LaTeX Template for NLP Final Project Report

Tom Arend
t.arend@phd.hertie-school.org

Nicolai Berk
nicolai.berk@gmail.com

Abstract

The measurement of ideology is one of the major applications of text analysis in political science. However, researchers often face scarcity of available labelled data to train supervised models for their specific domain. Manual annotation is costly and often severely affected by subjective bias. We propose the use of cross-domain learning to fine-tune transformer models on available, labelled political texts issued by political parties to obtain a classifier of political ideology. Using a unique dataset of newspaper articles authored by politicians, we test such an application in the German context. Comparing transformer neural networks fine-tuned on different data sets of party press releases, we present evidence on the feasibility of such an approach. This contributes to the broad literature on text analysis in the political domain, informing researchers on the limitations of training powerful deep learning models on political language with scarce training data. The experiments conducted by the authors indicate that the training on party labels cannot be easily translated to the domain of newspaper articles but that supervised models are highly sensitive to minor changes in the inputs. It is thus recommended to refrain from using deep learning in the absence of sufficient training data[1].

1. Introduction

The measurement of ideology and political bias[2] are the subject of much research on political texts [2, 11, 20]. Despite significant advances in our understanding and detection of ideology, most researchers still face significant constraints when working with text. Notably, they face a lack of appropriately labelled training data for supervised learning. This is linked to the high costs incurred by manually annotating a significant number of speeches, texts, or sentences. Often, the detection of ideological bias might be highly dependent on the coders' subjective assessment. Instead, researchers could train models on other sources of text with clear and available labels and subsequently apply them to the desired texts using cross-domain learning.

We conduct experiments with fine-tuned deep learning models to understand if and how cross-domain learning can be used to measure bias in the absence of abundant training data. We fine-tune a deep neural network to predict the authoring political party of German press releases. Once the model predicts the authoring party of a press releases sufficiently well, we apply it to estimate the bias of newspaper articles. To determine the best model we conduct experiments with different hyperparameter settings and compare different layer structures. While the application of transformers to measure bias is not entirely new in political science [23], we want to move beyond the current state-of-the-art by measuring the precise implications of different fine-tuning procedures.

Beyond testing the effectiveness of this two-step fine-tuning process, our project seeks to contribute to ongoing efforts to capture and measure bias/slant in newspaper articles. Newspapers represent an important institution in the political world, affecting phenomena ranging from polarization to voter turnout. Much like the shadows in Plato's cave allegory, news provide elites and citizens with a representation of a reality they are not able to see themselves [16]. The media have the power to affect voting behaviour [4, 10], as well as polarise the electorate [12] or motivate them to turn out to vote [6].

Given the importance of the news media for the study of politics, it is surprising that few papers deploy state-of-the-art deep learning technologies to classify ideological bias in news articles. Gentzkow and Shapiro estimate slant in US newspapers by identifying bi- and trigrams' indicative of a congressional speakers' party, and apply the resulting dictionary to newspapers to scale them [6]. More recently, Widmer et al. have assessed polarisation in the US media environment using a supervised model. They train a classifier on bigrams, predicting whether content was produced by a left-leaning network (CNN) or a right-leaning network (Fox news) [24]. We believe both approaches are likely inferior to more complex deep learning models, as such novel approach would not incorporate idiosyncratic phrases used

[1] GitHub repository: https://github.com/nicolaiberk/nlpdl_project

[2] 'Political bias', 'ideological slant' and variations of the two are used interchangeably in this report.

by the specific networks identified to train the data. Using party labels to train classifiers is more straightforward, as it places newspapers within the existing context. If existing approaches to classify newspaper slant can be improved upon, or even just complemented, we could provide an additional tool for researchers to study drivers and effects of media bias. A working state-of-the-art model might even renew interest in the subject matter and encourage researchers to find new and exciting applications for it.

Outside of academia, a confident and robust classifier of newspaper bias might help readers identify overly partisan articles. This would perhaps encourage them to approach certain news sources with more scepticism and hold news outlets to higher editorial standards. At the very least, it could encourage readers to think about their media consumption habits. In the long run, the highlighting of biases in articles might counteract the worrying polarization of entire electorates.

While the overall intention of the paper holds, our experiments do indicate that the training on party labels cannot be easily translated to the domain of newspaper articles. While our models show impressive performance in validation, out-of-sample test sets, and when censoring crucial information like party labels, they do not translate into stable and accurate classifiers of newspaper bias. While the broad expectations of newspapers' left-right placement are always confirmed, the predictions concerning an article's similarity to a given political party are highly sensitive to minor changes in the inputs.

2. Related Work

Measuring ideology is the prime use-case of automated text analysis in political science. While some work has used supervised learning to understand political polarisation [1] as well as media slant [24], this existing work estimates the polarisation of political actors or newspapers across time, or simply the similarity of one news source to others. Existing models have for the most part relied on bag-of-words approaches, including the scaling of texts using supervised [11] or unsupervised models [21]. Unsurprisingly, the performance of these models has so far proven unreliable, lagging behind the accuracy of human coders or expert judgement [3, 7, 8]. Recent applications have assessed the ability of deep learning models to replicate human coding of party manifestos [2] or estimate ideology directly using supervised classification [20]. These applications predominantly employ RNNs or CNNs, which we feel could be improved upon with state-of-the-art transformer neural networks.

Much less work has been devoted to the measurement of the ideological leaning of newspapers. While some work has assessed the impact of endorsements of specific candidates [4, 10], there have been few attempts to scale newspapers on ideology. Gentzkow and Shapiro estimate slant in US newspapers by identifying bi- and trigrams' indicative of a congressional speakers' party, and apply the resulting dictionary to newspapers to scale them [6]. More recently, Widmer et al. have assessed polarisation in the US media environment using a supervised model. They train a classifier on bigrams, predicting whether content was produced by a left-leaning network (CNN) or a right-leaning network (Fox news) [24].

We believe both approaches are likely inferior to more complex deep learning models. Our approach that fine-tunes Distilbert embeddings on political text (in the form of party press releases) should be more accurate and better suited to detect ideological leanings in other texts, such as news articles [19]. It places newspapers within an existing political context, rather than distinguishing them based on stylistic choices or simple (but ultimately meaning-less) word choice. However, any algorithm trained on one dataset and applied to another must be subject to careful validation.

A recent (yet unpublished) contribution proposes "cross-domain topic classification" to classify topics in parliamentary debates using a classifier trained on party manifestos ([14]). While this is very close to our approach, we improve upon this contribution in three distinct ways. First, while Osnabrügge *et al* employ relatively simple regression models using a bag-of-words approach, we propose the use of advanced deep learning techniques to uncover deeper meaning. Second, scaling similarity to political parties is a generalisable approach, as they are the major organising units of political conflict in contemporary Western societies. Additionally, party communication does not necessitate further labelling.

3. Proposed Method

We propose the use of cross-domain learning to identify newspaper slant: a pre-trained transformers model is fine-tuned to identify the authoring party of press releases. Then, this model is applied to a range of newspaper articles, indicating which parties' communication an article most resembles. This approach is useful, as political parties represent the major organisers of and reference points for political ideology. More importantly, party-issued political texts constitute data that already carry ideological labels and need no further labour for annotation. It therefor represents an efficient and broadly available source of data for cross-domain learning to identify ideological bias in newspaper articles. 'Bias' is here defined as the similarity of a set of text to a given party's communication. This is a reasonable definition, as political parties constitute the major points of reference and organizers of ideology in contemporary democracies.

In detail, we employ pre-trained German-language Dis-

tilBert transformer models .[3] and compare how three different training processes affect model performance. We expect optimal performance from a transformer neural network that was fine-tuned in two steps. First we use the pre-trained model to classify party press releases by issuing party. Secondly, the model is applied in its actual domain to estimate ideological bias in news articles. We subsequently validate the model using domain-specific survey data. These unique data have the advantage of carrying party labels, allowing a direct transfer of the categories from the initial fine-tuning process (first step) to the outcome of interest.

1. A baseline model using DisitlBert with a linear classification layer.

2. Several models with differently adjusted hyperparameters.

3. A model with an additional LSTM layer prior to classification.

This approach allows to assess by how much model performance is improved upon, when including information from both party press releases. If this model performs well, this constitutes strong evidence that party communication can be used to train classifiers which are applied in a diverse set of domains with scarce or no training data. If our model performs well, this would corroborate the idea that information from available sets of texts from political actors can be used to improve ideology measurements for other sets of text.

4. Experiments

Data: In this project, we planned to use three distinct data sources:

- A dataset of over 40,000 German party press releases issued between 2010 and 2019, collected by the SCRIPTS project[4].

- A collection of over 2 million German newspaper articles from six major newspapers, published between 2013 and 2019, collected by one of the authors in a previous project ([9]).

- A survey of newspaper readers asking them about the partisan bias of their newspaper, to generate a first indication of the expected results ([18]).

Software : All software for this project was self-written. However, we relied on the programming languages R ([17]) and Python ([22]), as well as corresponding packages, most importantly Huggingface's transfomers ([25]) and pytorch ([15]) for the training of our models. Any coding necessary for this project was done by either of the authors and relied on Google Colab's free web version. The visualisation of the different training runs was done through the wandb.ai platform.

Evaluation method:

Study 1: We use pre-trained DistilBert embeddings that we fine-tune on party press releases to see how well such a transformers model can classify political texts. We then experiment with different hyperparameter specifications to find the most accurate setup. Finally we see whether the addition of an extra layer (an LSTM in our case) might change the performance of our model.

Study 2: Given that the unexpectedly high performance , we 'blindfold' the model, removing party labels from the input data. This should help us to understand whether the classifier is able to distinguish press releases based on partisan rhetoric alone. Additionally, we predict a dataset of press releases following temporally on the training set. This out-of sample prediction should detect any over-fitting in our data; weak performance in the test set would be a strong indication of it.

Study 3: We use the models trained in study 1 to estimate the ideological bias in 4,000 newspaper articles from German newspapers FAZ, Spiegel Online, TAZ, and Welt. We compare the different models' output with survey data on the partisan slant by newspaper readers and assess the stability of the results.

Experimental details: In total, we trained 6 different models our input data. We use the 'distilbert-base-german-cased' pre-trained model from the Huggingface transformer library[5] as the underlying basis of our model. This smaller version of BERT has been trained to efficiently solve different classification tasks on 12 GB of German language data, including Wikipedia, legal data, and news. We fine-tune the German DistilBert model on our press release data. The classification layer in our transformers model consists of two linear transformations with a ReLU activation in between. For our baseline model, we decided to retain the default hyperparamater specification; three training epochs, a training batch size of 16, a weight decay of 0.01 for regularisation and a learning rate for the Adam optimizer of $5e-5$ [6].

[3] Link to the model page: https://huggingface.co/transformers/model_doc/distilbert.html

[4] https://www.scripts-berlin.eu/index.html. Special thanks go to Lukas Stötzer for the effortless (for us) provision of the data.

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

[6] The model's default settings can be found here: https://huggingface.co/distilbert-base-german-cased/blob/main/config.json.

In addition to the previous model, we also ran a number of experiments to determine the optimal hyperparameter settings. In our choice of experiments we largely followed the guidelines suggested by the authors of the original BERT paper [5]. We ran through the following setups:

- Training epochs: $2, 3, 4$
- Training batch size: $16, 32$
- Learning rate for the Adam optimizer: $5e-5, 3e-5, 2e-5$
- We apply a weight decay as a form of regularisation: 0.01
- Maximum sequence length (longer articles were truncated): 512

The results of the different experiments can be found in figures 1a through 1b. To run these experiments we split our data into 3 randomly sampled parts; a training set (80%), a validation set (10%) and a test set (10%). Following convention, we measured the performance during our experiments only on the validation set and held out the test set for an eventual performance evaluation of our chosen model. Regarding the different experiments, it became clear that the main difference relates to the learning rate.

When setting the learning rate lower than $5e-5$, the model performs worse across all the metrics. We believe that this a sign that it takes too long for the stochastic gradient descent algorithm (Adam) to converge to the optimum. The change in epochs also did not significantly impact the overall performance of the models. Given that the 3 epoch model performed only slightly worse than the model trained for 4 epochs, yet was computationally less demanding, we decided to retain the 3 epochs as our optimal model. When increasing the batch size from 16 to 32, the model improved slightly in accuracy. In the end we decided to stick with the first model specification (3 epochs, learning rate of $5e-5$ and batch size 16), which offers the best trade-off between accuracy and computational efficacy. However, we were not content to simply tune the different hyperparameters. We also wanted to understand how a more complex layers would affect the neural network, so we added an LSTM layer before the dense classification layer in our baseline model. Unfortunately, we did not realize at the time that this model conception is essentially overkill, given that transformers model capture many of the same contextual variables as an LSTM would. For this reason, we will not discuss this architecture in more detail.

Results, study 1: Having chosen our optimal model, we then computed the different performance metrics on the as of yet unseen test set. Its performance is shown in table 1. As the reader can see, the model performs at a very high, near-perfect level for all categories. The impressive performance of this model on the press releases was rather surprising to the authors. Indeed, this performance seemed too good to be true and raised a number of concerns, most notably regarding issues of overfitting. Therefore, we felt compelled to take took further steps to check the robustness of the model performance by changing the input data in two distinct ways.

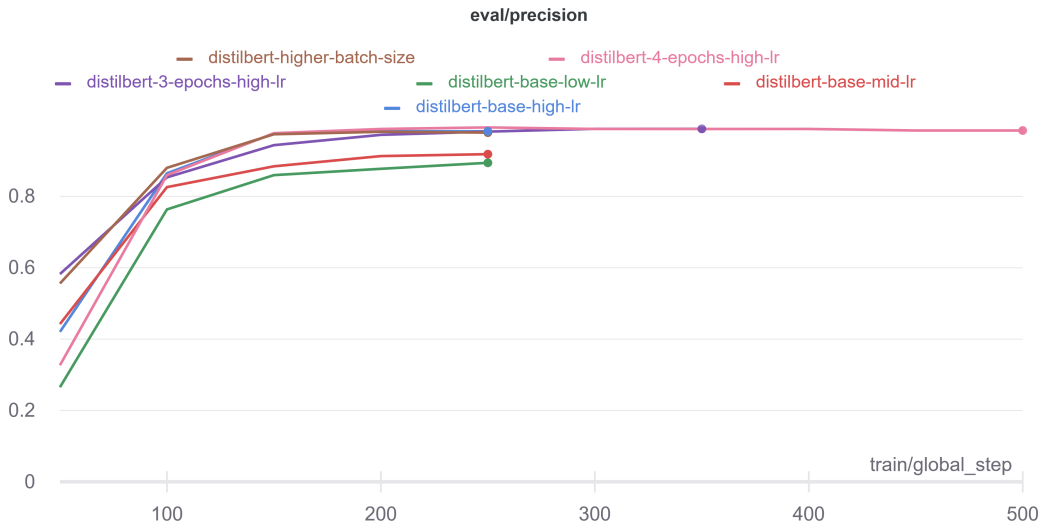
	label	class	f1	precision	recall	n
3	3	SPD	0.997927	0.997238	0.998617	1446
4	4	Linke	0.997237	0.997543	0.996931	1629
5	5	FDP	0.996416	0.996813	0.996019	1256
0	0	Greens	0.995416	0.996940	0.993898	1311
1	1	Union	0.990737	0.984224	0.997336	1126
2	2	AfD	0.989836	0.998423	0.981395	645

Table 1: Performance of default DistilBert model for the classification of party press releases.

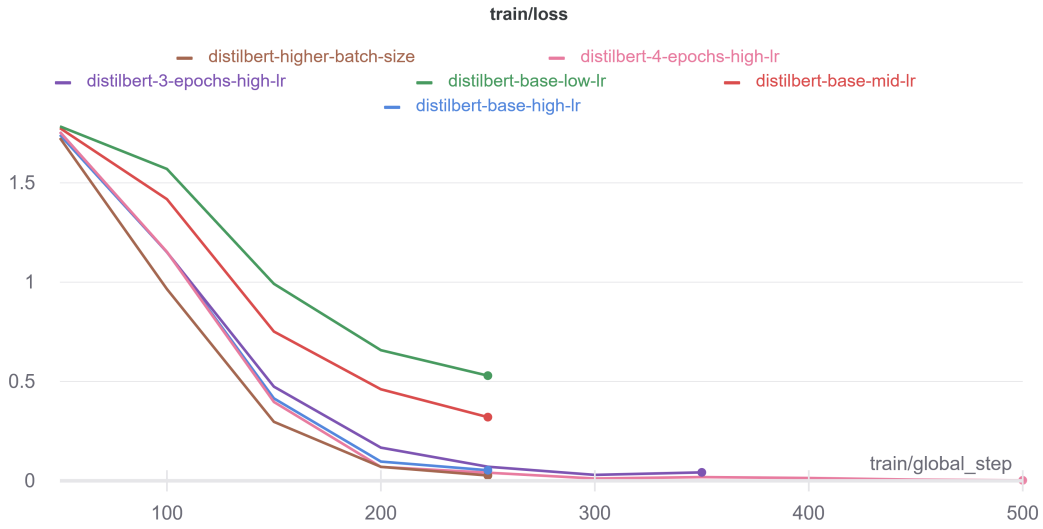
First, we "blindfolded" our model by excluding party labels and any words referring to the parties (such as party-related colours) from the analysis (the full list of terms excluded can be found in Appendix A). If the model performs drops off significantly, this would suggest that transformer picks up and incorporates party-specific clues or speech patterns that could artificially inflate its accuracy. Should the model continue to perform at such a high-level however, we would have to start considering that the model is indeed this accurate.

Unfortunately, even if our model remains accurate in the face of our "blindfolding" method, it will not dispel all the scepticism towards its performance. Therefore, we conducted second robustness check. This consists of a train-test split based on the date of the press release. The idea behind this is to simulate the model's performance on temporally unseen data. By training and evaluating the model's accuracy on press releases prior to 1st of January 2018, and then testing it on press releases that were released after January 2018, we can emulate real-world applications of the model. Hopefully, this would detect if the model were to overfit in the training process. A serious drop in performance on the out-of sample set would indicate that our model learned time specific speech-patterns and can only apply them to test data that is contemporary to its training data. Evidently this would dilute the usefulness of the model severely.

Results, study 2: Table 2 shows the performance of the blindfolded classifier predicting a validation set from the same sampling period as the training data. As the reader



(a) Precision



(b) Training-Loss

Figure 1: Performance across the different experimental models on a series of key metrics.

can see, the performance is virtually identical. This is very, very surprising, given that the most obvious cues - references to other parties. Have been removed. This indicates one of three things. First, the performance might accurately describe the classifiers ability to distinguish partisan differences in language. We will discuss this option further in a moment. Second, we might have missed some obvious cues in the blindfolding process. While our dictionary of party references is rather extensive (see Appendix A), it might still omit important cues that we did not think of. One way to address this issue in the future might be to use word embeddings to expand our dictionary for blindfolding. A third option is that BERT is too powerful, learning some spurious statistical cues to place party press releases with parties ([13]).

As stated previously, this performance, while impressive, will likely not suffice to convince us that no overfitting is taking place. We therefore move on to robustness check 2, which uses the model to predict labels of an out-of sample dataset. If the performance is strongly affected, we are overfitting on the training set. Table 3 shows the performance for this temporally delayed test sample. As we can see, the performance is still very high for many parties, nearly perfectly predicting press releases issued by the Greens, Linke, and Union. AfD and FDP show somewhat decreased but still great performance, although the recall for FDP press releases is not great at .72. However, the precision is severely decreased for the SPD at about .52.

	label	class	f1	precision	recall	n
0	0	SPD	0.998695	0.997392	1.000000	765
2	2	Linke	0.997509	0.995030	1.000000	1001
1	1	Greens	0.995935	0.997286	0.994587	739
4	4	FDP	0.995342	0.996890	0.993798	645
3	3	Union	0.992669	0.995588	0.989766	684
5	5	AfD	0.984874	0.983221	0.986532	297

Table 2: Performance of default BERT model for the classification of party press releases based on the **censored in-sample** data.

	label	class	f1	precision	recall	n
1	1	Greens	0.992078	0.991474	0.992683	820
2	2	Linke	0.960754	0.927273	0.996743	1228
3	3	Union	0.958419	0.923784	0.995751	706
5	5	AfD	0.880880	0.977959	0.801335	2547
4	4	FDP	0.824211	0.966667	0.718349	1090
0	0	SPD	0.686520	0.524342	0.993949	661

Table 3: Performance of default BERT model for the classification of party press releases based on the **censored out-of-sample** data.

Inspection of the cross tabulation for predicted and actual labels indicates that the classifier confuses AfD and FDP press releases for SPD press releases (c.f. Appendix figure 6).

It is somewhat hard to make clear statements based on this evidence, but given that AfD and SPD occupy opposite ends of the ideological spectrum and that one is an opposition, the other a governing party, a well-performing classifier should not confuse the two. It is thus more likely that the classifier picks up spurious relationships in the data that we do not understand. However, before concluding, we will assess whether the newspaper estimates produced by the DistilBert model are in line with our expectations and stable across different specifications.

Results, study 3: We classified a training set of 4,000 newspaper articles using the models described above. We expected articles from conservative newspapers FAZ and Welt to be more similar to Union and FDP and possibly the AfD, especially compared to the progressive TAZ. Spiegel Online (SPON) is expected to be equally similar to right- and leftwing newspapers.

For the three daily newspapers, we can validate our expectation using survey data from the 2017 German Longitudinal Election Study's Rolling Cross-Section (GLES-RCS; [18]). The study asked respondents whether they read a daily newspaper, and if so, whether they felt the newspaper's reporting favored certain parties over others. While these estimates are likely biased by the respondents own political stances (and as such likely underestimate the ideological extremity of their paper), they do provide a first clue. Table 4 shows the aggregated results of this question. We can see that - in line with our initial expectation, FAZ and Welt are placed close to the centre-right Union and FDP (but also SPD), while the more left-wing TAZ is placed close to the three centre-left and left parties (Linke, Grüne, SPD; but note that this estimate is based on only 89 respondents). None of the newspapers is considered to contain coverage particularly favorable for the AfD.

Paper	Linke	Grüne	SPD	Union	FDP	AfD	N
FAZ	0.01	0.02	0.11	0.33	0.11	0	390
TAZ	0.23	0.24	0.27	0.16	0.04	0	89
Welt	0.01	0.02	0.18	0.35	0.09	0.03	209

Table 4: Validation reference: readers' assessment of newspapers' bias

We estimated the similarity of reporting by German newspapers FAZ, Spiegel, TAZ and Welt using the simple BERT model, trained on the full data. The results can be seen in table 5, which shows the average probability for an article assigned by the classifier. As expected, FAZ and Welt are very similar to Union (81%) and FDP (58%/55%), but also rather similar to the Greens (36%/35%). While they show the highest similarity to the AfD (8%/7%), similarity to the radical-right party is generally on a very low level among all newspapers. Spiegel Online (SPON) shows lower similarity to the FDP and closer to Union and Greens, but is generally rather similar to the right-wing newspapers. As expected, the TAZ shows a comparatively different profile, being very similar to the Greens (average likelihood 70%), and less similar to the Union parties (41%), the FDP (32%), and the AfD (5%). Surprisingly, it also shows the lowest similarity to the Linke (13%). Maybe most surprising is the general low similarity to SPD press releases (4%/5%). It seems the party has a rather distinctive style in its press releases.

While the general placement of newspapers is rather similar to the readers' placement (TAZ more left-wing compared to Welt and FAZ), two things are not in line with expectations: the low representation of the SPD in the news articles, and the lowest score for the Linke in the TAZ articles - after all, this is the most left-wing publication.

We restricted the time-frame of the training data to reflect the time-frame of the newspaper articles (2013 - 2018). This

Paper	Linke	Grüne	SPD	Union	FDP	AfD
FAZ	0.15	0.36	0.05	0.81	0.58	0.08
Spiegel	0.11	0.39	0.05	0.89	0.34	0.07
TAZ	0.13	0.69	0.04	0.41	0.32	0.05
Welt	0.15	0.35	0.05	0.81	0.55	0.07

Table 5: Mean similarity estimate to each party by newspaper, based on training data including party labels.

should ensure that performance for all parties is similar and not driven by the higher prevalence of similar topics in the training data press releases. This minor change severely affected our results, as shown in table 6. Resemblance of all newspapers towards Greens and Union have strongly decreased, while resemblance to Linke, SPD, and especially FDP and AfD has strongly increased. All newspapers (including the left-wing TAZ) are now estimated to strongly resemble the FDP instead of the Union. While the higher estimates for SPD are welcomed and in line with readers' expectations, the strong resemblance with the FDP is puzzling, as are the high scores for the AfD and low scores for the Union.

Table 6: Placement of newspapers with date-restricted training data.

Paper	Linke	Grüne	SPD	Union	FDP	AfD
FAZ	0.20	0.20	0.26	0.19	0.85	0.40
Spiegel	0.13	0.17	0.23	0.07	0.94	0.16
TAZ	0.22	0.11	0.67	0.20	0.49	0.29
Welt	0.20	0.15	0.26	0.24	0.84	0.37

We think that - especially in case of the AfD - coverage about the party might be mis-classified as coverage similar to the parties' language. Hence, as in Study 1, we use the 'blindfolded' classifier to estimate the partisan slant of news coverage. Although it could be argued that party labels are a very important source of information as even negative coverage about a party can increase voters' awareness about the party and subsequently its electoral performance. Nevertheless, coverage about a party does not mean that that coverage is biased towards that party. More importantly, party labels that are highly indicative of press release authorship convey little information about ideological bias in news reporting.

Table 7 shows the estimates of this 'blindfolded' classifier. They are somewhat similar to the full but also time-restricted data presented before, but with few marked differences. The Linke is now strongly represented in the TAZ, especially compared to the other newspapers - an estimate that is in line with expectations. However, the Grüne are now least represented in all newspapers. Even the left-wing TAZ bears stronger similarities to the far-right AfD than the Grüne. The right-wing newspapers are now showing the highest similarity to the SPD, which is non-intuitive and in direct contradicition to our survey data. All newspapers show very weak resemblance to Union press releases, again contradicting the survey data. The high similarity to the FDP has even more increased, with the average news article in FAZ, Spiegel, and Welt being estimated to be over 90% similar to FDP press-releases. Even the left-wing TAZ shows the second highest resemblance to the economically right-wing party. This is maybe the strongest inconsistency with the survey data. Lastly, the similarity to the AfD was somewhat decreased. The right-wing newspapers show stronger resemblance with the far-right party than the others, but all estimates are higher than what is expected from the survey data.

Table 7: Placement of newspapers using blindfolded classifier.

Paper	Linke	Grüne	SPD	Union	FDP	AfD
FAZ	0.17	0.08	0.35	0.10	0.95	0.20
Spiegel	0.15	0.06	0.18	0.22	0.94	0.12
TAZ	0.62	0.08	0.23	0.07	0.60	0.17
Welt	0.16	0.11	0.31	0.10	0.93	0.20

5. Analysis

Despite the different experiments and the qualitative adjustments that we made to our model and input data, it looks like our chosen model performance continues to be extremely high and that across the different categories. Since none of our robustness checks seems to have had much of an impact, beyond the estimated drop-off in accuracy due to out-of-sample projection, we are tempted to conclude that our model truly is that well performing. Had we observed significantly different results for the blindfolded model or the temporally input data, this would have suggested problems of overfitting. While we cannot rule them out entirely, the overall picture does seem to suggest that transformers perform extremely well on party press releases. Looking at the training loss reduction in 1b we can also see that the smaller learning rate, had we trained it for more epochs, would have converged to the same accuracy levels as the highest learning rate. This is further suggestive evidence that we are not dealing with overfitting.

When it comes to the transfer to other text data, i.e. our newspaper articles, the picture is less clear. It is unfortunate, but so far no stability emerged in the estimates assessed. Here, small changes in the input data resulted in

vastly different estimates of newspaper bias. It seems that cross-domain applications of deep learning are highly sensitive to the form of the input data and that their application is not as straightforward. We will discuss potential reasons and remedies in the next section.

Perhaps a model that was fine-tuned immediately on articles labelled for bias would do a better job at detecting media slant. This is most likely due to the differences in structure, tone and sentence choice between press releases and newspapers. As it stands, the performance of our model is encouraging however for purely party political texts and offers a wide array of applications in political science.

6. Conclusions

The recommendation at this point is that researchers with cross-domain classification problems think hard about their input data, carefully validating their results, or apply simpler techniques where the estimates are more directly interpretable, such as regression. Nevertheless, cross-domain applications of deep learning are a promising avenue for further research. Future work here should assess how properties of the input data affect estimates in another domain and think carefully about validation to develop best practices for researches seeking to apply such methods.

7. Acknowledgements

We thank the vast community contributing to the development of machine learning tools, specifically huggingface. Additionally, we want to thank Lukas Stötzer for provision of the training data and Hauke Licht for providing us with a first notebook. Most importantly, we want to thank Slava Jankin, Hannah Bechara, and Huy Dang for providing us with the skills to pursue this project.

8. Contributions

All progress was discussed together, although some tasks were taken up more by one author. While the initial idea was developed together, Nico set up the infrastructure and produced a minor first draft of the proposal. This early proposal was mostly worked on by Tom afterwards. Nico then developed the first classifier and corresponding estimates. He wrote most of the midterm report, where the first experiments were presented and the new direction of the project was proposed. Tom proposed the validation through politicians op-eds and collected this data, although we never got to actually use it. In the final weeks of the project, the authors mostly worked on different tasks. Nico worked on a better understanding of the initial BERT model using different input data, assessed survey data for validation, and set up the final report, while Tom moved to more advanced models, adding LSTM and the hyperparameter optimisation of the DistilBert model. They jointly wrote the final report.

References

[1] E. Ash, M. Morelli, and R. Van Weelden. Elections and divisiveness: Theory and evidence. *Journal of Politics*, 79(4):1268–1285, 2017.

[2] A. Bilbao-Jayo and A. Almeida. Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks*, 14(11), 2018.

[3] T. Bräuninger, M. Debus, and J. Müller. Estimating hand-and computer-coded policy positions of political actors across countries and time. *Conference of the Midwest Political Science Association*, 2013.

[4] C. F. Chiang and B. Knight. Media bias and influence: Evidence from newspaper endorsements. *Review of Economic Studies*, 78(3):795–820, 2011.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186, 2019.

[6] M. Gentzkow and J. M. Shapiro. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71, 2010.

[7] F. Hjorth, R. Klemmensen, S. B. Hobolt, M. E. Hansen, and P. Kurrild-Klitgaard. Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research and Politics*, 2(2), 2015.

[8] J. Koljonen, V. Isotalo, P. Ahonen, and M. Mattila. Comparing computational and non-computational methods in party position estimation: Finland, 2003–2019. *Party Politics*, (May), 2020.

[9] W. Krause and N. Berk. Right-Wing Terrorist Attacks, the Media's Reactions, and Radical Right Party Support. 2021.

[10] J. M. D. Ladd and G. S. Lenz. Exploiting a rare communication shift to document the persuasive power of the news media. *American Journal of Political Science*, 53(2):394–410, 2009.

[11] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003.

[12] Y. Lelkes, G. Sood, and S. Iyengar. The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect. *American Journal of Political Science*, 61(1):5–20, 2017.

[13] T. Niven and H. Kao. Probing neural network comprehension of natural language arguments. *CoRR*, abs/1907.07355, 2019.

[14] M. Osnabrügge, E. Ash, and M. Morelli. Cross-Domain Topic Classification for Political Texts. 2020.

[15] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[16] Plato. *Republic*.

[17] R Development Core Team 3.0.1. A Language and Environment for Statistical Computing, 2013.

[18] S. Roßteutscher, H. Schoen, R. Schmitt-Beck, B. Weßels, C. Wolf, and A. Staudt. Rolling Cross-Section Wahlkampfstudie mit Nachwahl-Panelwelle (GLES 2017), 2019.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, Feb. 2020. arXiv: 1910.01108.

[20] A. Simoes and M. Del Mar Castaños. Fine-Tuned BERT for the Detection of Political Ideology Stanford CS224N Custom Project. 6(2017):1–5, 2020.

[21] J. B. Slapin and S. O. Proksch. A scaling model for estimating time-series party positions from Texts. *American Journal of Political Science*, 52(3):705–722, 2008.

[22] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[23] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv*, 2020.

[24] P. Widmer, E. Ash, and S. Galletta. Media Slant is Contagious. *SSRN Electronic Journal*, pages 1–42, 2020.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

9. Appendix

Appendix A Regular expressions excluded for 'blindfolding', separated by comma:

CDU, Christlich Demokratische Union, Christlich-[Dd]emokratische Union, CSU, Christlich-[Ss]oziale Union, Christlich Soziale Union, [Ss]chwarzes, [Ss]chwarzer, [Ss]chwarzem, [Ss]chwarzen, [Ss]chwarze, [Ss]chwarz, Union, SPD, [Ss]ozialdemokratische Partei Deutschlands, [Ss]ozialdemokratisch, [Ss]ozialdemokraten, [Ss]ozialdemokratinnen, [Ss]ozialdemokrat, [Rr]oter, [Rr]otes, [Rr]otem, [Rr]oten, [Rr]ote, [Rr]ot, Bündnis90/die Grünen, Bündnis90 / die Grünen, [Gg]rünes, [Gg]rüner, [Gg]rünem', [Gg]rünen, [Gg]rüne, [Gg]rün, [Ll]inke, Linkspartei, Freie Demokratische Partei, FDP, [Gg]elber, [Gg]elbes, [Gg]elbem, [Gg]elben, [Gg]elbe, [Gg]elb, A[fF]D, Alternative für Deutschland

Appendix B Performance across the different experimental models on a series of key metrics:

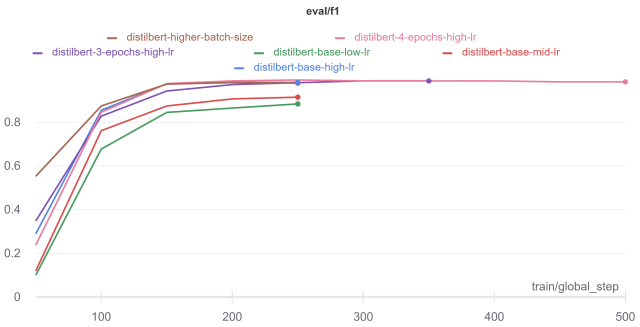


Figure 2: F1-metric

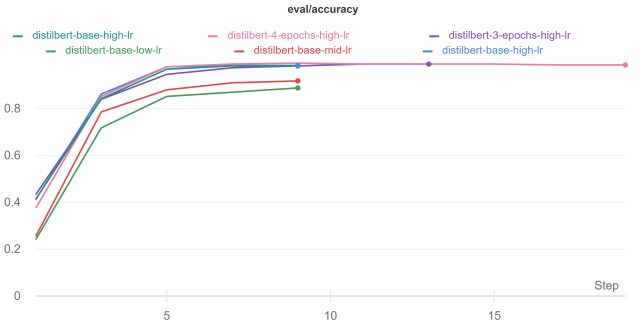


Figure 3: Accuracy

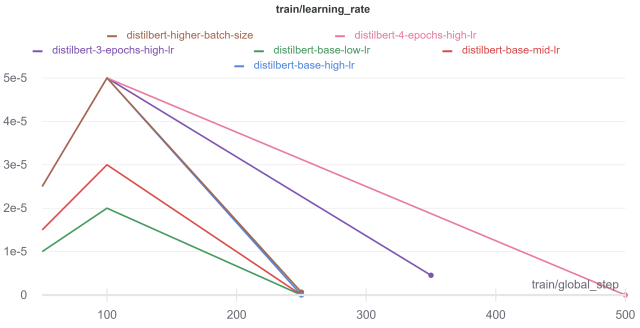


Figure 4: Learning Rate

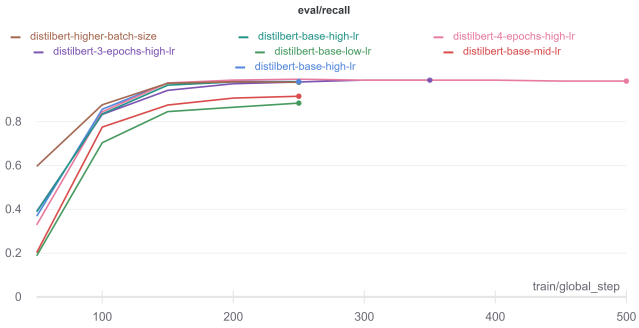


Figure 5: Recall

Appendix C Confusion Matrix Main Model:

Predicted Label True Label	AfD	FDP	Grüne	Link	SPD	Union
AfD	2087	1	0	3	657	0
FDP	0	809	0	0	0	0
Grüne	0	0	821	0	0	0
Link	0	0	0	1317	0	0
SPD	0	0	0	0	596	0
Union	0	0	0	0	0	761

Figure 6: Confusion Matrix - Main Model