Ideology Prediction of German Political Texts

Sinclair Schneider[1], **Florian Steuber**[1], **João A. G. Schneider**[1], **Gabi Dreo Rodosek**[1]
[1]Bundeswehr University Munich
Research Institute CODE
Carl-Wery-Straße 22
81739, Munich, Bavaria, Germany
sinclair.schneider@unibw.de, florian.steuber@unibw.de, joao.schneider@unibw.de, gabi.dreo@unibw.de

Abstract

Elections represent a crucial milestone in a nation's ongoing development. To better understand the political rhetoric from various movements, ranging from left to right, we propose a transformer-based model capable of projecting the political orientation of a text on a continuous left-to-right spectrum, represented by a normalized scalar, $d \in [-1, 1]$. This approach enables analysts to focus on specific segments of the political landscape, such as conservatives, while excluding liberal and far-right movements. Such a task can only be achieved with multiclass classifiers, provided that the desired orientation is incorporated within one of their predefined classes. To determine the most suitable foundation model among 13 candidate transformers for this task, we constructed four distinct corpora. One corpus comprised annotated plenary notes from the German Bundestag, while another was based on an official online decision-making tool, Wahl-O-Mat. The third corpus consisted of articles from 33 newspapers, each identified by its political orientation, and the fourth included 535,200 tweets from 597 members of the 20th and 21st German Bundestag. To mitigate overfitting, we used two distinct corpora for training and two for testing, respectively. For in-domain performance, DeBERTa-large achieved the highest F1 score ($F_1 = 0.844$) as well as for the X (Twitter) out-of-domain test ($ACC = 0.864$). Regarding the newspaper out-of-domain test, Gemma2-2B excelled ($MAE = 0.172$). This study demonstrates that transformer models can recognize political framing in German news at the level of public opinion polls. Our findings suggest that both the model architecture and the availability of domain-specific training data can be as influential as model size for estimating political bias. We discuss methodological limitations and outline directions for improving the robustness of bias measurement.

Code — https://github.com/SinclairSchneider/german_ideology_prediction

Bundestag/Wahl-O-Mat Datasets — https://doi.org/10.57967/hf/4924

German Media Datasets — https://huggingface.co/collections/SinclairSchneider/german-media-67dcb6c0bf4c007db3999153

Introduction

In February 2023, investigative journalists from the network "Forbidden Stories" uncovered a disinformation-as-a-service provider, working with social media bot accounts, known as "Team Jorge" (Andrzejewski 2023). This entity claims to have manipulated 33 elections, 27 of which were deemed successful. To demonstrate their capabilities, Team Jorge spread false rumors about a deceased emu (#RIP_Emmanuel), which ultimately led to real issues at the animal's farm. Although this is a particularly negative example, it highlights the considerable influence of social media on politics.

We believe that the robust tools of social media analysis can play a valuable role in helping political parties better understand the needs and preferences of their constituents, as well as in forecasting the trajectory of political discourse. To achieve this goal, the political ideology spectrum can be quantified on a continuous scale from -1 (left) to 1 (right). Assuming such a mapping is found, individuals' political ideology can be approximated from tweets on X. A range of $-1 \leq \theta \leq -0.9$ would yield left-wing topics such as the establishment of a single public healthcare system, the withdrawal of U.S. troops from Germany, a focus on social justice and climate protection, and an end to weapons exports. More centrist positions may be found in a range of $-0.1 \leq \theta \leq 0.1$, including principles against extremism, efforts to combat hate speech and misinformation, democratic values, military modernization, and digital strategies. Consequently, a threshold of $0.9 \leq \theta \leq 1$ might reveal right-wing topics such as the end of weapon supplies to Ukraine, claims of economic destruction linked to voting for the Green Party, viewing climate change as a business model, and the perception of immigration and Islam as threats to Western countries.

To achieve this, one could implement a topic modeling algorithm such as BERTopic (Grootendorst 2022). However, these approaches lack an essential component: the ability to dynamically focus on a specific political direction, which can only be addressed partially by classifiers with predefined categories. Therefore, this paper introduces a new algorithm that maps political texts onto a continuous scale ranging from -1 to 1, with a liberal orientation at 0.

This paper addresses three significant challenges: first, it aims to map text onto a continuous left-to-right spectrum

rather than simply categorizing it into discrete classes. Second, it seeks to adapt the generated algorithm to account for local political biases through a semi-supervised labeling approach. Third, it focuses on ensuring the algorithm's effectiveness by testing on distinct, out-of-domain datasets.

Approach The foundation for training a classifier that maps texts to a continuous left-to-right spectrum is the association of two-dimensional normalized vectors with political parties. An entirely left-wing party would be represented by a vector pointing to the left (-1, 0), while a right-wing party would have a vector directed to the right (1, 0). A centrist party would be indicated by an upward vector towards the center (0, 1). Intermediate positions are encoded by vectors of unit length at corresponding angles.

The output of a trained multilabel classifier, indicating the extent to which a party agrees with a given statement, is then multiplied by the corresponding vectors. At the end, all vectors are added, and the angle of the newly formed vector represents the classification result. To demonstrate that this approach is effective, it is finally tested on both crawled German newspapers and politicians' tweets, for which the political leanings are known. This outlines both the classifier's accuracy and its out-of-domain capabilities. In order to do so, we trained and tested 13 transformer classifiers.

Contribution The main contributions of this paper are the extension of previous approaches that used categorical variables with a continuous left-right spectrum between -1 and 1, as well as demonstrating the out-of-sample capabilities of our classifier. When tested against the 33 newspapers, our best classifier yielded a mean error (ME) of 0.17 on a scale between -1 and 1, which is an error of 8.58% on a survey-based benchmark dataset. Regarding the origin-prediction tweets, we found that accuracy increases to 0.864 when 100+ words are available. By using plenary speeches from the German Bundestag as one of the training sets, we ensured that our classifier is perfectly aligned with the German left-right spectrum without introducing the author's bias. With a total of four self-collected datasets, we also made sure that the out-of-domain accuracy is provided. By adapting the task of political stance prediction to a German context, we contribute to a more diverse array of training data and models, as this not only requires linguistic adaptation but also considers the unique political environment.

Related Work

Political ideology detection is typically done by building classes such as left, center, or right, using a manual annotation approach (Baly et al. 2020).

Different research projects approach the issue of such a limited political scale in various ways. Some focus solely on detecting (extreme) left-wing or right-wing opinions (Kiesel et al. 2019; Jakob et al. 2024), while others offer a broader spectrum (AllSides 2025). These broader approaches include classifications for "lean left" and "lean right", situated between the center and the two extremes. Others offer an even more fine-grained classification of seven or more classes (Preoţiuc-Pietro et al. 2017; Fagni and Cresci 2022), for instance, very conservative, conservative, moderately conservative.

Most foundational research is conducted in English, which often leads to an association with the United States. However, simply translating existing English-language datasets is insufficient for their application to German politics, given the diverse political views across countries. For this reason, researchers have begun to collect and label specific datasets in German, utilizing information from German newspapers (Aksenov et al. 2021).

The global nature of social media platforms, which span across borders and cultures, makes it difficult to develop generalizable models trained on tweets. For instance, methods that achieve over 90% accuracy on a carefully selected dataset can drop to approximately 65% when applied to different users within the same network (Cohen and Ruths 2013). Despite this, social media continues to be a focal point for transformer-based classification methods, particularly with models tailored for social media like BERTweet (Nguyen, Vu, and Tuan Nguyen 2020) and PoliBERTweet (Kawintiranon and Singh 2022).

Expanding beyond a text-only approach to ideology classification and incorporating users' networks opens up new opportunities for classification methods that utilize transformers, as demonstrated in previous research (Jiang, Ren, and Ferrara 2023).

Exploring publications analyzing German Bundestag speeches leads us to the work of Erhard et al. (2025), who investigated the rise of populism using these speeches. They identified four main categories: anti-elitism, people-centrism, left-wing ideology, and right-wing ideology. This framework enhances the traditional two-dimensional political spectrum by incorporating anti-elitism and people-centrism, while still relying on hand-labeled discrete categories.

Baly et al. (2019) adopt a similar approach by introducing trustworthiness as a second dimension on a three-point scale. Their work demonstrates that political orientation can be a useful factor in detecting misinformation, bias, and propaganda.

The issue of models trained on specific domains, such as news sites, performing poorly on other domains, like social media, in ideology classification has been noted by Volf and Simko (2025). They addressed this challenge by mixing datasets from multiple domains for the training process. Another way to improve the classifier's output is to build a dataset comprising the same stories told by news outlets with different political biases, providing a direct comparison of the same story across different political perspectives (Liu et al. 2022).

All approaches discussed so far are limited due to their categorical outputs. Specifically, ordinal scales cannot measure the extent to which left- or right-leaning perspectives are present. As there is no convention regarding the specific categories, model usage is limited to a predefined context. For instance, the concept of a left-wing opinion in the US may differ significantly from that in Germany.

Methodology

The processing pipeline was structured as follows: First, data from several sources was collected and further enriched to obtain generalizable models. Second, a binary political classifier and subsequent multi-label party classifiers were trained, using multiple BERT, Llama, and Gemma LLMs. Third, the multilabel output was converted to a continuous left-right spectrum (-1 to 1). Finally, in-domain and out-of-domain performance was evaluated using separate test sets, each drawn from an independent dataset. Furthermore, pre- and post-vector-optimization results are compared.

Datasets

Two independent sources (Bundestag, Wahlomat) were preprocessed for model training and testing. Despite artificially enriching and splitting the data (80:20 train-test split), models may overfit. This is why two additional datasets (newspapers, tweets) were used for model evaluation. For training and evaluation, the data of all datasets were either pre- or auto-labeled as explained below.

Bundestag Dataset All plenary debates of the German Bundestag are recorded in writing by stenographers and published (Deutscher Bundestag 2025). Besides the text of the speech, the speaker's name and party membership are minuted. This is also true regarding requests (question, party and name of the questioner) and all other potential speech interruptions, such as interjections, hissing, applause, etc. (type and party, resp. parties). All protocols were collected and processed for the period from October 2017 to September 2024. The raw speech data comprises 34,174 speeches.

Labeling The combination of speeches and interruptions constitutes a robust auto-labeling approach. All speeches were filtered for recorded interruptions. Speeches without any interruptions were discarded. For the remaining ones, the sentiment was extracted from the comments. The described extraction process is illustrated in Figure 6. This procedure yielded a dataset of 32,246 annotated statements (i.e., pro or contra opinions of parties). The association between parties based on the extracted sentiment is depicted in Figure 1 (upper triangle).

Data Enrichment In order for a classifier to correctly categorise not only political speeches but also political statements in general, the linguistic variance of the statements was artificially increased. For this purpose, a LLama 3.1 model was asked to summarize each text in five different versions: In the words of a child, of a teenager, of an adult, of an eloquent person, or as a social media post (tweet). The expanded dataset consisted of 449,209 statements. It was made publicly available (Schneider 2025b) after combining it with the Wahlomat dataset, which is described below.

Wahlomat Dataset The German multi-party system makes it difficult for voters to find the party that represents their interests best. Hence, a digital voters' guide called *Wahl-O-Mat* is released ahead of every federal and state election by the Bundeszentrale für politische Bildung (Federal Agency for Civic Education). It consists of several political statements that the user can agree or disagree with (viz. Fig. 5 for an example of the federal election in 2025). For this system to function, the respective party positions (approval, neutral, rejection) were officially surveyed in advance by the Federal Agency.

The used data is available online (Bolte 2025), comprising 1,751 unique statements regarding the elections between 1998 and 2021.

Labeling No annotation was needed as the data already consists of statements and attitudes of all parties. Attitudes were coded as 1 (approval), 0 (neutral), or -1 (rejection), respectively. Based on these values, the association between parties is illustrated in Figure 1 (lower triangle).

Data Enrichment The dataset was also synthetically enriched as described above, yielding 87,210 labelled statements. Table 6 presents an example of how the call for introducing a wealth tax could be expressed from various perspectives. The positions of the various parties regarding the original statement and thus also concerning the generated ones can be found in Table 4.

To ensure that the enriched sentences maintain similarity to the originals, we utilized the Qwen3-Embedding-8B model (Zhang et al. 2025) to map them into a vector space and calculated the cosine similarity against the original sentences. In contrast to parliamentary speeches containing substantial extraneous content (e.g., greetings), the Wahlomat dataset consists exclusively of condensed statements. Hence, only the latter was used for comparisons. The overall similarity of the paraphrased examples is 0.74, while the most similar sentences, paraphrased for a teenage audience, yielded an average cosine similarity of 0.78. To determine whether political bias was introduced during data enrichment, the cosine similarity distribution is assessed. As is common in statistics, the 5th percentile is computed. Since this extreme quantile is still sufficient with 0.54, we can assume that no fundamental bias has been introduced.

The combined training dataset (Bundestag+Wahlomat) consisted of 570,416 samples and is publicly available (Schneider 2025b).

Tweet Dataset To evaluate the performance of classifiers on short social media texts, we curated a dataset consisting of 535,200 tweets from 597 members of the 20th and 21st German Bundestag (Federal Parliament). Each political party is represented by 89,200 tweets, filtered to include only political content.

Labeling The labeling is based on the account owners' affiliation with the respective political party. Each tweet is assigned to a single political party only.

Newspaper Dataset Based on the assumption that the German media landscape sufficiently represents the political spectrum (cf. Maurer et al. 2024), a dataset of 33 newspapers was examined. From each source, at least 10,000 articles were collected, resulting in a representative dataset of approximately 10 million articles. An overview with precise numbers for all media is appended (cf. Table 5). Additionally, we retained metadata, such as news categories, to train

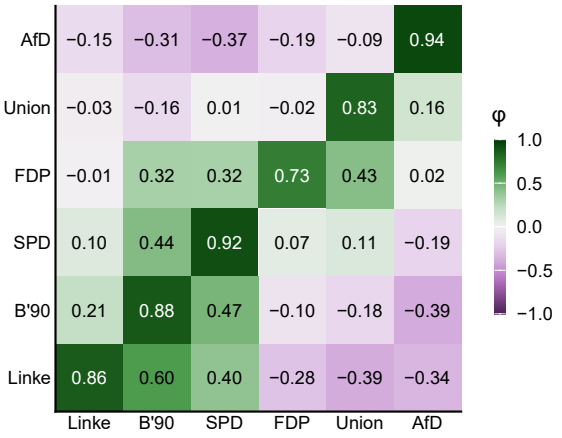

Figure 1: Associations between the parties based on Bundestag sentiments (upper triangle) and Wahlomat statements (lower triangle). Profile similarity (within, viz. diagonal) estimated Pearson's correlation of Phi measures per party between Bundestag and Wahlomat datasets.

a binary politics-non-politics classifier that serves as a filter later. The dataset was based on prior political classifications available for 39 newspapers (see below). Six newspapers were either discontinued or inaccessible due to technical issues.

Labeling The political stance of the articles was unknown, but several estimates exist at the newspaper level. The main one used here is based on $n = 1148$ participants who rated $k = 39$ newspapers on a scale from 1 (*extreme left-wing*) over 4 (*minimal party affiliation*) to 7 (*extreme right-wing*), with fake news and conspiracy theories falling under both extremes, respectively (Medienkompass.org 2025).

To verify the validity, we compared the ratings with the ones provided by two independent sources: Firstly, a comparable bias-rating platform that covers various international outlets (Mediabiasfactcheck.com 2025) and secondly, a scientific report about the German media landscape (Maurer, Kruschinski, and Jost 2024). Regarding both sources, appropriate association measures were computed using all pairwise complete cases to estimate convergent validity. We also report the respective measures for the subset of our sample.

Mediabiasfactcheck.com reports data for $k = 77$ media outlets, but only non-numeric labels in roughly half of the cases. The ratings are based on a scale from -10 (*extrem left*) over 0 (*least biased*) to +10 (*extreme right*). For better comparability, both considered scales were z–transformed. Note that this does not affect the correlation estimates but makes the scores directly comparable, as reported in Table 5 (mean values of zero with standard deviations of one). Both estimates were very highly correlated with $r = .90$ (resp. $r = .91$ regarding the sample). However, this estimate was based on the overlap of $k = 9$ outlets only ($k' = 7$ regarding our sample). To enlarge the intersection, the provided ordinal labels were converted into numerical values (i.e., *left* was assigned to -2, *left-center* to -1, *least biased* to 0, etc. with positive values for the right-hand side). Using Spearman's ρ for ordinal data yielded an even higher correlation of $\rho = .95$ for $k = 19$ pairs ($\rho = .96$ for $k' = 17$ regarding the sample).

Although the correlations are very high, it could be criticized that both ratings come from public platforms. Accordingly, the ratings from a scientific study were examined (2024, Maurer et al.), providing data for $k = 47$ media outlets by only $n = 9$ but extensively trained raters. Here, political ideology was rated using two separate five-point scales. As these showed a strong positive correlation ($r = .63$), both were reduced to a single dimension using principal component analysis (PCA; default settings, varimax rotation). From the resulting one-dimensional values, a subset of $k' = 21$ outlets was present at Mediencompass.org, yielding a very high correlation of $r = .95$ ($r = .94$ for the subset of $k' = 22$ regarding the sample).

Since ratings were shown to be very highly correlated with two independent sources, the validity of Mediencompass.org can be considered sufficient. This is also the case regarding our sample, which had approximately the same correlation coefficients.

Models

Foundation Models To effectively classify German political texts, we needed to select appropriate foundation models for this multilabel classification task. We used smaller encoder-only models with 0.21-2.1 billion parameters, alongside larger decoder-only models with 1.0-9.0 billion parameters.

For the encoder-only models, we chose DeBERTa Large (Dada et al. 2023), GottBERT Large (Scheible et al. 2024), GBERT and GELECTRA Large (Chan, Schweter, and Möller 2020), xlm-roberta Large (Conneau et al. 2020) and EuroBERT (Boizard et al. 2025). In contrast to the original DeBERTa model presented by He et al., Dada et al. trained a model on the same architecture but used a diverse German training corpus. This collection includes online encyclopedias, social media content, legal documents, medical texts, and fiction books, which collectively make the foundation model well-suited for a wide array of German-language applications.

The authors of GottBERT followed a similar approach, except that they employed a RoBERTa BASE architecture (Zhuang et al. 2021) combined with the OSCAR (Ortiz Suárez, Sagot, and Romary 2019) dataset.

GBERT and GELECTRA Large are developed by the same authors and use the BERT (Devlin et al. 2019) and ELECTRA (Clark et al. 2020) architectures, respectively, to build German foundation models from German text. The training data for these models is sourced from the OSCAR and OPUS corpora (Tiedemann 2012), as well as Wikipedia and OpenLegalData (Ostendorff, Blume, and Ostendorff 2020).

The EuroBERT series of models is also an encoder-only architecture trained on 5 trillion tokens across 15 European languages, including German. The family of decoder-only models, including the following, is best known for their generative capabilities but can also be used for classification tasks.

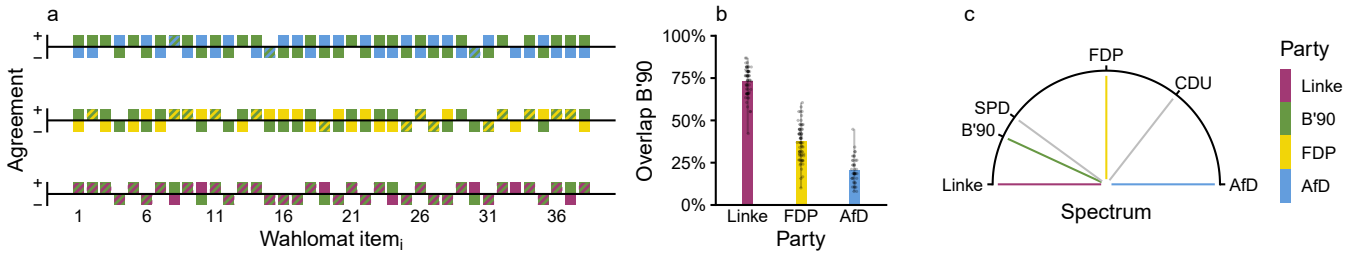

Figure 2: Exemplary comparison of the Green Party (B'90) against the right-wing (AfD), liberals (FDP), and the left party (Linke) in subplot a. Hachures indicate Wahlomat items on which both compared parties agree or disagree regarding the Brandenburg election in 2024. The mean overlap of all election results is displayed in subplot b. Results are mapped onto a left-right spectrum in subplot c regarding (dis-)similarity distance to the other parties.

The Gemma 2 models (Riviere et al. 2024) have primarily been trained using English texts, however, they use a significantly larger tokenizer inherited from the Gemini model (Gemini Team Google 2025), comprising 256,000 entries. This extensive vocabulary, combined with the multilingual nature of web data, enables the model to comprehend languages beyond English. A distinctive characteristic of the smaller Gemma models, specifically those with 2 and 9 billion parameters, is that they are trained via knowledge distillation from a larger teacher model. This methodology gives these models an advantage over others of similar size, enhancing their performance and efficacy in various applications.

In contrast, the authors of the Llama 3.2 (Dubey, Jauhri, and Pandey 2024) models used FastText to categorize the training data into 176 different languages, including German.

In summary, the smaller encoder-only models are mostly trained on German texts, whereas the larger decoder-only models were trained on multilingual data, including a German component.

Political Classifier Prior to the political orientation classifiers, we require an additional classifier to determine whether a text is political. This is crucial for assessing a newspaper's leaning, whether it is more left or right of center. If we were to classify all texts indiscriminately, we would also include non-political content, which could skew our average classification results toward the center. For this purpose, we used the metadata described in the previous newspaper section. By merging all political newspaper categories into a single political section and grouping other categories, such as entertainment, separately, we created a well-balanced dataset comprising 234,978 political and non-political texts (Schneider 2025c). This dataset is then used to train a German DeBERTa model (Dada et al. 2023) that can predict the probability that a text is politically related. The model is subsequently used with a threshold of 0.8, as suggested by the authors, ensuring that only political texts are processed further. The model achieves an F1 score of 0.99 on the test set, although a slight decline in performance on out-of-domain data is anticipated.

Political Party Classifiers To determine the appropriate political alignment of a text, we have trained 13 classifiers, including DeBERTa-large, EuroBERT, GBERT, XLM-RoBERTa, Llama, and Gemma, using a multilabel classification approach. This approach links an input text to one or more of the six major German political parties. After training, the best-performing classifier among the 13 candidates is evaluated using the out-of-domain newspaper data. The training process itself involves feeding lines similar to those in Table 4 into the pretrained foundation models and fine-tuning their weights for four epochs. For a detailed list of models and used parameter specifications, see Table 1. The six parties Die Linke, Bündnis 90 Die Grünen, SPD, FDP, CDU/CSU, and AfD were selected based on their consistent representation in the German parliament over the past few years. For the training, we focused on likes, excluding dislikes.

The training was conducted on various GPU servers, ranging from 4 A6000 Ada to 8 H200 GPUs. All training files are publicly accessible (Schneider 2025a), and the DeBERTA model was executed multiple times to identify the optimal hyperparameter configuration. Given that larger models required several days to train, we were unable to conduct a full training run for each configuration. Instead, we stopped training when the training loss no longer decreased and adjusted the parameters accordingly.

From Multilabel to a Continuous Scale

At this stage, it is necessary to explain why we have introduced a multilabel classifier while simultaneously representing a continuous output for political direction on a scale from -1 (left-wing) to 1 (right-wing). The key missing element is an adaptation model that translates the outputs of the multilabel classifiers (which correspond to six political parties) into the left-right spectrum. This adaptation model is based on the premise that each political party can be positioned on a left-right continuum, with varying degrees of liberalism. An alternative geometric representation is a semicircle. In this representation, we position three fixed points: Die Linke, the most left-wing party, on the far left; the FDP, an economically liberal German party, at the center; and the AfD, a strongly right-oriented party, at the far right end of the semicircle.

Classifier	Size	F_1	train time
DeBERTa-large[a]	435M	0.84	12:03:20
Gemma-2-9b[b]	9B	0.82	3d 05:04:34
EuroBERT-610m[b]	610M	0.79	1d 06:50:30
Gemma-2-2b[c]	2B	0.75	7d 12:39:57
GottBERT_large[b]	357M	0.74	09:20:56
gbert-large[b]	337M	0.73	09:32:50
EuroBERT-210m[b]	210M	0.73	09:06:01
gelectra-large[b]	336M	0.72	10:03:12
EuroBERT-2.1B[b]	2.1B	0.72	2d 16:01:31
xlm-roberta-large[b]	561M	0.71	09:37:06
Llama-3.2-3B[b]	3B	0.71	3d 09:40:08
Llama-3.2-1B[b]	1B	0.69	2d 22:43:17
gemma-3-1b[b]	1B	0.69	2d 14:11:47

Table 1: Overview of the used models, parameter sizes, evaluation metrics, training hours, and hardware used for training, i.e., *a.* 4 A6000 Ada GPUs (4×48GB vRAM); *b.* 8 H100 GPUs (8×80GB); *c.* 8 H200 GPUs (8×141GB)

The remaining task is to determine the positioning of the other three parties. Based on known political positions of the German parties, we know that the CDU is more conservative than the FDP, so it should be placed somewhere between the fixed points represented by the FDP and the AfD. Additionally, we recognize that the Grüne and SPD parties are more left-leaning than the FDP, indicating that they should be positioned between the fixed points of Die Linke and the FDP.

To begin our analysis with the party Die Grünen, we need to determine whether they align more closely with Die Linke or the FDP. To do so, we use the Wahlomat dataset described above. As it contains responses from political parties to the statements, we can compute the overlap across parties.

Consider the following scoring system for measuring agreement: assign a distance of 0.0 to two parties who provide identical answers, a distance of 0.5 to two parties whose responses differ slightly (one agrees or disagrees with a given statement while the other remains neutral), and a distance of 1.0 to two parties who are in complete disagreement. Figure 2a illustrates how the principle operates using the example of a particular election. Whenever there is an overlap in opinions, such as both parties endorsing the same statement, a striped pattern appears. The greater the number of striped boxes, the more similar the two parties are. In Figure 2b, we see that the Grüne party has the most overlaps with Die Linke. Meanwhile, Figure 2c accurately positions the Grüne party between Die Linke and the FDP, reflecting the calculated distances and angles.

The following calculation will be used for the sake of illustration. The Green Party and the Left Party collectively addressed 2,111 questions, providing identical responses to $I = 1,530$ of them. In $P = 284$ cases, one party took a neutral stance while the other either agreed or disagreed. Furthermore, on $O = 297$ questions, the two parties expressed differing opinions.

The Green Party and the Liberal Party (FDP) answered a total of 2,249 questions together. They fully agreed on $I = 828$ questions, partially agreed on $P = 383$ questions, and disagreed on $O = 1,038$ questions.

	Linke	B'90	SPD	FDP	CDU	AfD
$\theta°$	-90.0	-65.2	-53.9	0.0	37.9	90.0
v_x	-1.0	-0.9	-0.8	0.0	0.6	1.0
v_y	0.0	0.4	0.6	1.0	0.8	0.0

Table 2: Party vectors θ_i and unit vectors $\boldsymbol{v}$.

Let $d_{(a,b)} := (0.5 \cdot P + O)/T$ denote the estimated distance of two parties a and b, where $T = I + P + O$. Regarding the example, this yields $d_{B'90,Linke} = 0.208$ and $d_{B'90,FDP} = 0.547$. The relative proximity of a party a to the two reference parties b and c is then mapped onto the interval $[-90°, 0°]$ for left wing resp. $[0°, 90°]$ for right-wing parties by using $\theta_a = \varphi \cdot \big(d_{(a,b)}\big)/\big(d_{(a,c)} + d_{(a,b)}\big)$. Regarding the example of *B'90 - die Grünen* being a potentially left party, $\varphi = -90°$ is used, leading to $\theta_{B'90} \approx -65.2°$. The same reference parties are used for SPD, yielding $\theta_{SPD} \approx -53.9°$. However, using $\varphi = 90°$ the more right-wing CDU party is compared with FDP and AfD, resulting in $\theta_{CDU} \approx 37.9°$. For implementation, the arctan2 function is used to determine the quadrant of the circle. See Table 2 for an overview of all six party vectors.

Based on these angles θ_i, we calculate unit vectors $\boldsymbol{v}_i$ for each party $i \in \{\text{Linke}, \text{B}'90, \text{SPD}, \text{FDP}, \text{CDU}, \text{AfD}\}$ as $\mathbf{v}_i := \big(\sin(\theta_i), \cos(\theta_i)\big)$.

In the final step, the output p_i from the multilabel classifier is multiplied by the corresponding vectors $\mathbf{v}_i$, and all of these vectors are summed together. Formally, $\mathbf{v}_{\text{res}} = \sum_i p_i \, \mathbf{v}_i$. Finally, the angle is then calculated using $\text{atan2}(\mathbf{v}_{\text{res}})$ and divided through $\pi/2$ to transfer the resulting angle to a final classification score.

Overall Architecture

In this section, we combine all the building blocks that have been introduced in the previous sections. This is illustrated by processing the sentence: "Familienpolitik soll Wahlfreiheit ermöglichen: gute Kitas, Ganztagsschulen und flexible Arbeitsmodelle." (Family policy should enable freedom of choice: good daycare centers, all-day schools, and flexible working models). Since we are discussing social benefits, we expect a left-leaning result $(score < 0)$.

Political Classifier First, we check whether the example is political by using the DeBERTa political classifier introduced previously. This results in a score of 0.99, indicating that this statement is political.

Political Party Classifier If a text is classified as political, it is next processed using all 13 trained political party classifiers. For this example, we use gemma2-9b solely, which yields the following probabilities: $P(\text{party} = \text{'Linke'}) = .0307$, $P(\text{'B'90'}) = .2806$, $P(\text{'SPD'}) = .2743$, $P(\text{'FDP'}) = .4508$, $P(\text{'CDU'}) = .0698$, $P(\text{'AfD'}) = .0011$.

From Multilabel to a Continuous Scale We multiply the obtained model probabilities by the given party vectors and

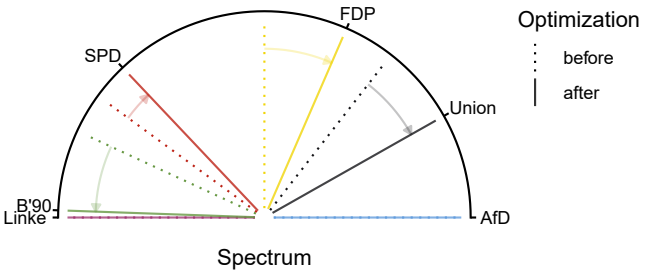

Figure 3: Comparison of the party vectors before and after the optimization for Gemma2-2b

calculate the combined vector $\mathbf{v}_{\text{res}} = (-0.159, 0.277)$.

From the combined result vector, we can accurately calculate the angle using the arctangent or atan2 function, which considers all four quadrants of the circle.

$$\theta_{\text{result}} = \text{atan2}(-0.159,\ 0.277) \approx -0.521\,\text{rad} \approx -29.851^\circ$$

To obtain a score s, scaled between -1 and 1 instead of -90 and 90, we must divide the result by 90 degrees or by $\pi/2$ when using radians: $\text{score} = \theta^\circ_{result}/90^\circ \approx -0.332$.

Our example regarding family politics has resulted in a slightly left-leaning vector, as assumed previously.

Evaluation Using Newspapers

The final step of the evaluation involves using the newspaper ratings from Mediencompass.org to compare its general orientation with the classifier's results. For instance, the newspaper Bild (Springer 2025) is rated 5.2 (0.4 on our scale, $[-1, 1]$) by the project, indicating a slightly right-wing orientation. Since our classifier operates at the per-article level, whereas the ground truth is at the per-newspaper level, we need to aggregate our classifier's results across all political articles and evaluate how closely the computed average aligns with the ground truth. As explained above, a filtering process is essential before assessing the political direction, ensuring that all articles meet a politicalness threshold of 0.8. If the average classification of all political articles in the newspaper Bild were 0.3, the error would be 0.1. The mean error across all 33 newspapers then reflects the quality of the specific trained classification model.

Let a be an article of a newspaper A out of the set of newspapers $\mathcal{A}$.

First, we compute the political leaning for each article a in a newspaper A and compare the result with the expected leaning based on Mediencompass.org L (newspaper level). The mean absolute error MAE of the tested model is computed as the average over the absolute differences between all newspapers tested by the respective model.

Final Optimization

In our final step, we aim to refine the vector model to better align with the newspaper data. Specifically, while we optimize for the evaluation data, we impose a constraint that limits the adjustment of each party vector to a maximum of 0.25 in either direction. This new optimization builds on our initial use of the Wahlomat responses, in which we anchored the liberal party, the FDP, at the top of the semicircle. Although this initial method involved several guiding assumptions, it did not guarantee an optimal outcome. By introducing these constraints, we acknowledge the value of our initial approach and seek to prevent the model from overfitting to the evaluation data.

Moreover, we position Die Linke on the left side and the AfD on the right side, while allowing for adjustments to the liberal FDP. This decision also has a technical basis. When we move the vectors representing the leftmost and rightmost positions upward, these positions cannot be reached through vector combinations without introducing negative contributions, which is not feasible, as the multilabel classifier only produces positive outputs ranging from 0 to 1. Thus, adjusting the leftmost and rightmost vectors would restrict the set of reachable vectors.

$$\min_{\{\Delta v_p\}_{p\in\mathcal{P}}} \text{MAE}(\tau) = \frac{1}{|\mathcal{A}|} \sum_{A\in\mathcal{A}} \left| \widehat{L}_A(\tau; \{v_p + \Delta v_p\}) - L_A \right|$$

$$\text{subject to } \|\Delta v_p\| \le \delta_p \quad \forall p \in \mathcal{P}$$

$$\text{with } \delta_p = \begin{cases} 0 & \text{if } p \in \{\text{Linke, AfD}\} \\ 0.25 & \text{else} \end{cases}$$

Classifier	Size	pre MSE	pre MAE	post MSE	post MAE
Gemma-2	2B	0.053	0.185	0.043	0.172
Gemma-2	9B	0.047	0.186	0.043	0.172
DeBERTa-l	435M	0.055	0.197	0.047	0.182
Llama-3.2	1B	0.061	0.209	0.053	0.183
GottBERT_l	357M	0.071	0.225	0.061	0.208
gbert-large	337M	0.062	0.196	0.062	0.192
gelectra-l	336M	0.066	0.211	0.067	0.202
xlm-roberta-l	561M	0.080	0.225	0.080	0.218
Llama-3.2	3B	0.087	0.250	0.082	0.248
gemma-3	1B	0.202	0.371	0.125	0.287
EuroBERT	610M	0.249	0.419	0.133	0.304
EuroBERT	210M	0.143	0.304	0.133	0.299
EuroBERT	2.1B	0.146	0.305	0.144	0.307

Table 3: Pre versus post optimization comparison of the used models by mean squared error (MSE) and mean absolute error (MAE), ordered by post MSE

Results

Findings

Upon reviewing the results, we conclude that our transformer models, when utilized with vectors, effectively identify political stances in German texts. The accuracy of our classifier closely aligns with that of public left/right polls, suggesting that its outcomes are consistent with those of human raters. Additionally, we found that a model's size does not necessarily guarantee effectiveness across all scenarios, as depicted and further explained in Figure 7. For instance, both Llama 3.2 models (with 1B and 3B parameters) performed worse on in- and out-of-domain classification than the significantly smaller DeBERTa-large model

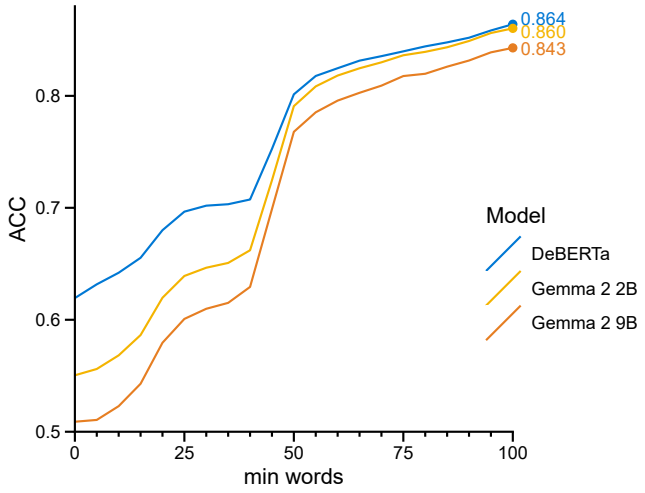

Figure 4: Classifier performance on tweets

(435M). The DeBERTa-large model achieved the highest in-domain performance, with an F_1 score of 0.84, while Gemma2-2b demonstrated the best out-of-domain results both before and after optimization, as illustrated Table 3.

In-Domain Model Performance

To evaluate the model's in-domain performance, we used the 20% Bundestag and Wahlomat test set parts. The models that performed best were DeBERTa-large ($F_1 = 0.84$), Gemma-2-9b ($F_1 = 0.79$), and EuroBERT-610m ($F_1 = 0.79$).

Out-of-Domain Model Performance on Tweets

The out-of-domain evaluation was carried out using posts from members of the German Bundestag. It is important to note that our knowledge of each tweet is limited to its author and their associated party; we do not have information about the tweet's political stance or how it might be received by others. Consequently, our evaluation focused solely on the accuracy of the three top-performing classifiers from the earlier in-domain classification task. We investigated a strong correlation between tweet length and the accuracy of our classifier of $0.96 \leq r \leq 0.97$ depending on the used model. This aspect is crucial, as tweets are often quite brief, and in some cases, the author may require the reader to rely on external context to fully understand a quote.

As illustrated in Figure 4, the accuracy for shorter tweets ranges between 50% and 65%. However, this accuracy increases to over 80% when tweets contain 50 or more words.

Out-of-Domain Model Performance on Newsmedia

The out-of-domain evaluation on newspapers was conducted using 33 German news outlets, ranging from left-wing publications like Jungle World to right-wing ones, such as Compact. Additionally, only news articles with a political score of 0.8 or higher were further processed and reviewed using a different model. The score was calculated as the MAE combined with the percentage relative to the total scale of 2 [-1, 1], providing a clearer perspective. The top-performing models were Gemma2-2b with an MAE of 0.1852 (9.26%), Gemma2-9b with an MAE of 0.1859 (9.29%), and gbert-large with an MAE of 0.1965 (9.82%). Notably, the ranking differs from the in-domain results, suggesting that some models generalize text more effectively than others.

Effect of the Vector Optimization

Following the initial out-of-domain test, we aimed to determine whether a different vector alignment could minimize the mean absolute error. We conducted a numerical optimization that allowed us to adjust each vector by ±0.25, except for the extreme positions held by Die Linke and AfD, which had to remain fixed to ensure the model's functionality. Our findings revealed that the optimizer successfully minimized the mean absolute error while keeping the adjustments within the specified range. Figure 3 illustrates the party vectors before and after the optimization process. Notably, the Grüne party shifted further to the left, while the FDP and CDU moved to the right. Post-optimization, each model possessed its own tailored set of vectors for all six parties.

After optimization, all models exhibited an average decrease in mean absolute error of 0.0239, corresponding to 1.1946%. EuroBERT-610m demonstrated the most substantial reduction, with a decrease of 0.304 or 5.73%, whereas EuroBERT-2.1B was the only model to perform worse, with a decline of 0.0016 or 0.08%.

Discussion

To train multiple BERT, Llama, and Gemma LLMs, a wide variety of data was used, including minutes from the German Bundestag and statements from a digital voters' guide (Wahlomat). Unlike the Bundestag data, which captures attitudes implicitly (e.g., via emotions), the Wahlomat dataset is based on explicit information provided directly by the parties. In order to further increase the (linguistic) variance, both datasets were artificially enriched. The models' multi-label responses were translated into a numerical continuum from -1 (left) to 1 (right). Performance was tested in (test set) and out of the domain (independent newspaper dataset and tweets).

We have observed that a trained multi-label transformer model, when combined with the appropriate party vector projection, can recognize political stances at a level comparable to that of polls for political classification. During our search for the optimal transformer model for our classifier, we discovered that the best-performing model within the specific domain (DeBERTa-large with an F1 score of 0.84) does not necessarily translate to being the best for other domains. In the out-of-domain test, DeBERTa-large was outperformed by Gemma-2-2B, indicating that a combination of model size, architecture, and training data is crucial for achieving superior out-of-sample accuracy.

The German-pretrained DeBERTa-large model demonstrated superior performance in in-domain classification, attributable to its extensive training corpus, as detailed in the models section of the methodology chapter. Unlike the conventional approach of relying solely on the OSCAR dataset

(Ortiz Suárez, Sagot, and Romary 2019), the developers of the German DeBERTa model (Dada et al. 2023) leveraged a more diverse dataset encompassing multiple domains. This strategic selection, incorporating formal, informal, legal, medical, and literary texts, enhanced the model's in-domain efficacy, underscoring the importance of data diversity in model training.

A compelling reason for the outstanding performance of the Gemma2-2b and Gemma2-9b models (Riviere et al. 2024) in the out-of-domain classification task is their unique training paradigms. By employing knowledge distillation from larger models, these models function as effectively condensed experts. This approach enables them to generalize significantly better than other models within the same size category.

Additionally, we found that providing the vectors from our initial approach to an optimizer, allowing it some flexibility in adjusting those vectors, can further enhance out-of-domain accuracy. Our method represents a novel approach compared to many existing techniques, which rely on discrete labels. Despite the evolving political landscape, our proposed methodology eliminates the need for manual labeling. This enables us to conduct periodic retraining by updating only the training datasets for our models.

Practical Implications

It is important to acknowledge that individuals exhibit biases to varying degrees. However, being aware of these biases can significantly enhance our understanding of the world. A classifier like the one introduced in this paper can assist in categorizing news outlets, authors, discussion threads, and various conversations, helping to prevent us from becoming trapped in echo chambers and encouraging a broader perspective. Additionally, a browser plugin could display the bias of every newspaper we visit.

An implementation could involve tracking a rolling score over a day or week to monitor the trajectory of topics, newspapers, or discussion threads. Warning signals could be triggered if a news outlet remains too entrenched in one extreme for an extended period.

Another application could involve conducting targeted searches for discussions aligned with left- or right-wing ideologies to analyze the topics with which specific groups engage. This would facilitate a deeper understanding of discourse patterns and sentiment within these ideological communities.

Our approach is highly adaptable to different countries and use cases because we avoid manual labeling and extract the political spectrum directly from the data. For instance, if there is a political shift in Germany, we can easily accommodate this by retraining the model. In contrast, comparable projects that rely on manual labeling require a significantly larger workforce. Our primary requirements are political texts from parties and fixed reference points, such as newspapers. A well-distributed representation of political parties within the spectrum is particularly beneficial, as it facilitates coverage of various points by combining those vectors. In contexts with only two major parties, achieving fine-grained classification can be more challenging because there is no liberal segment between them.

Our paper can also serve as a foundation for further research on social media. For instance, if a researcher has collected sufficient social media data and aims to track political shifts before, during, and after a governmental transition, they can use our tool.

Adapting the tool for other languages and cultural contexts is also possible. We introduce a new method for effectively classifying texts along the political spectrum, using comments and other indicators as proxies to construct a training corpus.

Limitations

The limitations can be categorized as model-related and methodological limitations. In our case, model-related limitations arise from using a classification model rather than a reasoning-based transformer model. At times, when quoting individuals, it can be challenging for the model to discern whether the quoted opinion is being critiqued. For example, consider the tweet: " 'Those who want human society must overcome male society.'Svenja Schulze (February 17, 2022) This was or is also stated in the SPD party program.". Based on the available information, we cannot determine whether the tweet's author agrees or disagrees with the statement. The classifier will categorize it as left-wing solely on the basis of the quoted content.

Another limitation of our analysis's non-reasoning capabilities lies in its evaluation of foundational statements. For instance, when we input Pierre-Joseph Proudhon's statement "Property is theft" (Proudhon 1840) into our gemma-2-9b classifier, it produces a score of 0.78, classifying it as a right-wing assertion. Despite Proudhon's typically being recognized as a figure of the libertarian left (Honeywell 2021; Levy and Adams 2019), some scholars have controversially interpreted him as a precursor to fascism (Schapiro 1945; Krier 2009). In this scenario, it would be intriguing to understand why the classification model categorized the statement as right-wing. However, because we are not using a reasoning model, we cannot examine the rationale for the classification. A possible explanation for the misclassification in this example is that such ideological statements are neither discussed in parliament nor part of the political agenda of mainstream political parties.

Given the increase in accuracy associated with a higher number of words per tweet, we conclude that the models may struggle with very short texts, particularly when background knowledge is required to interpret their meaning.

From a methodological standpoint, a one-dimensional projection may not offer enough entropy to accurately map political views in certain instances. This becomes evident when we examine Figure 1. One might intuitively assume that the left-most party, Die Linke, and the right-most party, AfD, exhibit the greatest distance. However, the negative correlations between the right-wing AfD and the Grüne (-0.32) and SPD (-0.34) parties are significantly stronger than that between the AfD and Die Linke (-0.18).

One possible interpretation for this phenomenon is that both the AfD and Die Linke respond to certain issues in sim-

ilar ways, albeit for distinct reasons. For instance, consider the issue of supplying arms to a nation under attack by its neighbor. A left-wing party such as Die Linke would oppose these arms deliveries from a pacifist perspective, whereas the right-wing AfD would object from a nationalist perspective, prioritizing Germany's interests and its trade relations with the aggressor state.

Another possible explanation is that both the left-wing Die Linke and the right-wing AfD are not part of the governing coalitions and instead belong to the opposition. In this role, opposition parties commonly critique the governing parties' policies and actions. Consequently, this can lead to the formation of similar opinions on both ends of the political spectrum.

The political shift occurring in numerous societies has been noted, but it also presents limitations. What may be viewed as slightly left or right today could be regarded as a liberal stance in the near future. To address this issue, it is recommended that the classifier be retrained with updated data whenever such a shift becomes apparent or at regular intervals.

A classifier's applicability is inherently limited to the linguistic and cultural context in which it was trained. For instance, the fundamentally divergent perspectives on public health care and labor rights observed in Germany and the United States underscore the pitfalls of deploying a classifier across different cultural frameworks without appropriate adaptation. The cultural context not only establishes limits for the classifier, but the diversity of the input data also plays a significant role. Positions further to the left of the party Die Linke, as well as those to the right of the party AfD, cannot be distinguished from the positions of these two parties since the scale based on the six parties is bounded by these parties.

It is essential to note that we employ a classification-based approach, which inherently lacks interpretability and reasoning. The classifier cannot provide insights into why it categorizes a specific text as belonging to the Grüne or SPD parties. Additionally, it does not explicitly model underlying principles of left- or right-wing politics, as it is not designed to serve as a reasoning model. While it effectively categorizes topics in the training data, it may struggle to classify novel concepts, as it lacks the capacity to reason about them. In summary, while the classifier can make errors, it should not be relied upon to block texts without additional verification or similar measures. To minimize the impact of misclassification, it is advisable to apply the classifier to a large volume of texts, such as those in a newspaper, as illustrated in the given example. This way, the significance of any single mistake is diminished.

Future Work

To address the limitations discussed above, it would be advantageous to develop a classifier capable of explaining why it assigns a text to a specific position on the left-right spectrum. This feature would help users understand the rationale for classifying a text as far-right or far-left and would also facilitate the identification of potential errors. Users could contest the classifier's assessment and dismiss the result in cases of inaccuracies, rather than accepting its conclusions unquestioningly. Additionally, a reasoning model would enhance its utility by enabling it to apply foundational concepts of left, liberal, and right-wing politics to new topics not encountered in its training data. Improved explainability and generalizability could benefit future developments.

Social Impact and Misuse

The developed models, once sufficiently advanced for everyday use, could potentially pose risks if individuals rely on them to filter news. Instead of expanding their perspectives, users might choose to exclude all news that exceeds a certain threshold in a political direction, whether left or right. As a result, what is intended to be a useful application could also be misused.

A second potential misuse of the model is the risk of discrimination against individuals based on their political beliefs. On a broader level, institutions could utilize the model to monitor live social media streams and identify individuals expressing dissenting opinions.

We firmly oppose all forms of discrimination on the basis of political viewpoints and strongly advocate for free speech. Additionally, we are committed to transparency by making all training data publicly available, ensuring that our models and their results are clear and accessible.

Ethical Statement

All of our training data has been sourced exclusively from publicly available materials and is intended solely for academic use. We did not bypass any safety mechanisms or use data behind a paywall. Additionally, we have not collected any personal data, except for political speeches by public figures.

We emphasize that the introduced classification method is still in its early stages, and errors cannot be discounted. Furthermore, our classifier should not be used to evaluate or discriminate against individuals on the basis of their political beliefs. We strongly advocate for a diverse political landscape and uphold the principles of free speech in respectful interactions, free from personal attacks.

Acknowledgments

The authors would like to thank the System Sciences Chair for Communication Systems and Network Security under the direction of Prof. Dr. Gabi Dreo Rodosek.

The authors acknowledge the financial support from the Federal Ministry of Education and Research of Germany in the program "Souverän. Digital. Vernetzt." Joint project 6G-life, project identification number: 16KISK002.

References

Aksenov, D.; Bourgonje, P.; Zaczynska, K.; Ostendorff, M.; Moreno-Schneider, J.; and Rehm, G. 2021. Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments. In Mostafazadeh Davani, A.; Kiela, D.; Lambert, M.; Vidgen, B.; Prabhakaran, V.; and Waseem, Z., eds., *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*, 121–131. Stroudsburg, Penn.: Association for Computational Linguistics.

AllSides. 2025. AllSides: Unbiased Balanced News. https://www.allsides.com/. Accessed 2025-09-10.

Andrzejewski, C. 2023. Team Jorge: In the heart of a global disinformation machine. https://forbiddenstories.org/team-jorge-disinformation. Accessed 2026-01-15.

Baly, R.; Da San Martino, G.; Glass, J.; and Nakov, P. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4982–4991. Online: Association for Computational Linguistics.

Baly, R.; Karadzhov, G.; Saleh, A.; Glass, J.; and Nakov, P. 2019. Multi-Task Ordinal Regression for Jointly Predicting the Trustworthiness and the Leading Political Ideology of News Media. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 2109–2116. Minneapolis, Minn.: Association for Computational Linguistics.

Boizard, N.; Gisserot-Boukhlef, H.; Alves, D. M.; Martins, A.; Hammal, A.; Corro, C.; Hudelot, C.; Malherbe, E.; Malaboeuf, E.; Jourdan, F.; Hautreux, G.; Alves, J.; Haddad, K. E.; Faysse, M.; Peyrard, M.; Guerreiro, N. M.; Fernandes, P.; Rei, R.; and Colombo, P. 2025. EuroBERT: Scaling Multilingual Encoders for European Languages. arXiv:2503.05500.

Bolte, F. 2025. qual-o-mat-data. https://github.com/gockelhahn/qual-o-mat-data. Accessed 2025-09-10.

Chan, B.; Schweter, S.; and Möller, T. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796. Barcelona, Spain: International Committee on Computational Linguistics.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv:2003.10555.

Cohen, R.; and Ruths, D. 2013. Classifying Political Orientation on Twitter: It's Not Easy! In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, volume 7, 91–99. Cambridge, Mass.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Online: Association for Computational Linguistics.

Cumming, G. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Dada, A.; Chen, A.; Peng, C.; Smith, K.; Idrissi-Yaghir, A.; Seibold, C.; Li, J.; Heiliger, L.; Friedrich, C.; Truhn, D.; Egger, J.; Bian, J.; Kleesiek, J.; and Wu, Y. 2023. On the Impact of Cross-Domain Data on German Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13801–13813. Singapore: Association for Computational Linguistics.

Deutscher Bundestag. 2025. Open Data. https://www.bundestag.de/services/opendata. Accessed 2025-09-10.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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, 4171–4186. Minneapolis, Minn.: Association for Computational Linguistics.

Dubey, A.; Jauhri, A.; and Pandey, A. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Erhard, L.; Hanke, S.; Remer, U.; Falenska, A.; and Heiberger, R. H. 2025. PopBERT. Detecting Populism and Its Host Ideologies in the German Bundestag. *Political Analysis*, 33(1): 1–17.

Fagni, T.; and Cresci, S. 2022. Fine-grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning. *Journal of Artificial Intelligence Research*, 73: 633–672.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gemini Team Google. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DeBERTa: Decoding-enhanced bert with disentangled attention. arXiv:2006.03654.

Honeywell, C. 2021. *Anarchism*. Key Concepts in Political Theory. Cambridge, United Kingdom: Polity Press.

Jakob, C.; Wenzel, P.; Mohtaj, S.; and Schmitt, V. 2024. Augmented Political Leaning Detection: Leveraging Parliamentary Speeches for Classifying News Articles. In Klamm, C.; Lapesa, G.; Ponzetto, S. P.; Rehbein, I.; and Sen, I., eds., *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, 126–133. Vienna, Austria: Association for Computational Linguistics.

Jiang, J.; Ren, X.; and Ferrara, E. 2023. Retweet-BERT: Political Leaning Detection Using Language Features and Information Diffusion on Social Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1): 459–469.

Kawintiranon, K.; and Singh, L. 2022. PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7360–7367. Marseille, France: European Language Resources Association.

Kiesel, J.; Mestre, M.; Shukla, R.; Vincent, E.; Adineh, P.; Corney, D.; Stein, B.; and Potthast, M. 2019. SemEval-2019

Task 4: Hyperpartisan News Detection. In May, J.; Shutova, E.; Herbelot, A.; Zhu, X.; Apidianaki, M.; and Mohammad, S. M., eds., *Proceedings of the 13th International Workshop on Semantic Evaluation*, 829–839. Minneapolis, Minn.: Association for Computational Linguistics.

Krier, F. 2009. *Sozialismus für Kleinbürger: Pierre Joseph Proudhon - Wegbereiter des Dritten Reiches*. Köln: Böhlau.

Levy, C.; and Adams, M. S., eds. 2019. *The Palgrave Handbook of Anarchism*. Cham: Springer.

Liu, Y.; Zhang, X. F.; Wegsman, D.; Beauchamp, N.; and Wang, L. 2022. POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 1354–1374. Seattle, Wash.: Association for Computational Linguistics.

Maurer, M.; Kruschinski, S.; and Jost, P. 2024. Fehlt da was? Perspektivenvielfalt in den öffentlich-rechtlichen Nachrichtenformaten. Technical report, Johannes Gutenberg-Universität Mainz, Institut für Publizistik, Mainz.

Mediabiasfactcheck.com. 2025. Media Bias/Fact Check - Search and Learn the Bias of News Media. https://mediabiasfactcheck.com/filtered-search/?country=DE. Accessed: 2026-01-04.

Medienkompass.org. 2025. Deutsche Medienlandschaft. https://medienkompass.org/deutsche-medienlandschaft/. Accessed: 2025-07-07.

Nguyen, D. Q.; Vu, T.; and Tuan Nguyen, A. 2020. BERTweet: A pre-trained language model for English Tweets. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14. Online: Association for Computational Linguistics.

Ortiz Suárez, P. J.; Sagot, B.; and Romary, L. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Bański, P.; Barbaresi, A.; Biber, H.; Breiteneder, E.; Clematide, S.; Kupietz, M.; Lüngen, H.; and Iliadi, C., eds., *Proceedings of the Workshop on Challenges in the Management of Large Corpora*, 9–16. Cardiff, United Kingdom: Leibniz-Institut.

Ostendorff, M.; Blume, T.; and Ostendorff, S. 2020. Towards an Open Platform for Legal Information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 385–388. Online: Association for Computing Machinery.

Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, 729–740. Vancouver, Canada: Association for Computational Linguistics.

Proudhon, P.-J. 1840. *Qu'est-ce que la propriété?, ou, Recherches sur le principe du droit et du gouvernement: Premier mémoire*. Brocard.

Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; and Bhupatiraju, S. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.

Schapiro, J. S. 1945. Pierre Joseph Proudhon, Harbinger of Fascism. *The American Historical Review*, 50(4): 714.

Scheible, R.; Frei, J.; Thomczyk, F.; He, H.; Tippmann, P.; Knaus, J.; Jaravine, V.; Kramer, F.; and Boeker, M. 2024. GottBERT: a pure German Language Model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21237–21250. Miami, Fla.: Association for Computational Linguistics.

Schneider, S. 2025a. german_ideology_prediction. https://github.com/SinclairSchneider/german_ideology_prediction. Accessed 2025-09-15.

Schneider, S. 2025b. trainset_political_party_big. https://doi.org/10/qvxb. Revision 444f2da, Accessed 2025-09-10.

Schneider, S. 2025c. trainset_political_text_yes_no_german. https://doi.org/10/qvw9. Accessed 2025-09-12.

Springer. 2025. BILD.de. https://www.bild.de. Accessed 2025-12-22.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. In Calzolari, N.; Choukri, K.; Declerck, T.; Doğan, M. U.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–2218. Istanbul, Turkey: European Language Resources Association (ELRA).

Volf, M.; and Simko, J. 2025. Political Leaning and Politicalness Classification of Texts. arXiv:2507.13913.

Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; Huang, F.; and Zhou, J. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176.

Zhuang, L.; Wayne, L.; Ya, S.; and Jun, Z. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In Li, S.; Sun, M.; Liu, Y.; Wu, H.; Liu, K.; Che, W.; He, S.; and Rao, G., eds., *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 1218–1227. Huhhot, China: Chinese Information Processing Society of China.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see the Ethical Statement
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Methods section
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Methods, Subsection *Dataset*

(e) Did you describe the limitations of your work? Yes, see the Discussion, Subsection *Limitations*

(f) Did you discuss any potential negative societal impacts of your work? Yes, see the Discussion, Subsection *Practical implications*

(g) Did you discuss any potential misuse of your work? Yes, see the Discussion, Subsection *Social impact and misuse*

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see the Discussion, Subsection *Social impact and misuse*

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see the linked GitHub repository, folder 05_train_new_model

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, see Figure 7

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Table 1

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the Methods

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see the Discussion, Subsection *Limitations*

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, see the References and the Methods, Subsection *Dataset*

(b) Did you mention the license of the assets? Yes

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we included links to the code and datasets

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see the Ethical Statement

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see the Ethical Statement the Methods, Subsection Datasets

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, see the model cards on HuggingFace

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendices

Tables and Figures

No	Linke	B'90	SPD	FDP	CDU	AfD
1	yes	yes	yes	no	no	no
2	yes	yes	yes	yes	yes	no
3	yes	yes	yes	no	no	no

Table 4: (Dis)agreement of various parties regarding three exemplary statements: 1. *A tax is to be reintroduced on high net worth individuals*, 2. *Germany should keep the euro as its currency.*, 3. *A minimum wage should be introduced.*

	Source A		Source B				Source C			
Media	x	$z(x)$	x	y	PC_1	$z(PC)$	ord.	x	$z(x)$	Size
Achgut	5.20	1.19								72.1K
Augsburger Allgemeine			2.74	2.87	0.46	0.46				
Bayerische Rundfunk	4.40	0.39	2.62	3.17	0.69	0.69	0	-1.70	-0.62	40.7K
Berliner Zeitung			2.47	2.66	-0.17	-0.17	-1	-3.00	-1.01	
Bild (Springer)	5.20	1.19	2.89	3.49	1.46	1.46	1	3.60	0.97	1.47M
Cicero	4.90	0.89					1			13.5K
Compact	6.00	1.99								11.8K
der Freitag	2.70	-1.31								26.8K
Deutschlandfunk	3.80	-0.21	2.94	2.44	0.18	0.19	0			89.0K
FAZ	4.50	0.49	2.85	2.47	0.11	0.11	1	3.40	0.91	316.0K
Focus	4.90	0.89	3.71	2.89	1.78	1.79	1	2.80	0.73	33.5K
Frankfurter Rundschau	3.40	-0.61	2.22	1.98	-1.38	-1.38	-1			165.0K
General-Anzeiger Bonn			2.67	2.75	0.22	0.22	0	1.50	0.34	
Handelsblatt	4.30	0.29	3.69	2.58	1.36	1.36	1	2.10	0.52	
junge Freiheit	5.80	1.79	3.14	3.97	2.40	2.41				56.0K
junge Welt	2.40	-1.61	1.84	2.14	-1.67	-1.67	-2			91.7K
Jungle World	2.30	-1.71								57.7K
Linksunten (indymedia)	2.00	-2.01								74.0K
MDR	4.10	0.09	2.69	2.35	-0.28	-0.28				63.2K
MM News	5.10	1.09								197.0K
Münchner Merkur			2.75	2.98	0.61	0.61	0	1.50	0.34	
NDR	3.70	-0.31								35.2K
Neues Deutschland	2.60	-1.41	1.71	1.80	-2.28	-2.29	-2			283.0K
NTV	4.30	0.29	2.56	2.60	-0.13	-0.13	1			548.0K
NachDenkSeiten	3.10	-0.91								20.4K
RT Deutsch	5.10	1.09								44.2K
RTL	4.50	0.49	2.50	3.14	0.49	0.49				187.0K
Rheinische Post			2.49	2.49	-0.36	-0.36	1	2.30	0.58	
Saarbrücker			2.72	2.76	0.30	0.30	0	-0.70	-0.32	
Spiegel	3.50	-0.51	2.40	2.37	-0.63	-0.63	-1			1.11M
Stern	3.80	-0.21								414.0K
Süddeutsche	3.50	-0.51	2.36	2.34	-0.73	-0.73	-1	-3.40	-1.13	1.89M
T-Online			2.75	2.62	0.16	0.16	-1	-2.20	-0.77	
TAZ	2.80	-1.21	1.86	1.98	-1.86	-1.87	-2			725.0K
Tagesschau (ARD)	3.70	-0.31	2.59	2.66	0.00	0.00	0			55.8K
Tagesspiegel	3.60	-0.41	2.50	2.74	-0.03	-0.03				1.20M
Tichys	5.50	1.49	3.62	3.99	3.07	3.08	2	6.80	1.93	33.3K
Vice	2.80	-1.21								53.8K
WDR	3.50	-0.51	2.00	2.73	-0.70	-0.70				45.8K
Web.de			2.56	2.60	-0.12	-0.12	-1	-2.20	-0.77	
Welt	4.80	0.79	3.11	3.29	1.49	1.50	1	3.60	0.97	164.0K
Zeit	3.60	-0.41	2.47	2.71	-0.11	-0.11	-1			343.0K

Table 5: Overview of the German media landscape, including several online versions of newspapers like *Frankfurter Allgemeine Zeitung* (FAZ), *die tageszeitung* (TAZ); television channels like *Mitteldeutscher Rundfunk* (MDR), *Norddeutscher Rundfunk* (NDR), *Westdeutsche Rundfunk* (WDR), *Radio Télévision Luxembourg* (RTL), and various other online news media formats. Media bias estimates were collected from three sources: A. ratings of $k = 39$ media outlets from $n = 1148$ participants on a seven-point Likert scale from *extrem left* (1) to *extrem right* (7), provided by Medienkompass.org (2025); B. ratings of $k = 47$ media outlets from only $n = 9$ extensively trained raters on two correlated five-point scales, provided by Maurer, Kruschinski, and Jost (2024); and C. ratings regarding $k = 77$ outlets rated on a scale from *extrem left* (-10) to *extreme right* (10) retrieved from Mediabiasfactcheck.com (2025). For source B, both correlated scales were reduced using principal components analysis (PCA), yielding one principal component (PC). Numeric ratings were z–transformed for comparability (standardised, i.e. $M = 0$, $SD = 1$). Regarding source C, numeric scores were missing for various media outlets; ordinal (ord.) scores were estimated from the given labels accordingly. Appropriate association estimates for pairwise completed cases showed high correlations, indicating convergent validity. Based on the media list from source A, $k = 33$ (approx. 85%) media outlets were scraped, yielding a dataset of approx. 10M articles; 74K per outlet on average.

Type	Statement
original	A tax is to be reintroduced on high net worth individuals.
child	Soon there will be a tax again for people who have a lot of money. This means that they will have to give part of their money to the government. The government can then use this money to buy important things for everyone.
teenager	The federal government plans to reintroduce a tax on high net worth individuals. This means that people with large fortunes will have to hand over part of their wealth to the state. The revenue will be used for important public tasks.
adult	The reintroduction of a wealth tax for high net worth individuals is currently under discussion. This measure is intended to help increase government revenue and reduce social inequality. By taxing large fortunes, the government can finance important investments in infrastructure and social programs.
eloquent	There are growing calls for a tax to be levied on high net worth individuals in order to combat social inequality in our society. This measure could make an important contribution to the redistribution of wealth.
tweet	A wealth tax would also stimulate the economy by promoting consumption and investment. A positive effect for everyone!

Table 6: Example of paraphrasing an original statement in the words of different persona

Figure 5: Exemplary statement 1/38: *Germany should continue to provide military support to Ukraine*, sourced from the Wahlomat service regarding the German federal elections in 2025 (www.wahl-o-mat.de/bundestagswahl2025). Screenshot a shows the user view with response options (approval, neutral, disapproval), b depicts the stance of selected parties (disapproval by the most left-wing and right-wing parties, approval by the others).

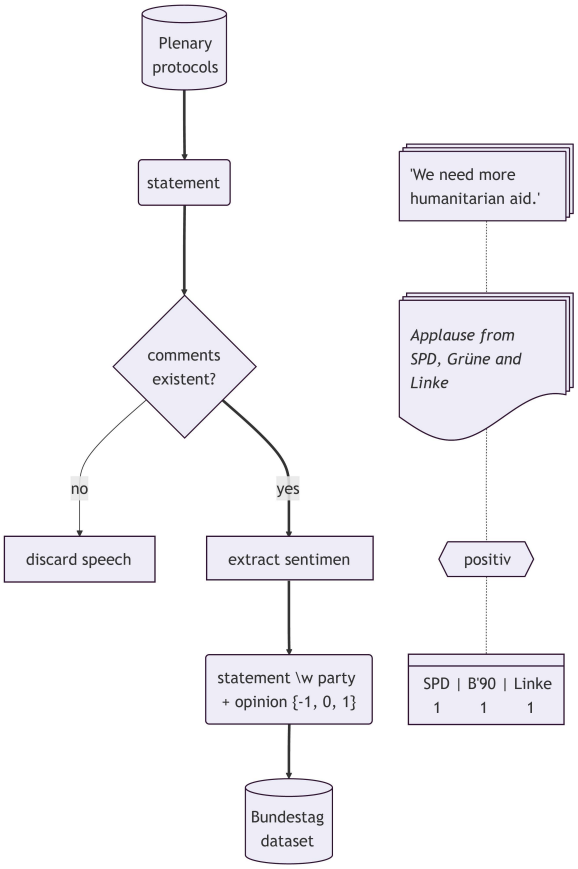

Figure 6: Flowchart of sentiment extraction

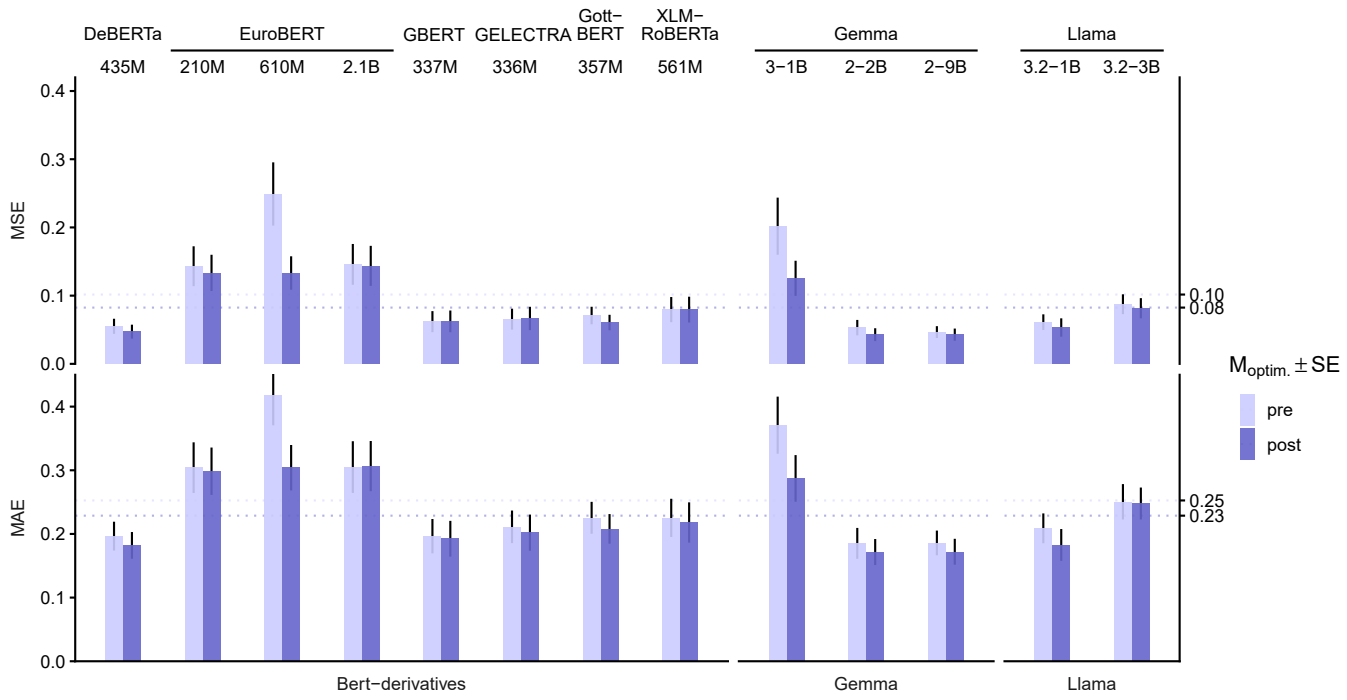

Figure 7: Depicted is the effect of optimization across all 13 models and 33 news media outlets, measured using the mean absolute error (MAE, lower panel) and mean squared error (MSE, upper panel). Error bars represent standard errors (SE). Values are sorted by model class and, within each class, by parameter size for various Gemma, Llama, and Bert derivatives. The color contrast highlights the effect of the optimization: values before optimization (light bars) are generally higher than those after optimization (dark bars). This reduction in error metrics due to the optimization is evident from the dashed horizontal lines, which represent the mean values across the models. The differences indicate moderately strong effects, which we report as d_{av} according to Cumming (2013) with 95% confidence intervals (CI). Specifically, the optimization had an estimated effect of $d_{av} = 0.37$, $CI_{95\%}[0.08, 0.66]$ as measured by the MAE, and an only slightly smaller $d_{av} = 0.36$, $CI_{95\%}[0.00, 0.73]$ with respect to the MSE. These effects were largely consistent across models and metrics, as reflected in high pre-post correlations ($r_{\mathrm{MAE}} = .91$ and $r_{\mathrm{MSE}} = .88$). Two exceptions stand out: EuroBERT-610M and Gemma-3-1B, for which optimization had a stronger effect regardless of the metric considered. These are also the models whose initial values were clearly above the average values (cf. the upper dashed line in both panels). No clear effect of model size (in terms of the number of parameters) on performance is evident; for both metrics and measurement points, size and error correlated only weakly with $r \approx -.25$ (for all metrics before and after optimization, with only post-optimization MSE showing a slightly higher correlation of $r = -.27$). In other words, a higher number of parameters tends to produce smaller errors across models, though this does not necessarily hold true for individual models. For example, the smaller Llama-3.2 with 1B parameters consistently yields lower errors than the much larger model with 3B parameters. The results suggest that model size alone is not a reliable predictor. At this point, it should be noted that these findings are reported purely descriptively; our setup did not have the primary goal of demonstrating an effect of model size but rather aimed to identify the best model. Here, Gemma2-2B yielded the lowest errors, regardless of the optimization or metric. However, the error bars suggest that the performance of much smaller models such as GBERT-337 or DeBERTa-425M does not differ significantly. No pairwise tests were calculated.