Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

Modelling Cultural and Socio-Economic Dimensions of Political Bias in German Tweets

Aishwarya Anegundi[1] and **Konstantin Schulz**[1] and **Christian Rauh**[2] and **Georg Rehm**[1]
[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
aishwarya.anegundi@dfki.de, konstantin.schulz@dfki.de, georg.rehm@dfki.de
[2] Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin, Germany
christian.rauh@wzb.eu

Abstract

We introduce a new bi-dimensional classification scheme for political bias. In particular, we collaborate with political scientists and identify two important aspects: cultural and socio-economic positions. Using a dataset of tweets by German politicians, we show that the new scheme draws more distinctive boundaries that are easier to model for machine learning classifiers (F1 scores: 0.92 and 0.86), compared to one-dimensional approaches. We investigate the validity by applying the new classifiers to the whole dataset, including previously unseen data from other parties. Additional experiments highlight the importance of dataset size and balance, as well as the superior performance of transformer language models as opposed to older methods. Finally, an extensive error analysis confirms our hypothesis that lexical overlap, in combination with high attention values, is a reliable empirical predictor of misclassification for political bias.

1 Introduction

Political radicalization is linked to a society's sense of insecurity (Bartoszewicz, 2016). Such a feeling may arise especially in times of crisis, such as financial crashes, large migration movements, or pandemics. In this setting, citizens' trust in a country's government or into the political system more generally can decline quickly (Easton, 1975; Dostal, 2015), leading to further radicalization.

The effects of such a development are visible not only in terms of elections (Funke et al., 2016; Recuero et al., 2020) and media coverage (Bender et al., 2021), but also in general public political discourse and corresponding language use: Politically biased texts tend to exhibit a wording that is different from their neutral counterparts (Krestel et al., 2012; Fairbanks et al., 2018). At times, this lexical deviation is hard to detect because the texts are positioned in seemingly neutral environments like technological or scientific sections of a newspaper (Kang and Yang, 2022). Furthermore, there are additional factors beyond wording: The filtering and selection of information to be presented in a given spot is a bias in its own right, but can directly affect or reflect political discourse: Presenting quotes by famous hyperpartisan politicians often serves as a subtle disguise for an author's own political motives (Fan et al., 2019). Besides, the media coverage of political parties or crime-related ethnical aspects is indicative of the current government, the popularity of specific parties (Lazaridou and Krestel, 2016) and the trust in the executive's impartiality (Pfeiffer et al., 2018).

By training language models on such tendentious texts, we tend to reproduce and spread their bias (Bender et al., 2021), even if the resulting models are used in rather neutral contexts (Liu et al., 2021). Since political bias (PB) is closely related to credibility (Su et al., 2020; Vargas et al., 2020; Aksenov et al., 2021; DeVerna et al., 2021; Saltz et al., 2021) and trustworthiness (Viviani and Pasi, 2017), such language models will suffer from reduced acceptance and utility unless we can reasonably detect and decrease their bias. The same applies to traditional media content: There is no way to holistically analyze media credibility without considering the PB of respective outlets. Thus, we make the following contributions:

- We introduce a new classification scheme for PB adapted to recent insights of political science.
- Using the Polly corpus (De Smedt and Jaki, 2018), a dataset of German tweets, we train and evaluate transformer-based classifiers with our new scheme. Polly corpus does not provide the labels with respect to political dimension; instead provides a political party

label. Although there are large annotated datasets incorporating fine-grained schemes for parliament speeches and interviews (Blätte and Blessing, 2018; Rauh and Schwalbach, 2020), there are none for social media such as Twitter. Hence we use party affiliations as a proxy for the dimensions. We represent the extremes of cultural dimension with political parties *Grüne* and *AfD* and the extremes of socio-economic dimension with *Die Linke* and *FDP*.

- Using the classifiers, we test four hypotheses:
  1. The current one-dimensional schemes are overly simplistic models of PB. Integrating socio-economic and cultural dimensions of political conflict is more effective for classifying PB.
  2. Adding more data and balancing the dataset leads to better PB classification results.
  3. Misclassified texts often exhibit lexical overlap with the opposing end of the respective dimension.
  4. In misclassified texts, words from the opposing end of the respective dimension receive high attention from the transformer model.

We make our source code[1] and models[2] publicly available. In the following, we describe our conceptual model of PB, the annotations in the dataset and the architecture of our classifiers, as well as their training and the corresponding evaluation.

2 Related Work

Previous machine learning approaches to PB detection have mostly conceptualized it as binary text classification: Given an input text, the algorithm assigns a label indicating the presence or absence of PB. Similarly, the binary choice can also be used to model the direction of bias on continuous scales (Iyyer et al., 2014; Fairbanks et al., 2018; Liu et al., 2021), moving the desired outputs closer to seminal applications of text-based ideological scaling in the political sciences (Laver et al., 2003; Slapin and Proksch, 2008; Rheault and Cochrane, 2020; Sältzer, 2022).

As in many cases of language modeling, binary decisions are easy to set up and learn. On the downside, they do not properly reflect all nuances of complex concepts like PB. That is why some approaches use more fine-grained classification schemes: They extend the left-right spectrum to incorporate more intermediate positions (Aksenov et al., 2021) or reuse datasets that originally proceeded this way (Fairbanks et al., 2018). Such advanced schemes may be more accurate than the simple binary models, but are also harder to annotate. In cases where this kind of data does not yet exist, many researchers fall back to using other documented phenomena as proxies for PB: Preference of specific political parties (Krestel et al., 2012; Kang and Yang, 2022), membership in such parties (Iyyer et al., 2014) and social interactions of the authors on Twitter (Li and Goldwasser, 2019) are prominent examples in that regard.

All in all, existing computational approaches to PB detection are still mostly one-dimensional, thereby reducing the political conflict to a single 'left-right' dimension. In political science, however, there is a growing agreement that political conflict is at least two-dimensional. The conventional left-right dimension comprising of socio-economic preferences regarding the relative power of markets and the state is increasingly complemented by a separate 'cultural' dimension of political conflict (Hooghe et al., 2002; Kriesi et al., 2008; Bornschier, 2010; Zürn and de Wilde, 2016; Lengfeld and Dilger, 2018). This dimension captures disagreements on culturally 'liberal' versus 'conservative' value orientations, compounding political stances on the openness of borders, migration, minority protection, environmentalism, or gender and sexuality questions. This two-dimensional structure has been shown to map onto political competition among partisan elites (Kriesi et al., 2008) and is also reflected in attitudes and vote intentions of citizens (Lucassen and Lubbers, 2012; Lengfeld and Dilger, 2018; Norris and Inglehart, 2019).

3 Methodology

This section discusses our conceptual model of PB, and different ways of classifying PB, followed by methods used to explain cases of misclassification.

Conceptual Model: To provide a more sophisticated model of PB, we follow recent insights from the field of political science and abandon the overly simple one-dimensional perspective. Instead, we

[1] https://github.com/konstantinschulz/political-bias-classification

[2] https://live.european-language-grid.eu/catalogue/tool-service/18689

use a two-dimensional approach aimed at capturing both socio-economic and cultural conflict lines. Unfortunately, to our knowledge, there is no dataset of German texts with readily available aggregate annotations on these two dimensions. Therefore, we use party affiliation as a proxy for the two dimensions. The intuition is that certain political parties in Germany represent the extremes on each of the two separate dimensions. This assumption is consistent with extant party-classification schemes in the political sciences (Polk and Rovny, 2017; Volkens et al., 2021) and is a common makeshift solution in PB classification suffering from annotation scarcity.

Domain and Register: We build on previous work analyzing social media because this forum of public discourse is known to be associated with PB (Badjatiya et al., 2019; Li and Goldwasser, 2019; Recuero et al., 2020) and corresponding disinformation (Gallotti et al., 2020; Keller et al., 2020; Sharma et al., 2020; Zhou et al., 2020; DeVerna et al., 2021; Mattern et al., 2021; Weinzierl and Harabagiu, 2021). This decision has important consequences for our trained classifiers: They will be well-adjusted to the short, rather colloquial texts on social media, but may fail when confronted with more formal registers and longer texts. The key challenge here is domain divergence (Kashyap et al., 2021), which we cannot reliably address without having access to multiple comparable datasets. Considering the political science work on correspondences between social media communication and parliamentary behavior of politicians (Silva and Proksch, 2021; Sältzer, 2022), one step into this direction would be the application of our classifiers to German parliamentary speeches (similar to the approach by Krestel et al., 2012). In that case, the domain would still be political, but the register drastically differs. We plan to evaluate this setup in future studies.

Classification: As a baseline, we chose to encode the tweets using FastText embeddings (Bojanowski et al., 2017) and train traditional machine learning (ML) models. FastText embeddings are learned with a method built on top of the continuous skip-gram model (Mikolov et al., 2013) overcoming the limitation of assigning a different vector for every word of the vocabulary by considering sub-word information. Hence, FastText embeddings perform better for morphologically rich languages like German and are suitable for our classification problem. We obtain the FastText embeddings for each word in the tweet, average them and feed them into ML models. We train different classifiers based on Random Forests (Breiman, 2001), Logistic Regression (Cox, 1958), Multi-Layer Perceptrons (MLP, Ramchoun et al., 2016), and Support Vector Machines (SVM, Cortes and Vapnik, 1995) with a linear kernel. Random Forest is an ensemble classification algorithm whose output is based on predictions of several decision trees constructed at training time. The Logistic Regression algorithm classifies a data point by computing log-odds on the linear combination of independent variables. MLP is a simple feed-forward neural network trained with backpropagation. SVMs construct a hyperplane in a high-dimensional space separating the two classes. The location of the data points on either side of the hyperplane determines their class.

FastText embeddings only incorporate distributional semantic relations between words but fail to consider the context of a word in a sentence, such as word order. We use transfer learning from pre-trained language models such as GBERT (Chan et al., 2020) to overcome this limitation. We chose GBERT-base model for our classification task due to the limited amount of data. GBERT has the same architecture as BERT (Devlin et al., 2019), but it is pre-trained on a large German corpus and has achieved impressive performance on various natural language processing tasks. The architecture of BERT is based on the multi-layer bidirectional transformer encoder with a multi-head attention mechanism (Vaswani et al., 2017). The base version consists of 12 layers, a hidden size of 768, and 12 attention heads, making up $110M$ parameters.

Error Analysis: For the error analysis, we are mainly interested to find out how well the model can learn the data distribution. Hence, we analyze attention scores (hypothesis 4) as an approximation of token importance (Wiegreffe and Pinter, 2019; Tutek and Šnajder, 2020), in combination with association scores (hypothesis 3) derived from the dataset. To identify the most important words associated with a particular class, we use a custom **word importance** WI metric which includes Pointwise Mutual Information (PMI) and Term Frequency–Inverse Document Frequency (TF-IDF), weighted by relative word frequency. Both measures have been shown to be useful approximations of association strength (Bouma, 2009; Krestel et al.,

2012; Fan et al., 2019). The distance between association scores for different classes gives higher scores to the words frequent in one class and infrequent in the opposite class. Normalizing by relative word frequency helps us avoid high scores for words with rare occurrences. The formula is

$$WI(c, w) = (\alpha(c, w) - \alpha(\hat{c}, w)) \cdot f(c, w) \quad (1)$$

where $\hat{c}$ is the opposing class, α is either PMI or TF-IDF and $f(c, w)$ is the relative frequency of w within class c. We create two vocabularies for each class consisting of important words, one identified with the WI metric using PMI as α (PMI vocabulary) and the other using TF-IDF as α (TF-IDF vocabulary). Furthermore, we compute attention scores for each word in the tweet, summing up the attention scores for all sub-tokens forming the word. We average the attention score over all the attention heads across all the layers.

To verify hypothesis 3, we analyze the percentage of confusing words in each tweet. A word is confusing if $WI(c, w) - WI(\hat{c}, w)$ is positive, indicating that the word is more important in the opposite end of the dimension. We analyze the amount of tweets above a certain threshold percentage of confusing words and examine how this number changes for varying minima. We compare the ratio of wrong and correct predictions for each threshold to confirm the hypothesis. Further, to verify the hypothesis 4, we rank the confusing words according to the magnitude of $WI(c, w) - WI(\hat{c}, w)$ and check if the topmost confusing words receive the highest attention from the model. Again, we compare the ratio of false and correct predictions to confirm the hypothesis. We repeat the process for the vocabularies in both dimensions.

4 Experiments

Dataset: We trained our classification models on a subset of the Polly corpus (De Smedt and Jaki, 2018). The corpus focuses on the 2017 German Federal Election and consists of 125K tweets collected from August 2017 to December 2017. It comprises seven subgroups denoting tweets by fans, by politicians, about politicians, containing the phrase *ist ein* ("is a"), hate speech, emojis, and random tweets. In our study, we used the subset containing tweets by politicians also denoted as "By-Party" currently in their Google Sheet[3]. Each tweet in the By-Party subset also provides metadata such as likes, timestamps, names of the politicians, and their associated political parties. The By-Party subset has about 14.2K tweets from seven different parties: CDU, CSU, SPD, Die Linke, Die Grünen, FDP, and AfD. With respect to gender, it contains tweets from 13 female and 22 male politicians selected based on their popularity.

Following extant party-classification schemes in the political sciences (Polk and Rovny, 2017; Volkens et al., 2021) we exploit the following party labels. For the dimension capturing conflict between culturally liberal and conservative stances, we consider tweets from Die Grünen (the rather cosmopolitan German Green party) and the Alternative für Deutschland (AfD, a populist far-right party) as representations of the most extreme stances. We anchor the socio-economic left-right dimension on tweets from Die Linke (a far-left party) and the FDP (taking market-liberal stances). This results in about 4.5K tweets for each dimension. The data distribution for the socio-economic dimension is $1.96k$ tweets for Die Linke and $2.52k$ tweets for FDP (Die Linke = 43.82%, FDP = 56.7%). Similarly, the distribution for the cultural dimension is $2.16k$ tweets for Die Grünen and $2.4k$ tweets for AfD (Die Grünen = 47.33%, AfD = 52.66%). Given the limited data points, we split the collection of tweets into train and test data at a 90:10 ratio. We then preprocess the tweets to remove mentions, URLs and the retweet string "RT @mention". While we retain the emoticons for the classification using the BERT model, we remove them for the FastText embeddings because FastText does not contain meaningful embeddings for them. We always downsample the majority class to achieve class balancing before training the model.

Baseline Model: For our classification task, we download the 300-dimensional pre-trained vectors for the German language[4], provided by Facebook[5] to initialize the FastText model using the Gensim library[6]. We normalize and tokenize the tweets using the ICU-Tokenizer[7]. To obtain the final embedding, we average the FastText word embeddings of each token in the tweet. The resulting vectors are used to train the ML classifiers with the scikit-learn library.

[3] https://docs.google.com/spreadsheets/d/1c5peNMjt24U0FcEMSj8gD_JjzumqXTWbPWa_yb2nNt0/edit. URLs were all last accessed on 2022-06-09.

[4] https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.de.zip

[5] https://fasttext.cc/docs/en/pretrained-vectors.html

[6] https://radimrehurek.com/gensim/models/fasttext.html

[7] https://github.com/mingruimingrui/ICU-tokenizer

The Random Forest classifier is trained with the Gini criterion with 100 trees as estimators. The MLP classifier comprises 12 layers and is trained with the Adam optimizer, ReLU activation and early stopping. We use a linear kernel for the SVM classifier and Stochastic Average Gradient solver for the Logistic Regression.

GBERT Model: We fine-tune the GBERT-base model on the Polly By-Party subcorpus using the HuggingFace transformers library[8]. Before fine-tuning, we tokenize the tweets using the AutoTokenizer for GBERT from the same library. The GBERT model encodes the tweets, and these encodings are fed into an output feed-forward network, followed by a softmax layer. This is achieved by using the AutoModelForSequenceClassification class from the transformers library. We train the model with the AdamW optimizer, with a learning rate of 5e-5 and a batch size of 8 for five epochs.

5 Results

Classification: Tables 1 and 2 show the accuracy, micro-averaged precision, recall and F1 scores for different classification models over cultural and socio-economic dimensions. We use micro-averaging for the evaluation to be consistent with our additional experiments on class imbalance (see below). GBERT-base performs best for both dimensions, although the performance is much higher for the cultural dimension with 92% accuracy than for the socio-economic dimension with 86%. The better performance of GBERT in comparison to ML algorithms can be explained by the fact that GBERT has been pre-trained on large German text corpora. Besides, it takes into consideration the context of a word in both directions. Its large number of parameters enables it to model a complex underlying function. All the ML algorithms perform the same, more or less, and the varying model sizes can explain the slight differences. In contrast, the GBERT model trained on a traditional left-right dimension with Die Linke on the left end and AfD on the right end of the spectrum as proxies has an accuracy of 87.02% (micro F1 = 86.4%). Hence, deviating from the traditional one-dimensional approach leads to higher classification performance, supporting our hypothesis 1.

Table 3 shows the results of the GBERT model trained with reduced data for balanced and unbalanced scenarios. For both dimensions, the model's performance reduces when trained with half the data, supporting hypothesis 2. We can see that the majority class (FDP) is easier to classify for the socio-economic dimension. Hence, the accuracy drops after balancing. Meanwhile, for the cultural dimension, both classes are equally hard to classify, and increasing the relative importance of the minority class (Die Grünen) through balancing leads to a slight increase of overall accuracy. We hypothesize that, after the balancing intervention, the model uses a larger share of its weights and biases to model the (former) minority class, which increases the performance for that class.

Application: We apply the two trained classifiers to the whole dataset (see Figure 1). Each tweet gets a cultural and a socio-economic score. The score for a specific party is the average of all its associated tweets. We observe that, as expected, the four proxy parties (AfD, FDP, Die Grünen, Die Linke) are close to the respective extreme of the dimension that they represent. Interestingly, these proxy parties form two pairs: The distance from the Left to the Green party is smaller than to the liberal or conservative party. The same goes for the liberal party, which has a small distance to the conservative party, as opposed to the Left or Green. Finally, we note that most parties are situated in the lower left quadrant (open, socialist), while the remaining two occupy outlier places (liberal and/or conservative). This could be an indication of political isolation. However, the dataset is a sample of just a few dozen politicians with a moderate bias regarding the distribution of gender, and possibly age or other important factors. Thus, our results

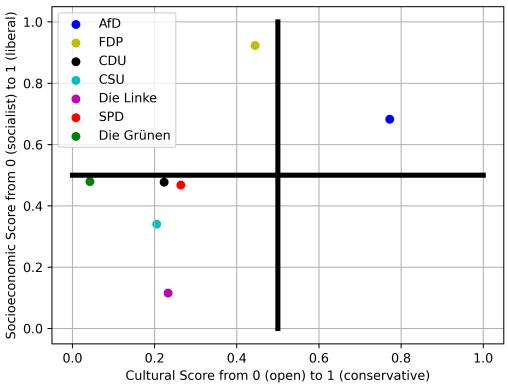

Figure 1: Cultural and Socio-economic Scores of German Political Parties

[8] https://huggingface.co/bert-base-german-cased

are not necessarily representative of each party as a whole. Instead, they can serve as general tendency that needs to be investigated more thoroughly in future studies.

TF-IDF Vocabulary: Figures 2 and 3 show the percentage of tweets consisting of a minimum number of confusing words (threshold) for the TF-IDF vocabulary. For the cultural dimension (Figure 2), we can infer that, on average, 10.6% more tweets meeting the threshold are misclassified, compared to the correct predictions. Although not consistent over all the thresholds, we see similar behavior (Figure 3) for the socio-economic dimension, between the 10% and 35% thresholds, with 1.3% more tweets meeting the threshold and being misclassified, compared to the correct predictions on average. Furthermore, misclassified tweets make up a larger share of the dataset (+19.3%) compared to the correctly classified ones, with at least one confusing word receiving the highest attention for the cultural dimension (Figure 4). We see a different behavior when we consider only a few of the top confusing words up to a minimum of 30%, after which the trend reverses. The same trend emerges for the socio-economic dimension (see Appendix A) when we consider at least the top 25% of confusing words. The behavior is not as strong as in the cultural dimension, with only 2% of wrong predictions consisting of a confusing word receiving highest attention in comparison to 1.5% for the correct predictions. Some lexical examples of commonly confused words in a TF-IDF vocabulary are as in Table 6.

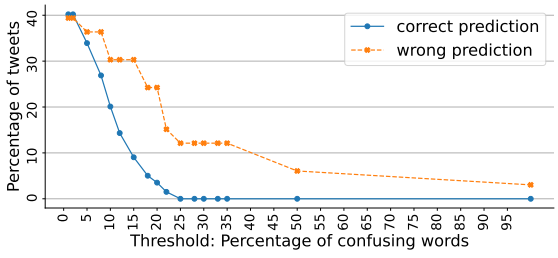

Figure 2: Percentage of wrong predictions and correct predictions for varying thresholds of confusing words computed using the TF-IDF vocabulary for the cultural dimension.

PMI Vocabulary: Analogous to our analysis using TF-IDF, we also observe the variation in the percentage of wrong and correct predictions for the PMI vocabulary. For the cultural dimension (Appendix A), at any given threshold, the percentage of misclassified tweets meeting the threshold exceeds the correctly classified tweets by 11.6% on average. For the socio-economic dimension, we observe the same trend up to the 28% threshold, with wrong predictions meeting the threshold exceeding the correct predictions by 5.31% on average. Also, similar to the TF-IDF vocabulary, on average, 12.7% more misclassified tweets than correct ones in the cultural dimension includes at least one confusing word that receives the highest attention (Figure 5). We observed the same trend when considering only a few of the top confusing words. The behavior is not so evident for the socio-economic dimension, with wrong predictions constituting only 2% more than correct predictions on average. The trend reverses when we consider more than 55% of the top confusing words (Appendix A). For lexical examples of commonly confused words in a PMI vocabulary see Table 6.

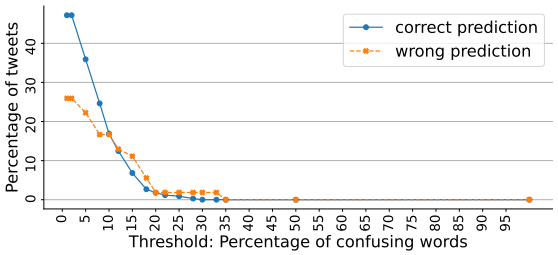

Figure 3: Percentage of wrong predictions and correct predictions for varying thresholds of confusing words computed using the TF-IDF vocabulary for the socio-economic dimension.

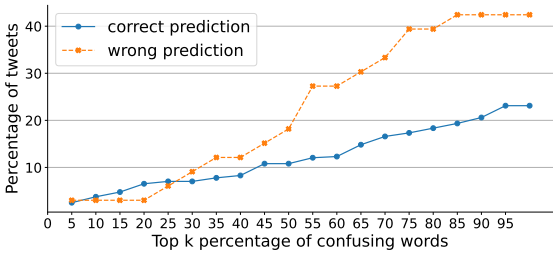

Figure 4: Percentage of tweets consisting of a confusing word receiving the highest attention from the model for the cultural dimension with the TF-IDF vocabulary.

For both TF-IDF vocabulary and PMI vocabulary, hypothesis 3 holds for the cultural dimension over all the thresholds. In contrast, hypothesis 4 is confirmed with a larger margin for the PMI vocabulary compared to the TF-IDF vocabulary (Figures 4 and 5). For the socio-economic dimension, hy-

Model	Accuracy	Precision	Recall	F1
GBERT-base	**0.92**	**0.93**	**0.92**	**0.92**
Logistic Regression	0.80	0.81	0.80	0.80
SVM	0.83	0.83	0.83	0.83
Random Forests	0.81	0.81	0.81	0.81
MLP	0.82	0.82	0.82	0.82

Table 1: Comparative evaluation of classification: GBERT-base with ML classifiers for the cultural dimension (Die Grünen vs. AfD) on Polly test data.

Model	Accuracy	Precision	Recall	F1
GBERT-base	**0.86**	**0.89**	**0.83**	**0.86**
Logistic Regression	0.68	0.68	0.68	0.67
SVM	0.71	0.71	0.71	0.71
Random Forests	0.73	0.73	0.73	0.73
MLP	0.70	0.70	0.70	0.69

Table 2: Comparative evaluation of classification: GBERT-base with ML classifiers for the socio-economic dimension (Die Linke vs. FDP) on Polly test data.

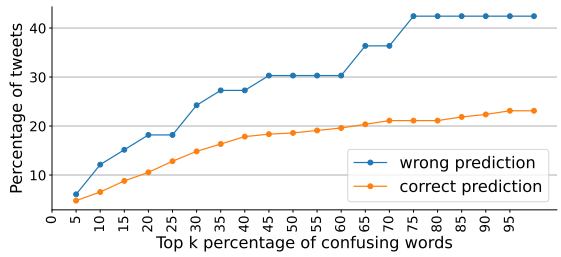

Figure 5: Percentage of tweets consisting of a confusing word receiving the highest attention from the model for the cultural dimension with the PMI vocabulary.

pothesis 3 holds over a specific range of thresholds only, although the distinction is more explicit in the PMI vocabulary than in the TF-IDF vocabulary. Similarly, the PMI vocabulary shows a clearer difference between wrong and correct predictions for hypothesis 4 than the TF-IDF vocabulary. Furthermore, hypothesis 4 holds when we consider more confusing words for the TF-IDF vocabulary in contrast to fewer confusing words in the case of the PMI vocabulary for the socio-economic dimension (Appendix A).

6 Conclusions

We have shown that PB can be reliably analyzed in two dimensions. In particular, we follow recent insights from political science and abandon one-dimensional scales like 'left vs. right'. Instead, we use separate dimensions for cultural and socio-economic conflict lines to model different aspects of PB. Due to a lack of appropriately annotated datasets for this new scheme, we use party affiliation as a proxy for the dimensions: The German political parties *Grüne* and *AfD* represent different extremes of the cultural dimension, while *Die Linke* and *FDP* span up the socio-economic conflict line. We use GBERT to train separate binary classifiers for tweets by each of those parties' members, showing that the cultural distinction is easier to model in our setup. In both cases, the deep learning approach is superior to other ML baselines like SVM or Random Forests.

We conduct additional experiments to explain classification errors. The classifiers struggle when many words from the opposing political spectrum are used and receive high attention by the transformer model. This is particularly true for the cultural dimension, but only partially for the socio-economic cleavage. We hypothesize that, in the latter case, the language use of the different parties is more similar to each other, blurring the lexical boundaries and thus reducing the risk of classification errors based solely on the presence of specific words. This may be related to a long-standing political science debate on position- vs. salience-based party competition (Dolezal et al., 2014): in the former perspective, parties compete with different stances on the same topics, which would mean that they share a high number of words. In the latter perspective, parties compete by emphasizing different topics, which should be related to greater lexical diversity across tweets from different parties.

Dimension	Data Distribution (%)			F1		Accuracy
socio-economic		Die Linke	FDP	Die Linke	FDP	
	unbalanced	43.82	56.7	0.825	0.861	0.845
	balanced	50	50	0.841	0.833	0.837
cultural		Die Grünen	AfD	Die Grünen	AfD	
	unbalanced	47.33	52.66	0.884	0.894	0.889
	balanced	50	50	0.898	0.897	0.897

Table 3: Evaluation of the GBERT model trained on only half of the Polly train data. For each dimension, we see the model's performance in balanced and unbalanced setups indicated by per-class F1 score and overall accuracy. The two classes for each dimension are the two extremes of the dimension represented by political parties.

In terms of future work, we plan to evaluate our classifiers on other datasets of political language, such as extant collections of German parliamentary speeches (Blätte and Blessing, 2018; Rauh and Schwalbach, 2020). Besides, we need to empirically explore possible reasons for the different classification performance in our two dimensions. Furthermore, creating new annotations specifically for our proposed model of PB would enable researchers to train classifiers with a higher construct validity. Finally, while our bi-dimensional scheme for PB detection is better than the single-dimensional scheme, exploring other dimensions is worthwhile following new political science research.

Acknowledgments

The research presented in this paper is funded by the German Federal Ministry of Education and Research (BMBF) through the project PANQURA (grant no. 03COV03E).

References

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131, Online. Association for Computational Linguistics.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. *The World Wide Web Conference on - WWW '19*, pages 49–59.

Monika Gabriela Bartoszewicz. 2016. Festung Europa: Securitization of migration and radicalization of European Societies. *Acta Universitatis Carolinae Studia Territorialia*, 16(2):11–37.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Andreas Blätte and Andre Blessing. 2018. The GermaParl Corpus of Parliamentary Protocols - ACL Anthology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association (ELRA).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Simon Bornschier. 2010. *Cleavage Politics and the Populist Right: The New Cultural Conflict in Western Europe*. Temple University Press.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

C. Cortes and V. Vapnik. 1995. Support vector networks. *Machine Learning*, 20:273–297.

David R. Cox. 1958. The regression analysis of binary sequences (with discussion). *J Roy Stat Soc B*, 20:215–242.

Tom De Smedt and Sylvia Jaki. 2018. The Polly corpus: Online political debate in Germany. In *Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)*, pages 33–36, Antwerp.

Matthew R DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden. 2021. CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)*, pages 992–999, Virtual. AAAI.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Dolezal, Laurenz Ennser-Jedenastik, Wolfgang C. Müller, and Anna Katharina Winkler. 2014. How parties compete for votes: A test of saliency theory. *European Journal of Political Research*, 53(1):57–76. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-6765.12017.

Jörg Michael Dostal. 2015. The Pegida movement and German political culture: Is right-wing populism here to stay? *The Political Quarterly*, 86(4):523–531.

David Easton. 1975. A Re-Assessment of the Concept of Political Support. *British Journal of Political Science*, 5(4):435–457.

James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility assessment in the news: Do we need to read? In *Proc. of the MIS2 Workshop Held in Conjunction with 11th Int'l Conf. on Web Search and Data Mining*, pages 1–8, Marina Del Rey. ACM.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. *arXiv preprint arXiv:1909.02670*.

Manuel Funke, Moritz Schularick, and Christoph Trebesch. 2016. Going to extremes: Politics after financial crises, 1870–2014. *European Economic Review*, 88:227–260.

Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour*, 4(12):1285–1293.

Liesbet Hooghe, Gary Marks, and Carole J. Wilson. 2002. Does Left/Right Structure Party Positions on European Integration? *Comparative Political Studies*, 35(8):965–989. Publisher: SAGE Publications Inc.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122.

Hyungsuc Kang and Janghoon Yang. 2022. Quantifying perceived political bias of newspapers through a document classification technique. *Journal of Quantitative Linguistics*, 29(2):127–150.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. Domain Divergences: A Survey and Empirical Analysis. *arXiv:2010.12198 [cs]*.

Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2020. Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*, 37(2):256–280.

Ralf Krestel, Alex Wall, and Wolfgang Nejdl. 2012. Treehugger or Petrolhead? Identifying bias by comparing online news articles with political speeches. In *Proceedings of the 21st International Conference on World Wide Web*, pages 547–548.

Hanspeter Kriesi, Edgar Grande, Romain Lachat, Martin Dolezal, Simon Bornschier, and Timotheos Frey. 2008. *West European Politics in the Age of Globalization*. Cambridge University Press, Cambridge.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting Policy Positions from Political Texts Using Words as Data. *The American Political Science Review*, 97(2):311–331.

Konstantina Lazaridou and Ralf Krestel. 2016. Identifying political bias in news articles. *Bulletin of the IEEE TCDL*, 12(2).

Holger Lengfeld and Clara Dilger. 2018. Kulturelle und ökonomische Bedrohung. Eine Analyse der Ursachen der Parteiidentifikation mit der „Alternative für Deutschland" mit dem Sozio-oekonomischen Panel 2016: Cultural and Economic Threats. A Causal Analysis of the Party Identification with the "Alternative for Germany" (AfD) using the German Socio-Economic Panel 2016. *Zeitschrift für Soziologie*, 47(3):181–199.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks for political perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating Political Bias in Language Models through Reinforced Calibration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14857–14866.

Geertje Lucassen and Marcel Lubbers. 2012. Who fears what? Explaining far-right-wing preference in Europe by distinguishing perceived cultural and economic ethnic threats. *Comparative Political Studies*, 45(5):547–574.

Justus Mattern, Yu Qiao, Elma Kerz, Daniel Wiechmann, and Markus Strohmaier. 2021. FANG-COVID: A New Large-Scale Benchmark Dataset for Fake News Detection in German. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 78–91, Dominican Republic. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Pippa Norris and Ronald Inglehart. 2019. *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*. Cambridge University Press. Google-Books-ID: I8aGDwAAQBAJ.

Christian Pfeiffer, Dirk Baier, and Sören Kliem. 2018. Zur Entwicklung der Gewalt in Deutschland. Schwerpunkte: Jugendliche und Flüchtlinge als Täter und Opfer. Zentrale Befunde eines Gutachtens im Auftrag des Bundesministeriums für Familie, Senioren, Frauen und Jugend (BMFSFJ). Technical report, Zürcher Hochschule für Angewandte Wissenschaften, Zürich.

Jonathan Polk and Jan Rovny. 2017. Anti-Elite/Establishment Rhetoric and Party Positioning on European Integration. *Chinese Political Science Review*, pages 1–16.

Hassan Ramchoun, Youssef Ghanou, Mohamed Ettaouil, and Mohammed Amine Janati Idrissi. 2016. Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(1):26–30.

Christian Rauh 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.

Raquel Recuero, Felipe Bonow Soares, and Anatoliy Gruzd. 2020. Hyperpartisanship, disinformation and political conversations on Twitter: The Brazilian presidential election of 2018. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 569–578.

Ludovic Rheault and Christopher Cochrane. 2020. Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*, 28(1):112–133. Publisher: Cambridge University Press.

Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. 2021. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School Misinformation Review*, 2(5):1–25.

Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. 2020. Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv preprint arXiv:2003.12309*.

Bruno Castanho Silva and Sven-Oliver Proksch. 2021. Politicians unleashed? Political communication on Twitter and in parliament in Western Europe. *Political Science Research and Methods*, pages 1–17. Publisher: Cambridge University Press.

Jonathan Slapin and Sven-Oliver Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective. *Natural Language Processing Research*, 1(1-2):1–13.

Marius Sältzer. 2022. Finding the bird's wings: Dimensions of factional conflict on Twitter. *Party Politics*, 28(1):61–70. Publisher: SAGE Publications Ltd.

Martin Tutek and Jan Šnajder. 2020. Staying True to Your Word: (How) Can Attention Become Explanation? *arXiv preprint arXiv:2005.09379*.

Luis Vargas, Patrick Emami, and Patrick Traynor. 2020. On the detection of disinformation campaign activity with network analysis. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, pages 133–146, Virtual. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Marco Viviani and Gabriella Pasi. 2017. Credibility in social media: Opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 7(5):e1209.

Andrea Volkens, Tobias Burst, Werner Krause, Pola Lehmann, Theres Matthieß, Sven Regel, Bernhard Weßels, Lisa Zehnter, and Wissenschaftszentrum Berlin Für Sozialforschung (WZB). 2021. Manifesto Project Dataset. Type: dataset.

Maxwell A. Weinzierl and Sanda M. Harabagiu. 2021. Automatic Detection of COVID-19 Vaccine Misinformation with Graph Link Prediction. *arXiv:2108.02314 [cs]*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3205–3212, Virtual Event Ireland. ACM.

Michael Zürn and Pieter de Wilde. 2016. Debating globalization: cosmopolitanism and communitarianism as political ideologies. *Journal of Political Ideologies*, 21(3):280–301.

A Detailed Results

In this section, we provide additional plots and information that further strengthen the discussions provided in the main paper.

A.1 Error Analysis

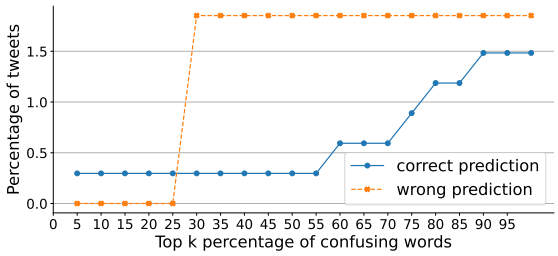

Figure 6: Percentage of tweets consisting of a confusing word receiving the highest attention from the model for the socio-economic dimension with the TF-IDF vocabulary.

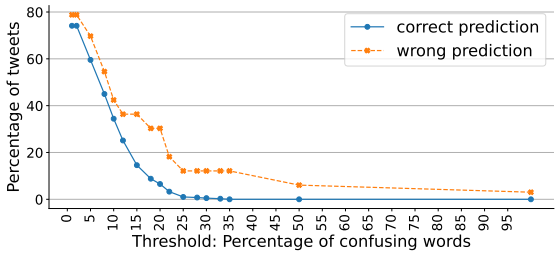

Figure 7: Percentage of wrong predictions and correct predictions with varying thresholds of confusing words computed using the PMI vocabulary for the cultural dimension.

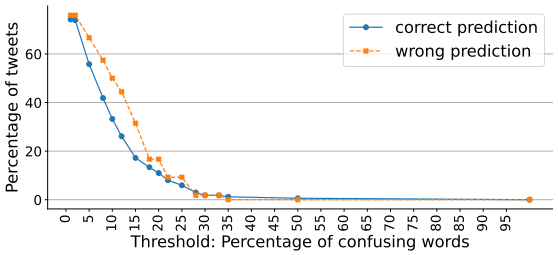

Figure 8: Percentage of wrong predictions and correct predictions with varying thresholds of confusing words computed using the PMI vocabulary for the socio-economic dimension.

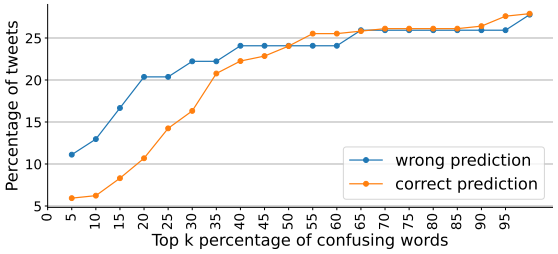

Figure 9: Percentage of tweets consisting of a confusing word receiving the highest attention from the model for the socio-economic dimension with the PMI vocabulary.

Die Linke	FDP	Die Grünen	AfD
btw17	cl	darumgruen	afd
heute	tl	darumgrün	traudichdeutschland
linke	btw17	btw17	btw17
mehr	denkenwirneu	get	merkel
merkel	fdp	heute	mehr
spd	jamaika	mehr	zeit
menschen	beer	katrin	wer
cdu	heute	geht	fdp
müssen	mal	klimaschutz	eu
soziale	mehr	jamaika	morgen

Table 4: Top 10 important words based on WI with TF-IDF as α.

Die Linke	FDP	Die Grünen	AfD
linke	fdp	klimaschutz	afd
soziale	netzdg	kohleausstieg	traudichdeutschland
merkel	cl	sondierungen	dr
btw17	tl	bdk17	merkel
gerechtigkeit	sondierung	sondierung	guten
cdu	kurdistan	umwelt	bitte
spd	freut	jamaika	grenzen
arbeit	denkenwirneu	zukunft	spitzenkandidatin
menschen	digitalisierung	grün	bundestag
rente	trendwende	klima	zeit

Table 5: Top 10 important words based on WI with PMI as α.

TF-IDF as α		PMI as α	
Cultural	**Socio-Economic**	**Cultural**	**Socio-Economic**
zeit	mal	btw17	btw17
statt	bt	mehr	mal
fdp	geht	statt	mehr
mal	ab	zeit	müssen
berlin	dank	mal	warum
merkel	klar	gibt	jamaika
ganz	besser	jamaika	menschen
immer	genau	fdp	eu
politik	interview	politik	wohl
warum	bildung	merkel	brauchen

Table 6: Examples of some commonly confused words for each dimension.