful feedback on an earlier version of the manuscript.