Predicting political party affiliation from text

Felix Biessmann* **Pola Lehmann†** **Daniel Kirsch** **Sebastian Schelter ‡**

Abstract

Every day a large amount of text is produced during public discourse. Some of this text is produced by actors whose political colour is very obvious. However, though many actors cannot clearly be associated with a political party, their statements may be biased towards a specific party. Identifying such biases is crucial for political research as well as for media consumers, especially when analysing the influence of the media on political discourse and vice versa. In this study, we investigate the extent to which political party affiliation can be predicted from textual content. Results indicate that automated classification of political affiliation is possible with an accuracy better than chance, even across different text domains. We propose methods to better interpret these results, and find that features not related to political policies, such as speech sentiment, can be discriminative and thus exploited by text analysis models.

1 Introduction

Analysis and classifications of political text is and has been a very important tool to generate political science data [8]. Traditionally, experts conduct such classifications by reading and labelling the text of interest[1]. This is, however, a very time consuming task and thus sets various limits on the possible amount of data that a few experts can analyse. The growing field of automated text analysis, which allows for 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 than human coders can achieve [9].

A major problem with automated text analyses is generalisation to text domains other than that on which the system has been trained [15]. While political experts can read texts from different domains and are able to detect political bias appearing in a variety of contexts and styles, machine learning algorithms are prone to poor performance generalisation across text domains if the training data is biased towards one domain only. Unfortunately, good unbiased training data is difficult to obtain. One of the best sources for automated political text analysis systems are plenary debates of the parliament: many studies are based on this type of data, as it consists of a large source of text that can be clearly associated with a party. We examine to what extent models trained on this data can generalise their predictions to other text domains, such as party manifestos and texts from social media. We discuss the effects of text length and domain shifts of text data, and investigate some possible reasons for the differences in classification performance.

We investigate the predictions of the models with three strategies: first, we test the influence of text length on the prediction accuracy. Second, we use sentiment analysis to investigate whether this aspect of language has discriminatory power. Third, univariate measures of correlation between text features and party affiliation allow us to relate the predictions to the kind of information that political experts use for interpreting texts.

In this article, section 2 gives an overview of the data acquisition and preprocessing methods, section 3 presents the model, training and evaluation procedures, in section 4 we discuss results and section 5 concludes with interpretations of the results.

*felix.biessmann@gmail.com
†pola.lehmann@wzb.eu
‡sebastian.schelter@tu-berlin.de

[1] See for example the Manifesto Project, the Comparative Agendas Project or Poltext.

2 Data Sets and Feature Extraction

We ran all experiments using publicly available data sets of German political texts, and applied 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 three sources: a) the plenary debates held in the German parliament (*Bundestag*) b) all manifestos of parties winning seats in the election to the German parliament and c) facebook posts from all parties. The texts from plenary debates were used to train a classifier and evaluate it on this in-domain data. We employed the latter two data sources to test the generalisation performance of the classifier on out-of-domain data.

Parliament discussion data Parliament texts are annotated with the respective party label. The protocols of plenary debates are available through the website of the German Bundestag [3]; we leveraged an open source API to query the data in a cleaned and structured format [2]. Each uninterrupted part was treated as a separate speech.

Party manifesto data The party manifesto text originates from the Manifesto Corpus [12]. The data released in this project mainly comprises the complete manifestos of all parties that have won seats in a national election. Each statement or *quasi-sentence*[2] 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 [1]. Each quasi-sentence has two types of labels: the party affiliation and the manually assigned policy issue aimed at in each quasi-sentence. The length of each annotated statement in the party manifestos is rather short. The median length is 95 characters, or 12 words[3]. In order to increase the length of the texts for classification, we used the policy labels 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*. In this setting, each party had just one data point for each of the topics.

Facebook post data We crawled the facebook page of each party [4, 7, 5, 6] and extracted the post texts, excluding all comments and other information. Like the manifesto data, these texts are very short. As aggregation per topic was not possible for this data, we aggregated the texts by splitting all texts into parts of 1000 words.

2.2 Bag-of-Words Vectorisation

We tokenised all text data and transformed it into bag-of-word (BOW) vectors as implemented in scikit-learn [13]. Several options for BOW vectorisations were tried, including term-frequency-inverse-document-frequency normalisation, n-gram patterns up to size $n = 3$ and different cut-offs for discarding words which were too frequent or infrequent.

3 Classification Model and Training

We leveraged bag-of-words feature vectors to train a multinomial logistic regression model. Let $y \in \{1, 2, \ldots, K\}$ be the true party affiliation and $\mathbf{w}_1, \ldots, \mathbf{w}_K \in \mathbb{R}^d$ the weight vectors associated with the kth party. Then the party affiliation estimate is modelled as

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

3.1 Optimisation of Model Parameters

The model pipeline contained a number of hyperparameters that we optimised using gridsearch cross-validation. To this end, we split the parliament speech data into training and validation sets in a 90%/10% ratio; we trained the pipeline with each parameter setting on the training set and validated its performance on the validation set. We chose the parameters of the best performing model to train a model on the training and validation set data. None of the data in the separately held back in-domain test data nor the out-of-domain test data sets was used for this hyperparameter optimisation.

3.2 Sentiment analysis

We extracted sentiments via a publicly available key word list [14]. A sentiment vector $\mathbf{s} \in \mathbb{R}^d$ was constructed from the sentiment polarity values in the sentiment dictionary. We compute the

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

[3] The longest statement is 522 characters (65 words) long, the 25%/50%/75% percentiles are 63/95/135 characters or 8/12/17 words, respectively.

sentiment index for attributing positive or negative sentiment to a text as the cosine similarity between BOW vectors and sentiment vector.

3.3 Interpreting bag-of-words models

Interpreting coefficients of linear models (independent of the regulariser used) implicitly assumes uncorrelated features; this assumption is violated by the text data used in this study. Thus direct interpretation of the model coefficients $\mathbf{w}_k$ is problematic, see also [17, 10]. In order to allow for better interpretation of the predictions and to assess which features are discriminative, we computed correlation coefficients between each word and the party affiliation label.

4 Results

The following section gives an overview of the results for all political bias prediction tasks. Predictions compared with the manifesto data were computed using models trained on texts from the 17th Bundestag, predictions obtained for facebook post texts were computed with models trained on the 18th Bundestag[4].

4.1 In-domain predictions

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. We list the evaluation results for the political party affiliation prediction on in-domain data (held-out parliamentary speech text) for the 17th Bundestag in Table 1. These results are comparable to those of [11] who report a classification accuracy of 0.61 on a five class problem predicting party affiliation in the European parliament.

4.2 Out-of-domain predictions

For out-of-domain data obtained from manifesto data, the models yield significantly lower precision and recall values between 0.3 and 0.4, see Table 2. We observe a similar effect for the facebook post data. The short texts resulted in poor prediction accuracies of 0.51 on average. Additionally, classes were highly unbalanced in this setting, since some parties have an order of magnitude more posts than others.

[4]We leveraged the speeches from the 17th legislative period for the first task as this legislature is already completed and offers more data. Results for the 18th Bundestag are similar but omitted for brevity. We employ the speeches of the 18th legislative period for the facebook posts as the posts were more recent.

Table 1: **In-domain classification performance** for data from the 17th legislative period on in-domain data. N denotes number of data points in the evaluation set.

	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

Table 2: **Out-of-domain** classification performance (quasi-sentence level) on **manifesto data** of a classifier trained on speeches of the 17th legislative period of the Bundestag.

	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

4.3 Influence of text length on accuracy

A key factor that made the prediction in the out-of-domain prediction task particularly difficult was the short length of the texts to classify, see also section 2. In order to investigate the effect of text length, we aggregated the data into longer texts, and grouped manifesto data into political topics. Table 3 shows the topic level prediction results. We obtain F1 scores of above 0.8 for all parties except for the SPD . As the facebook posts lacked topic labels, we conducted the aggregation of these texts by first concatenating all facebook posts of a party into one long text; this text was then partitioned into segments of 1000 words each. For each party 50 random segments were selected for classification. The results are shown in Table 4. Prediction accuracies comparable to the in-domain case can also be achieved for these texts. This increase is in line with previous findings on the influence of text length on political bias prediction accuracy [11].

Table 3: **Out-of-domain** classification performance (topic level) on **manifesto data**. Compared to quasi-sentence level predictions (Table 2), the predictions made on the topic level are more reliable.

	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

Table 4: **Out-of-domain** classification performance on 50 randomly selected **facebook posts** of respective party (text length: 1000 words). The average prediction performance is comparable to that on in-domain test data.

	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

4.4 Misclassification and policy change

Automatic political text analysis requires a profound understanding of the models used. One way to better understand these models is to inspect the misclassifications of a model. A potential explanation for the misclassifications could be that parties change their policy positions over time. The confusion matrix for the 17th Bundestag in Table 5 shows that the SPD manifesto texts are often predicted as belonging to the CDU/CSU on the topic level. This was the the legislative period when the CDU under chancellor Merkel was making a strong left move with respect to socioeconomic issues.

4.5 Predicting government status

We also trained a model on government membership labels, in order to a better compare against other studies that predict party affiliation in a two party system. Table 6 shows the results for the 17th legislative period. While the in-domain prediction accuracy is close to 0.9, the out-of-domain evaluation on manifesto data drops again to a performance close to chance. This is in line with results on binary classification of political bias in the Canadian parliament [16]. The authors report classification accuracies between 0.80 and 0.87, and find a pronounced drop in performance on texts from a different domain (e.g. older texts or texts from another chamber). In our results, the aggregation into topics did not increase the accuracy in this binary setting when classifying manifesto texts. The drop in accuracy of the binary classifier on facebook data (aggregated analogous to the party affiliation case) was less pronounced: accuracies were above 0.70.

Table 5: **Topic level confusion matrices** of manifesto texts.

		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

4.6 Discriminative features

Another important question when analysing automatic text classification models is whether the difference between the features of each party stems from different policies or from other aspects of the text. To address this point we analysed features that are discriminative for government membership and for parties.

Sentiment correlates with political power The drop in prediction accuracy in the government prediction task was more pronounced for manifesto texts than for facebook posts. What do facebook posts and plenary debates have in common? In contrast with the authors of manifestos, both the speakers in the parliament as well as the authors of facebook posts know which party is in government. A language feature that might capture this is sentiment. Indeed our results in Table 7 show that positive sentiment strongly correlates with government membership and the number of seats in the parliament. Previous studies also find that text features which are discriminative in a two party system are not necessarily related to policies but more to language of defence and attack [11].

Table 6: Classification accuracy on the binary prediction problem, categorising texts into government and opposition. Out-of-domain accuracy again drops close to chance performance for the manifesto data but remains higher for the facebook post texts.

	In-Domain	**Out-of-Domain**	
	Parliament	Manifestos	Facebook Posts
Accuracy	0.88	0.60	0.76

Table 7: Correlation coefficient between the 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

Correlations between words and parties In order to determine further discriminative features, we quantified which words were preferentially used by each party by measuring the correlation of single words with the party label. Unspecific stopwords were excluded. We find clear differences between the parties, which are in line with the parties ideologies.

Left party (linke) Frequent words include referrals to big companies (*konzerne*) and their profits (*profite*), the working class *beschaeftigte*, the social welfare program *hartz iv* as well as war (*krieg*).

Green party (gruene) Uses words related to environmental damage (*klimaschaedlichen*), exploited low wage employees (*leiharbeitskraefte*) and pensions (*garantierente*).

Social Democratic Party (SPD) Uses mostly unspecific words related to the parliament and governmental processes (*staatssekretaerin, kanzlerin, bundestagsfraktion*) and some words related to cutting of expenses (*kuerzungen*).

Christian Democratic Union/Christian Social Union (CDU/CSU) Often used words relate to a pro-economy attitude, such as competitiveness or (economic) development (*wettbewerbsfaehigkeit, entwicklung*) and words related to security (*sicherheit, stabilitaet*).

5 Conclusions and Limitations

We find that automated political bias prediction is possible with an accuracy better than chance, even beyond the training text domain. These results suggest that such systems could be helpful as assistive technology, for example for human annotators in an active learning setting.

In line with previous findings [16, 11], we find a large effect of text length and text domain on the generalisation performance of the classifier. The first effect, that longer texts are easier to classify, intuitively makes sense. Also humans are challenged when judging the political bias of shorter texts out of context [8]. However, short texts are a realistic challenge for automated political bias prediction systems: political texts from social media data and other web sources are often very short and hence difficult to analyse for both human annotators and algorithms. Both political education and science can benefit from automatic analyses of these very data streams, as these fields have a strong influence on public opinion and yet cannot be analysed by humans alone, due to the volume of data.

The second effect, i.e. the drop in generalisation performance on out-of-domain data, appears to be correlated to the first one: it can be alleviated in some cases by aggregating texts into longer segments. In the case of party affiliation prediction, the out-of-domain classification is on a par or even better than the prediction accuracy on in-domain data. However in the binary classification setting (government membership prediction), text aggregation does not help as much: aggregating manifesto data, written without the knowledge of which party would be member of the government, into longer texts does not counteract the effect of out-of-domain accuracy drop. We attribute this effect in part to the fact that sentiment appears to be a discriminative feature for government membership.

Acknowledgements

We would like to thank Friedrich Lindenberg for factoring out the `https://github.com/bundestag/plpr-scraper` from his Bundestag project. Michael Gaebler provided helpful feedback on an earlier version of the manuscript.

References

[1] Manifesto codebook `https://manifesto-project.wzb.eu/information/documents?name=handbook_v4`.

[2] `https://github.com/bundestag`.

[3] `https://www.bundestag.de/protokolle`.

[4] `https://www.facebook.com/B90DieGruenen/`.

[5] `https://www.facebook.com/CDU/`.

[6] `https://www.facebook.com/linkspartei/`.

[7] `https://www.facebook.com/SPD/`.

[8] Kenneth Benoit, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Rewview*, Forthcoming.

[9] Kenneth Benoit, Michael Laver, and Slava Mikhaylov. Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science*, 53:495–513, 2.

[10] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Biessmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

[11] Graeme Hirst, Yaroslav Riabinin, Jory Graham, and Magali Boizot-Roche. Text to ideology or text to party status? In Isa Maks Bertie Kaal and Annemarie van Elfrinkhof, editors, *From Text to Political Positions: Text analysis across disciplines*, pages 47–70, 2014.

[12] Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Regel, and Annika Werner. *Manifesto Corpus*. WZB Berlin Social Science Center., `https://manifestoproject.wzb.eu/information/documents/corpus`, 2015.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] R. Remus, U. Quasthoff, and G. Heyer. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, 2010.

[15] Jonathan B. Slapin and Sven-Oliver Proksch. Word as data: Content analysis in legislative studies. In Shane Martin, Thomas Saalfeld, and Kaare W. Strøm, editors, *The Oxford Handbook of Legislative Studies*, pages 126–144. Oxford University Press, Oxford, 2014.

[16] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48, 2008.

[17] Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. The feature importance ranking measure. In *ECML/PKDD*, pages 694–709, 2009.