Automated Political Stance Identification in Political Texts: A Transparent, Reproducible, Scalable and Theory-Driven Approach

Juan S. Gómez Cruces, Yorick Scheffler, and Ewan Thomas-Colquhoun

March 2026

This is a working paper currently under review. Please do not cite without the authors' permission.

This paper introduces a transparent, reproducible, scalable, and theory-driven method for automatically identifying political stances in text using the `Political DEBATE` model (Burnham et al. 2024), a light-weight natural language inference (NLI) classifier specialized for political discourse. Building on established political science theories, we adapt conceptual definitions from expert surveys into natural-language hypotheses which are evaluated against premise texts in a *zero-shot* setting—meaning the model makes predictions without any task-specific labeled examples—to identify ideological entailment. Using this approach, we classify texts along three key dimensions of political competition: economic left–right ideology, populist versus pluralist rhetoric, and support for liberal democratic values. We validate the method by applying the model to a diverse sample of political texts and comparing its outputs with expert-coded assessments. For the economic left–right and populism–pluralism dimensions, the model produces scores within the range of expert evaluations. Unlike supervised or fine-tuned approaches, our method requires no additional labeled training data, reduces computational and financial costs, and provides high transparency, as every stance score can be traced back to explicit hypotheses grounded in political science theories. By making political stance detection transparent, reproducible, scalable, and theory-grounded, this work strengthens the potential of researchers, practitioners, civil society organizations, policymakers, and the broader public to monitor, compare, and promote democratic accountability and public debate in an increasingly complex information environment.

Introduction

Social scientists have long been interested in understanding and measuring the political stances of parties and leaders using texts (see Laver and Garry 2000). With the emergence of powerful computational tools for natural language processing (NLP), these methods are increasingly relied upon to automatically extract meaningful information from large volumes of unstructured political texts, such as party manifestos, speeches, news reports, and social media posts (see Hou and Huang 2025; Marwala 2023). Computational approaches have complemented traditional methods by using NLP and machine learning to automate stance detection across various types of political texts. By applying techniques such as sentiment analysis, topic modeling, and text classification, researchers have aimed to identify political stances (Burnham 2025), track democratic backsliding (Mochtak 2025), and systematically analyze ideological positions (Nikolaev, Ceron, and

Padó 2023), among other tasks.

Following these attempts to take advantage of an increased amount of text available for analysis, this paper introduces a novel, transparent and theory-based method for leveraging large language models (LLMs), built on the transformers architecture, to automate the identification of political stances in texts across three different dimensions: economic left–right ideology, support for liberal democratic values, and populist/pluralist rhetoric. This approach builds on high-quality expert survey codebooks and employs a zero-shot approach using the `Political DEBATE` model (Burnham et al. 2024). By combining theory-driven hypothesis construction with a lightweight, auditable scoring procedure, our method aims to enrich the toolkit available to researchers, practitioners, and the general public, offering a means of accessing reliable, transparent information on the positions and policy preferences of political parties and leaders that existing supervised and generative approaches cannot provide at comparable cost and interpretability. We validate the accuracy of our method by applying the model to a sample of political texts and comparing its outputs with expert-coded assessments from a survey with 147 responses. The results show no significant differences between the model's scores and those of human experts, suggesting that our approach performs at least as well as expert-coded assessments.

In the following sections, we situate our work within the broader literature on political stance identification. Then we provide a detailed explanation of the model we draw from, including its training data, validation, and potential sources of bias, before evaluating its performance in comparison to other models and approaches. We then describe the process that translated expert codebooks into the model hypotheses providing for the conceptual transparency of our approach. As both an illustration and a validation exercise, we present the results of applying the model to a sample of 30 texts from speeches, party manifestos, and social media posts, and compare them with scores obtained from human experts. We conclude with a discussion of the potential applications and limitations of this approach, as well as the contributions to future research on political stance identification.

Political Stance Detection and Natural Language Processing

Within computational approaches to political science, stance detection[1] is defined as a classification task for identifying an actor's public position on a given target, usually political actors, policies, or values, using textual data (Burnham 2025; ALDayel and Magdy 2021). Distinct from sentiment analysis, which focuses on affective valence, identifying positive and negative *emotion*, stance detection seeks to identify an author's *position* to a given target through analysis of textual features (Bestvater and Monroe 2023; Küçük and Can 2020). The applications of such an approach in political contexts are myriad. A plurality of studies have, for instance, classified users' stances towards politicians based on tweets (Lai et al. 2017; Darwish, Magdy, and Zanouda 2017; Darwish, Stefanov, et al. 2020), detected fake news among news articles and other online texts (Lillie and Middelboe 2019; Ghanem, Rosso, and Rangel 2018), and categorized politicians' stances towards policy on the basis of their speeches on the US congressional floor, as well as at parliamentary debates across the United Kingdom and India (Rohit and Singh 2018; Abercrombie and Batista-Navarro 2022).

While previously computational approaches to stance detection relied on dictionaries, word-count features, and supervised classifiers, such as support vector machines and Naive Bayes (Küçük and Can 2020), the advent of artificial neural networks and, in particular, the rise of transformer-based architectures has seen researchers increasingly turn to deep-learning NLP for this task (Vaswani et al. 2023; Bahdanau, Cho, and Bengio 2016). These architectures have allowed researchers to approach larger, less-structured datasets, widening the range and quantity of data that can be analysed, and opening the possibility for novel subjects of focus and larger-N studies of political speech. Within such studies, three approaches have emerged which are most frequently applied for stance detection, including, in decreasing popularity, 1) supervised deep-learning using labeled corpora; 2) prompt-based in-context zero-shot classification using generative large language models;

[1] Taken here as synonymous with 'stance classification', 'stance identification', 'stance prediction', 'debate-side classification'.

and 3) natural language inference and entailment models trained for specific domains.

Each approach presents a distinct set of trade-offs, governed by empirical constraints and research priorities. Within political science approaches, however, six aspects emerge as critical for methodological and model selection: i) the availability and quality of labeled data; ii) model size and resulting compute and energy costs; iii) maximizing precision and recall attainment; iv) political domain specialization; v) output interpretability and inferential transparency; and vi) ethical suitability. Choosing any single approach necessarily represents a compromise across these factors. This section, therefore, introduces each of the three named approaches to stance detection, describing their application in previous studies of political stance detection, and evaluating their utility across these aspects for our defined usage case.

Supervised Text Classification

Supervised fine-tuning of large language models has become the most common approach for maximizing domain accuracy for text classification tasks within natural language processing, and has achieved human-level accuracy for labeling in the political domain across a number of proposed benchmarks (Howard and Ruder 2018). In this approach, a pretrained base language model, frequently the Roberta, Deberta and XLMBerta variants of Google AI's Bidirectional Encoder Representations from Transformers model (BERT), is fine-tuned on a labeled dataset that reflects the target task, pairing text snippets with classification labels (Devlin et al. 2019). Fine-tuning adapts these pretrained models to more closely match the downstream data distribution in subsequent output by updating model parameters to produce outputs which more closely fit the training data (Houlsby et al. 2019; Ding et al. 2023). Such approaches require high-quality base-labels for model fine-tuning with researchers applying existing datasets from the political science literature for pre-training, including collections of English-language parliamentary speeches (Bosley et al. 2023), as well as fine-tuning from the Ideological Books Corpus, Comparative Manifestos Project, and the Political Tweets Datasets (Chen, Mizuno, and Doi 2024).

With domain-specific training and fine-tuning, these models can achieve accuracy beyond that of general models using in-context training, and as smaller models are more compute efficient per inference, reducing resource requirements of researchers and ethical concerns around energy use (e.g., 110 million parameters for BERT vs 8 billion for base Llama variants). Yet this approach is not without issues. Producing quality labeled corpora for novel applications requires significant human time investment, with accuracy still an issue within areas requiring contextual understanding above the linguistic – such as fake news detection. Finally, and crucially, these gains come with an interpretability cost: the internal activations and distributed parameter patterns that produce a given prediction are not directly interpretable, with the "black box" problem a frequently cited issue within machine learning research. Post-hoc explanation methods exist, but provide limited, sometimes misleading guarantees, making it hard to trace which learned features or token interactions actually drove a decision — a fundamental obstacle for auditing, debugging, and trusting machine learning systems in sensitive applications (Lipton 2016; Doshi-Velez and Kim 2017).

Generative Pre-Trained Transformers and Zero-Shot Classification

With the expansion in the capability of LLMs developed using generative pre-trained transformers (GPTs), researchers are increasingly exploring the zero-shot prompt-based capabilities of models such as OpenAI's ChatGPT, Anthropic's Claude, Google's Gemini, and Meta's Llama, for text classification tasks. *Zero-shot classification* refers to the ability of a model to make predictions on tasks or categories it has never been explicitly trained on, relying instead on broad pre-trained knowledge and flexible instruction (Brown et al. 2020). In contrast, *few-shot* or *single-shot classification* provides the model with one or a small number of labeled examples to guide its predictions before it is applied to new cases; this can substantially improve performance on domain-specific or ambiguous tasks.

Zero-shot and single-shot approaches can mitigate the data-availability constraints of supervised systems by leveraging broad, pre-trained knowledge and flexible instruction

tuning to perform classification with little or no task-specific annotation (Brown et al. 2020; Liu and Shi 2025). This makes such methods expedient for exploratory classification and for generating candidate labels with minimal engineering, meaning they are more readily applicable to novel domains and accessible to researchers without programming experience (Wang, Qu, and Ye 2024). Their principal drawbacks for text classification are reproducibility and auditability; model outputs vary with prompt formulation, temperature and model version, and closed-weight models preclude making attempts at replication. Proprietary models with closed weights and training parameters, such as ChatGPT, have also seen concerns raised around the legality of the company's acquisition of training data, a key ethical issue researchers should be aware of before deploying such models (Cole and Oltermann 2025; Grynbaum and Mac 2023).

Moreover, while contemporary generative models can be prompted to provide reasoning for decisions made, the rhetorical explanations such models produce are not structured evidence in the sense that academic work can reliably review; they are plausible-sounding natural-language justification rather than decomposed entailment scores linked to explicit data. Further drawbacks also concern their efficiency. Their accuracy can still struggle in comparison to smaller domain-specific models, particularly in the political domain, whilst using significantly more compute and energy resources (Bosley et al. 2023; Wang, Qu, and Ye 2024). The efficient use of domain-specialized LLMs with model parameters magnitudes smaller than the latest GPTs is increasingly the trend of cutting-edge research. Hsieh et al. (2023) show that a 770M parameter can outperform a 540B parameter using only 80% of available training data by leveraging LLM-generated rationales as additional supervision. In another study, Lehman et al. (2023) demonstrate that smaller models pre-trained on clinical text substantially outperform much larger general-purpose LLMs across clinical NLP tasks — even under limited annotation conditions. Finally, ZHAO et al. (2023) argue that applying general-purpose LLMs in specialized domains is limited by domain-specific data heterogeneity, complex knowledge structures, unique objectives, and regulatory and ethical constraints. These authors also highlight a growing shift toward smaller, domain-specialized LLMs that offer greater efficiency and better alignment

with specialized requirements than large, general GPT-scale models.

Natural Language Inference (NLI) for Classification

Natural Language Inference (NLI) models offer something of a compromise between these two approaches. Developed as a baseline task for evaluating the performance of models built upon the transformers architecture, NLI describes the process whereby a LLM must identify whether a logical relationship exists between a *premise* sentence and a *hypothesis* sentence. Given the pairing of premise-hypothesis, an NLI model assigns one of (typically) three labels:

- **Entailment:** Hypothesis logically follows from premise.
- **Contradiction:** Hypothesis is logically incompatible with premise.
- **Neutral:** Hypothesis is neither entailed nor contradicted by premise (insufficient information).

Developing models capable of producing entailment labels follows the same process of supervised learning as described before — pre-training base transformer models such as BERT and ROBERTa on standard datasets such as SNLI and MulitNLI — before fine-tuning on a subset of data specialized to the given domain. Finally, within implementation, hypothesis prompts are engineered to best fit the scenario required, in this case the political axes which are the focus of our defined task.

The benefit of such an approach, as Burnham et al. (2024) argue, is that "nearly any classification task can be broken down into this structure [...] a single language model trained for natural language inference can function as a universal classifier and label documents across many dimensions without additional training." Within political contexts, these authors demonstrate how this structure can allow for a wide range of hypotheses the classifier can evaluate, including for sentiment analysis: "Does the author of the text use positive language?" named-entity analysis: "Does the statement refer to a specific person?" and topic assignment: "The text is about politics." In this way,

NLIs retain the lightweight-applicability of supervised training models, while retaining the zero-shot capabilities of large GPT models and ability to be applied to wider domains.

For text classification tasks, this means that the process of breaking complex categories into defined hypotheses and expressing them in a manner best interpretable for the model is non-trivial. It requires strong understanding both of the political and theoretical contexts surrounding the selected domain, as well as of the model's training data and limitations. Nevertheless, these authors demonstrated that these models can perform as well or better than GPTs, while remaining magnitudes smaller than their larger model alternatives. From an efficiency standpoint, therefore, NLI models can be as accurate while requiring a fraction of the energy and compute resources.

As such, our research sets about implementing and testing the applicability of such an approach within a highly complex area - that of ascribing a political stance to speakers based on short textual snippets along the dimensions of economic left vs. economic right, pluralist vs. populist rhetoric, and support for liberal democratic values. As we found within model validation, defining a "ground-truth" score for the position of text along these axes is not straightforward - experts surveyed for providing validation produced strongly divergent scores - showing that even traditional non-automated approaches to text labeling are not uncomplicated.

The Political DEBATE Model

Burnham et al. (2024) developed the `Political DEBATE` model as a domain-specialized NLI classifier for political text. Their pipeline began with Laurer et al.'s (Laurer et al. 2022) multilingual checkpoint, `mDeBERTa-v3-base-xnli-multilingual-nli-2mil7`, itself derived from Microsoft's `mDeBERTa-v3-base` transformer and pre-trained on 2,730,000 NLI premise–hypothesis pairs across 26 languages. Burnham and colleagues then fine-tuned this foundation on their own Pol_NLI dataset, consisting of 200,000 premise–hypothesis pairs drawn from social media, news outlets, congressional bills, court-case summaries, and other politically relevant documents. The tasks covered by Pol_NLI in-

clude stance detection, topic classification, hate-speech and toxicity detection, and event extraction. The result is a model available in two sizes—base (86 million parameters) and large (304 million parameters)—capable of zero-shot inference across a wide range of political classification tasks without further fine-tuning. Beyond the team's own validation, Volf and Simko Volf and Simko 2025 find the model to "perform remarkably" comparable large-language models in the specific task of identifying the "politicalness" of a text - that is to say the extent to which a given premise text concerns politics.
The `Political DEBATE` model is open source, offering transparency in ways that are impossible with proprietary models. It also provides, alongside each label, a probability estimate that can be used to gauge the model's confidence in its output. In addition, its specialization for NLI tasks makes it far smaller than comparable proprietary models—86 million and 304 million parameters for the base and large variants, respectively—thereby enabling deployment on consumer-grade hardware.

The `Political DEBATE` model demonstrates strong zero-shot performance. However, Burnham and colleagues have yet to apply it to the specific theoretical dimensions central to comparative politics, such as economic left–right ideology, populist versus pluralist rhetoric, and support for liberal democratic values. Our approach integrates NLI capabilities with established political science theory by operationalizing codebooks from the Global Party Survey, CHES, and V-Party into testable hypotheses. In doing so, we ensure our model remains interoperable with existing broadly accepted datasets. Second, we introduce a deterministic, auditable scoring methodology that aggregates entailment probabilities across these balanced hypothesis sets in a manner that is fully traceable to individual hypotheses. Third, we validate the resulting scores against 147 expert survey responses, demonstrating that the method performs within the expert range on two of the three dimensions—something Burnham et al. (2024) did not attempt. Together, these contributions transform a general-purpose political NLI tool into a theory-grounded measurement instrument for comparative political research.

From Theory to Natural Language Inference

Within our defined text classification task we aim to achieve accuracy levels comparable to those of experts from established academic sources. Accordingly, we ground our model theoretically in codebooks from highly cited and reputable expert surveys, including the *Global Party Survey* (Norris 2020), the *Chapel Hill Expert Survey* (Jolly et al. 2022), and the *Varieties of Party Identity and Organization project* (Lindberg et al. 2022).

The three dimensions for which we score texts are key areas of political competition – typical areas of analytical interest for researchers – and were thus selected to complement existing research approaches. First, the economic left–right dimension represents the most established and consistently documented axis of political conflict in comparative politics. Since Anthony's Downs (1957) *An Economic Theory of Political Action in a Democracy*, the economic left–right divide has been shown to structure voter preferences, party positioning, and government formation across a wide range of democracies (Dalton 2006). Its centrality to political competition makes it an indispensable baseline dimension for any stance identification framework.

Second, the populism–pluralism dimension captures a form of political rhetoric that has become increasingly salient across both established and emerging democracies over the past two decades (Mudde 2017; Norris and Inglehart 2019). Populist rhetoric represents a qualitatively distinct mode of political logic (Laclau 2005) that cuts across the traditional economic left–right axis (Mudde 2004): left-wing and right-wing parties alike have adopted populist frames, meaning that economic ideology alone is insufficient to characterize the full range of contemporary political positioning. Including this dimension, therefore, adds explanatory power to the economic axis.

Third, in the context of growing challenges to liberal democracy from ascendent illiberal powers and with sustained academic interest in democratisation and backsliding, support for liberal democratic values has become an increasingly contested dimension of political competition as a growing number of parties and leaders across different regions have adopted positions that challenge core democratic norms (**lhrmann2019third**; Levitsky and Ziblatt 2018). Measuring this dimension is of particular importance for tracking

processes of democratic backsliding, which have been documented with growing frequency in the political science literature (Mochtak 2025).

Taken together, these three dimensions capture variation along the substantive ideological and rhetorical axes that jointly structure much of contemporary political competition. They also correspond directly to the conceptual frameworks underlying three of the most widely used expert survey instruments in comparative politics, the Chapel Hill Expert Survey, the Global Party Survey, and the Varieties of Party Identity and Organization project, thus lending our operationalization established scholarly authority and enabling future comparisons with existing cross-national datasets.

Recently, populism has become a widely used term in academic literature, yet its definition remains contested. The most widely accepted formulation describes populism as "a thin-centered ideology that considers society to be ultimately separated into two homogeneous and antagonistic groups, 'the pure people' and the 'corrupt elite,' and which argues that politics should be an expression of the volonté générale" (Mudde 2004, p. 543). This definition implies the existence of two core components of populist logic: a people-centric element and an anti-elite element.

We argue that, for the purpose of identifying political stances, it is more appropriate to treat these two components separately. A text may, for instance, invoke the sovereign authority of the people without explicitly attacking an elite, as is common in civic or nationalist rhetoric, while anti-elite appeals can appear without accompanying people-centric language, as in technocratic critiques. Collapsing these into a single dimension would obscure such variation and risk misclassifying texts that exhibit only one feature of the populist ideal type (Mudde 2004). A movement, party, or leader may at times be people-centric without being explicitly anti-elite, and vice versa. However, when populist actors employ a strongly people-centric rhetoric, they are, by definition, anti-pluralist. Accordingly, to identify populist stances, we draw on the question used in the *Global Party Survey* (Norris 2020) to classify parties as employing either populist or pluralist rhetoric:

"Parties can also be classified by their current use of POPULIST OR PLU-

> *RALIST rhetoric. POPULIST language typically challenges the legitimacy of established political institutions and emphasizes that the will of the people should prevail.*
>
> *By contrast, PLURALIST rhetoric rejects these ideas, believing that elected leaders should govern, constrained by minority rights, bargaining and compromise, as well as checks and balances on executive power."* (Norris 2020, p. 6, 7)

The distinction between economic left and economic right ideology is perhaps clearer than that of populism. Nevertheless, there are multiple ways to conceptualize this divide, with some scholars adopting minimalistic approaches and others preferring multidimensional frameworks (Federico 2019). While we acknowledge the value of incorporating additional dimensions, we adopt a more parsimonious approach here. Following Norberto Bobbio (1996)'s understanding of the economic left–right divide, we base our approximation of economic stances on the view that the economic left supports redistribution, state intervention, and egalitarian policies, whereas the economic right opposes redistribution, limits state intervention, and accepts or reinforces social and economic inequalities. Accordingly, to identify economic left/right ideology, we use the following question from the *Chapel Hill Expert Survey*:

> *"Parties and presidents can be classified in terms of their stance on ECONOMIC ISSUES such as privatization, taxes, regulation, government spending, and the welfare state. Parties on the economic left want government to play an active role in the economy. Those on the economic right want a reduced role for government."* (Jolly et al. 2022, p. 22)

Support for liberal democracy is conceived here as support for the core values that constitute this model of governance. Many of these values were articulated by Robert Dahl (2008) in his notion of *polyarchy*. According to Dahl, a polyarchy must include the following requirements: free and fair elections, freedom of expression, freedom to form and join organizations, the right to vote, eligibility for public office, access to alternative sources of information, and institutions that ensure government policies depend on votes

and other expressions of citizen preference (Dahl 2008, p. 3). Accordingly, we draw on the following question from the Varieties of Party Identity and Organization project's codebook, which closely reflects these principles:

> *"Prior to this election, to what extent was the leadership of this political party clearly committed to free and fair elections with multiple parties, freedom of speech, media, assembly and association? Clarification: Party leaders show no commitment to such principles if they openly support an autocratic form of government without elections or freedom of speech, assembly and association (e.g. theocracy; single-party rule; revolutionary regime). Party leaders show a full commitment to key democratic principles if they unambiguously support freedom of speech, media, assembly and association and pledge to accept defeat in free and fair elections."* (Lindberg et al. 2022, p. 27)

We use each of these questions and adapt them to the different Natural Language Identification tasks and approaches. First, we reword the questions to remove references specific to party positioning, since our goal is to identify stances in texts from a variety of sources (e.g. speeches, social media posts, party manifestos). We adapt these definitions to our specific hypotheses-based approach.

For the hypotheses-based approach, we dissect these definitions to create a list of possible hypotheses. For instance, in the case of economic left–right ideology, we include hypotheses like these ones:

```
# Economic Left Positions
    The author of this text believes corporations should pay higher
    taxes: (1.0, left)
    The author of this text believes wealthy individuals should
    pay higher tax rates: (1.0, left)
    The author of this text believes minimum wage laws should
    be strengthened: (1.0, left)
```

```
# Economic Right Positions
    The author of this text believes corporate tax rates should be
    lowered: (1.0, right)
    The author of this text believes income taxes should be reduced:
    (1.0, right)
    The author of this text believes unions hurt economic
    competitiveness: (1.0, right)
```

In total, we have 30 hypotheses for the economic left/right, with 15 for each side. The hypotheses cover a broad range of economic factors traditionally attributed to the sides as described in the definition. The sets are built in a way that we have a corresponding thesis on the opposite side, as can be seen in the example with the thesis addressing taxes. The hypotheses are not exclusively selected in a way that we have one hypothesis for each topic. The hypotheses overlap in the topic they are addressing, since this benefits our scoring methodology as explained in the next section. In the case of economic left policies, we have, for example, "The author of this text believes unemployment benefits should be expanded", and "The author of this text believes social safety nets should be expanded".

For each of our three cases, we built hypothesis sets for the three aspects guided by the definitions of this characteristic.

For the response-based approach, we adapt these hypotheses to present a description and a premise. For instance:

```
# Economic Left Positions
    'description': Economic Left. Wants the government to play an
    active role in the economy.
    'premise': Includes higher taxes, more regulation and government
     spending and a more generous welfare state.
# Center Economic Positions
```

```
        'description': Center. Balances market freedom with government
        intervention for stability and fairness.
        'premise': Supports a mixed economy with both private enterprise
         and public services.
    # Economic Right Positions
        'description': Economic Right. Emphasizes a reduced economic
        role for government.
        'premise': Includes privatization, lower taxes, less regulation,
        less government spending, and a leaner welfare state
```

Hypothesis-Based Scoring Methodology

After establishing the structure of transforming codebooks from political science into hypothesis sets, the following section describes the process of scoring a statement across our policy dimensions economic right vs. economic left, liberal vs. illiberal, and populism vs. pluralism. For each scorer, we use a balanced hypothesis set. A hypothesis acts as a premise for the inference call to the model. The output scores we utilize further are the entailment probability that the DEBATE model gives as output to an input text statement. As a mathematical foundation for combining the scores per hypothesis, we utilize a weighted average. For this, the balanced hypothesis set is important because otherwise one side would gain an unfair advantage in the scoring process. In the score calculation, we compute the weighted average of the entailment scores of the model for each hypothesis of each side of the policy dimension. The overall output score is then shifted from the center point depending on which hypothesis set the NLI model has in total a stronger entailment for. That also has the side effect of statements that have mixed signals, triggering some hypotheses on each side, being kept more centered, and accompanying nuances in natural language. This scoring methodology treats all hypotheses as equally weighted. This is a deliberate choice intended to avoid imposing

researcher-defined hierarchies on the relative importance of different ideological cues. Equal weighting is, of course, itself a methodological assumption. However, it requires fewer substantive judgments than differential weighting, which would demand a principled justification for assigning distinct values to particular cues. Providing such a justification is difficult without reintroducing the theoretical subjectivity that this approach seeks to minimize (Grimmer and Stewart 2013).

The rationale for this approach is twofold. First, keeping the evaluation in the textual domain may reduce the need for the model to map semantic judgments onto an externally imposed numerical scale, allowing the DEBATE model to utilize its large latent space rather than funneling cues through a numerical bottleneck before producing a human-readable score. Second, it allows the scoring procedure to capture the fact that a single statement can simultaneously support multiple hypotheses.

General Framework

In the following the general framework behind each scorer is described based on the example of the economic left vs economic right scorer. For a given political text T and a set of hypothesis statements $H = \{h_1, h_2, \ldots, h_n\}$, we first obtain entailment probabilities using a Natural Language Inference (NLI) model:

$$p_i = P(\text{entailment}|T, h_i), \quad i = 1, 2, \ldots, n \tag{1}$$

Each hypothesis h_i is associated with a weight $w_i = 1$ and a directional label $d_i \in \{\text{economic left}, \text{economic right}\}$ (or corresponding labels for other dimensions).

Scoring

Weighted Average Calculation

The weighted probabilities for each political direction are calculated as:

$$P_{\text{economic left}} = \{p_i \cdot w_i : d_i = \text{economic left}\} \tag{2}$$

$$P_{\text{economic right}} = \{p_i \cdot w_i : d_i = \text{economic right}\} \tag{3}$$

The directional averages are computed as:

$$\overline{p}_{\text{economic left}} = \frac{1}{|P_{\text{economic left}}|} \sum_{p \in P_{\text{economic left}}} p \tag{4}$$

$$\overline{p}_{\text{economic right}} = \frac{1}{|P_{\text{economic right}}|} \sum_{p \in P_{\text{economic right}}} p \tag{5}$$

Final Score Calculation

The final economic left–right score is calculated using the difference between averages:

$$\text{difference} = \overline{p}_{\text{economic left}} - \overline{p}_{\text{economic right}} \tag{6}$$

$$S_{\text{LR}} = \text{clip}(5 - (d \times 5),\ 0,\ 10) \tag{7}$$

where d is the difference score and $\text{clip}(x, a, b) = \max(a, \min(x, b))$ constrains the result to $[0, 10]$.

The general scoring process, as shown above, which calculates overall scores straightforwardly given the entailment for each of the hypotheses. The framework has two benefits. On the one hand, it is really transparent because each calculated score for a text is given to it in a deterministic way. The assigned overall score can be understood all the way back to the score per hypothesis in the respective hypothesis set. The hypothesis set can be easily extended to cover aspects that might have been missed in the currently used ones or adapted to other contexts besides the three studied ones here.

Validation strategy

To validate the scores obtained by the two scoring approaches described in the previous section, we relied on evaluations from social science experts who scored a sample of texts. This section describes the text sample used for validation, the expert recruitment process, and the survey administered to obtain their assessments.

Although the social science literature has extensively examined rhetoric related to populism, economic left–right ideology, and support for liberal democratic values, only a few studies publish the datasets of texts they analyze. Moreover, the theoretical foundations used to classify such texts often vary across studies and sometimes differ from our own conceptualization of these topics. For this reason, we collected texts from multiple sources and manually reviewed each one to ensure that they aligned as closely as possible with the categories assigned. Still, we treat these as weak labels, since in many cases the categorization criteria were not fully explained by the original authors or were inferred indirectly—for example, by using the political party expressing the text as a proxy. After completing the collection process, we drew a representative sample of 20 texts for each topic. Table A1 in the appendix provides detailed information on the specific sources included in our dataset.

We recruit experts for each of the three dimensions. To do so, we identified authors who had published articles on these dimensions in leading political science journals—*Political Communication* and *Comparative Political Studies*—within the past five years. In the case of populism, we additionally included scholars affiliated with the Team Populism group, given their prominence in this area. Our final list comprised 198 authors for the populism texts, 117 authors for the economic right/left ideology texts, and 120 authors for the texts related to support for liberal democracy.

We contacted these authors and invited them to complete a brief survey. Each author was presented with a random selection of texts and asked to adjust a slider to indicate the score that best reflected the content of a given text based on the definition provided for the corresponding dimension. In total, we received 147 responses: 71 for the populist/pluralist dimension, 37 for economic right/left ideology, and 38 for support for

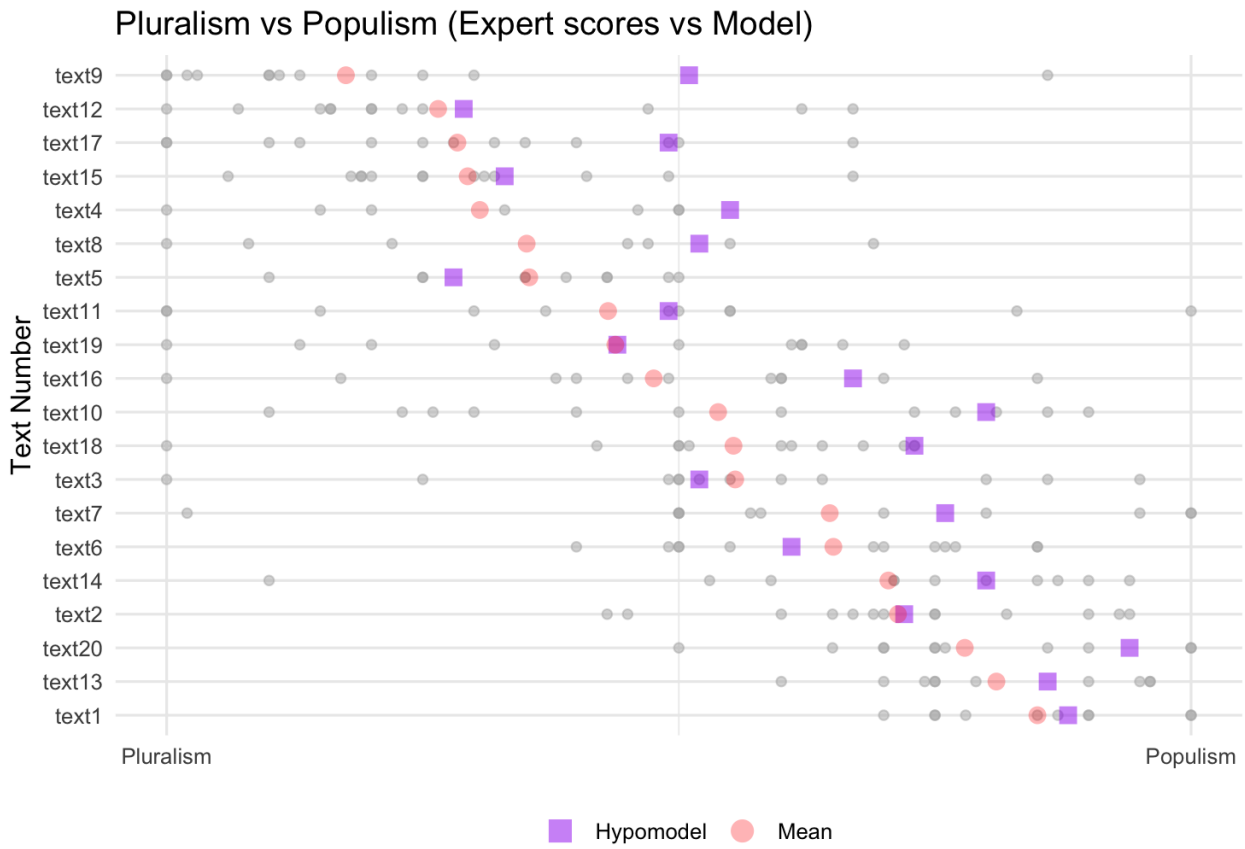

Figure 1: Pluralism/Populism

liberal democracy. With the input of these experts, we assess the performance of our approach. The results of this assessment are presented in the next section.

Results

In this section, we assess the performance of our approach across the three tasks of interest. Our evaluation compares expert-provided survey scores with the scores generated by our model. Overall, the results show that the model performs within the range of expert evaluations, with a few outlier cases that may require targeted adjustments. This applies primarily to the tasks of populism/pluralism and economic left–right ideology. For the task measuring support for liberal democracy, however, the model scores deviate substantially from expert judgments, indicating the need for a major reconsideration of the modeling approach.

We begin by presenting a visual representation of the scores assigned by each model

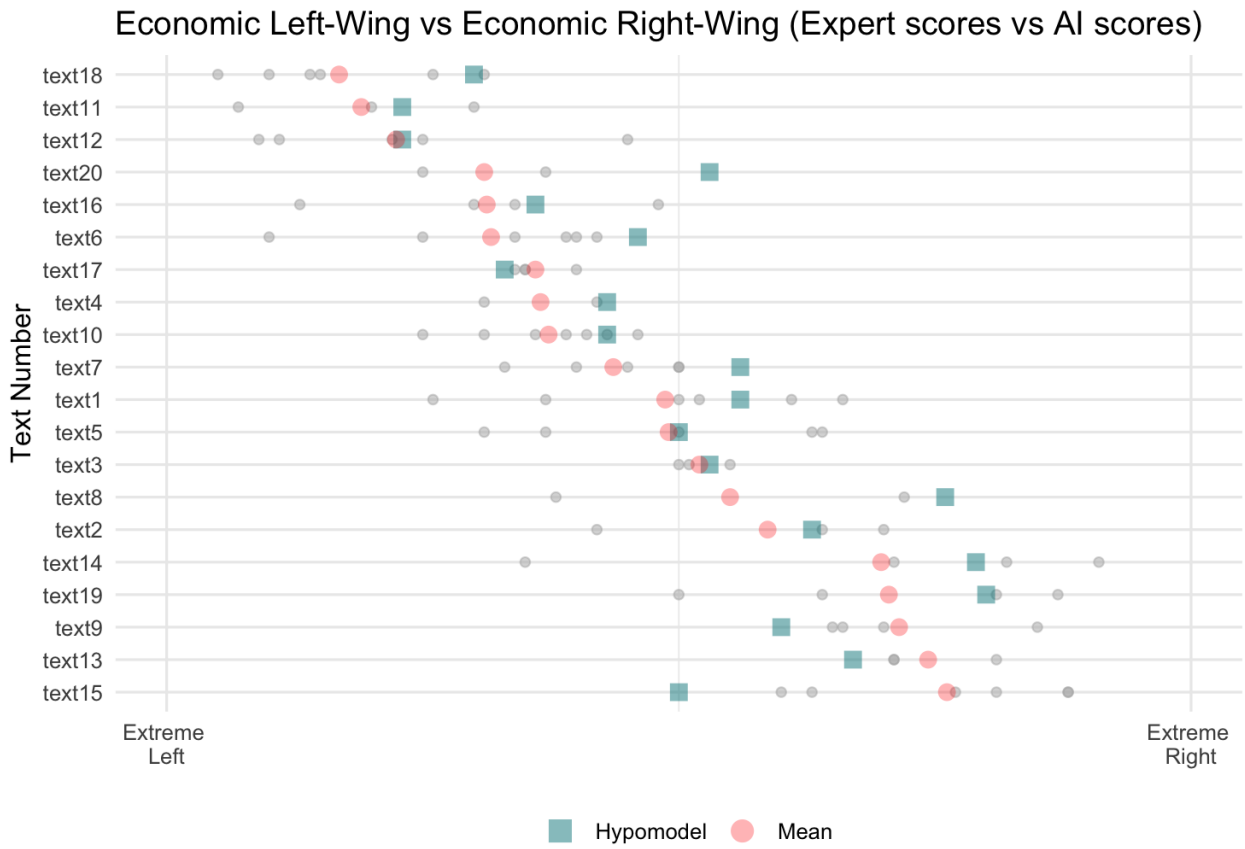

Figure 2: Economic Left/Right

in comparison with those provided by experts. Figure 1 displays this comparison for the pluralism–populism rhetoric scale. As shown, the models' scores generally fall within the range of expert evaluations. With many scores falling close to the mean such as in texts 19, 2, 12, or 1. Figure 2 presents the comparison for the economic left–right texts. Here some scores such as in text 20, 7, and 15 lie outside the experts' judgments, with a range of expert evaluations somewhat narrower. Yet, the model aligns closely with the experts' assessments, with again some scores falling very close to the mean, such as texts 12, 17, 15, 3, and 2. Finally, Figure 3 shows the comparison for texts expressing support for—or opposition to—liberal democratic values. Here, we observe a greater number of scores that fall well outside the expert range. Yet, some cases such as texts 4, 1, 15, and 14 are close to the mean.

We further analyze our model scores by providing correlation and error metrics. Table 1 shows the Pearson correlation coefficient and the Spearman's rank correlation coefficient

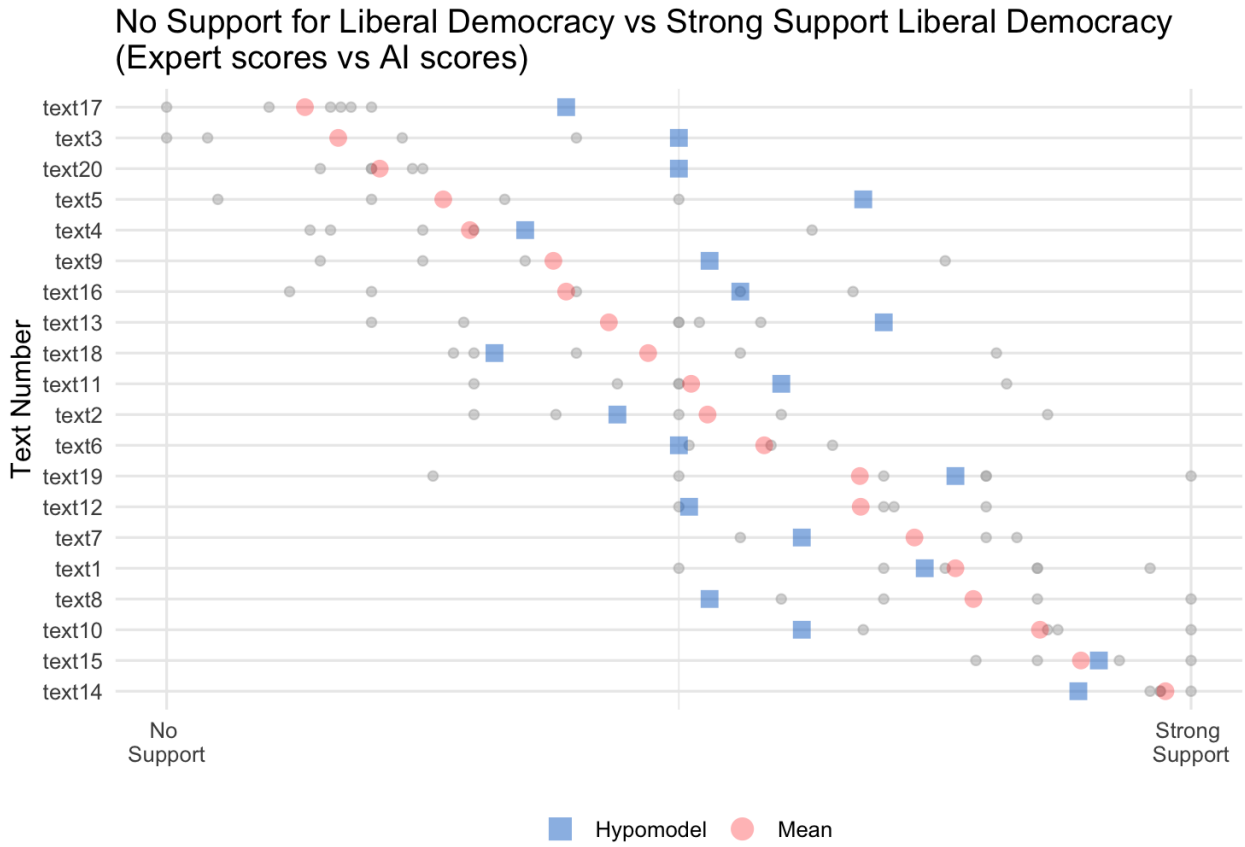

Figure 3: Support for Liberal Democratic Values

between the model predictions and expert scores. The correlation coefficients show that the model provides moderately strong correlation with expert judgments across tasks, although the strength varies by task. For Populism/Pluralism, the model displays relatively high correlations (Pearson = 0.84; Spearman = 0.81), indicating strong agreement with expert scores in both linear association and rank ordering. A similar pattern is observed for the Economic Left/Right dimension, where correlations remain comparatively strong (Pearson = 0.81; Spearman = 0.79), suggesting consistent predictive performance. For Support for Liberal Democracy, correlations are lower overall (Pearson = 0.62; Spearman = 0.58), indicating weaker alignment with expert assessments on this dimension. Overall, performance is strongest for Populism/Pluralism and Economic Left/Right Ideology, whereas Support for Liberal Democracy appears to be more challenging for the model.

Regarding error metrics, we provide measurements for the mean absolute error (MAE) and the root mean square error (RMSE). Table 2 shows these metrics across the three

Model / Text Type	Pearson correlation coefficient	Spearman's rank correlation coefficient
Populism/Pluralism Rhetoric	0.84	0.81
Economic Left/Right Ideology	0.81	0.79
Support for Liberal Democracy	0.62	0.58

Table 1: Correlation performance (Pearson and Spearman) for hypothesis-based approach across tasks.

Text Type	Mean absolute error	Root mean square error
Populism/Pluralism Rhetoric	11.57	14.92
Economic Left/Right Ideology	9.24	11.62
Support for Liberal Democracy	16.68	19.62

Table 2: Error metrics (MAE and RMSE) for hypothesis-based approach across tasks.

tasks. The model achieves relatively low prediction error for Populism/Pluralism (MAE = 11.57; RMSE = 14.92), indicating a close correspondence with expert scores and limited large deviations. This pattern is also observed for the Economic Left/Right Ideology scores, where the approach shows even lower error overall (MAE = 9.24; RMSE = 11.62), indicating comparatively strong accuracy. In contrast, performance is weaker for the Support for Liberal Democracy task, where both MAE (16.68) and RMSE (19.62) are higher, suggesting greater dispersion between predictions and expert evaluations. Overall, these results show that the hypothesis-based method shows a strong performance on the Populism/Pluralism rhetoric and Economic Left/Right Ideology dimensions, while Support for Liberal Democracy remains the most challenging task.

While these results are promising, the nature of our model offers an unusual opportunity to address some of the issues encountered. Accordingly, we pursue two mitigation strategies. First, for the Populism/Pluralism and Economic Left/Right tasks, we plan to refine the underlying hypotheses to better reflect how the model interprets and evaluates the stances expressed in the texts. Second, for the Liberal/Illiberal dimension, we will obtain additional expert responses to achieve a clearer understanding of how experts

assess these texts and, in turn, produce a more reliable benchmark for evaluation.

Qualitative Bias Analysis

To identify potential sources of bias within the scores obtained, we conducted a qualitative analysis of the underlying `Political DEBATE` model. The analysis employs template-based bias testing, a method in which identical statements are tested, with the only difference being the addressed group or the political figure to whom the statement is attributed. The statements are spread across the whole spectrum of the score. After the full statement set is scored, the average deviation per group or per political figure is computed.

The results for the political figures are presented in Figure 4. It is worth noting that the economic left–right dimension remains largely agnostic to who the statements are attributed to. However, the populist–pluralist rhetoric dimension seems very sensitive in this test. Statements attributed to Donald Trump or Jair Bolsonaro shift the score more than one point towards a populist stance. The bias is also considerable for right-wing leaders such as Narendra Modi, but is likewise observable for the economic left-wing U.S. politician Bernie Sanders. Support for liberal democracy also reveals several notable patterns of bias. In a somewhat counterintuitive finding, leaders frequently characterized as illiberal, such as Narendra Modi and Viktor Orbán (Smilova 2025), display levels of bias in favor of liberal democracy comparable to those of leaders commonly associated with pro-liberal values, such as Angela Merkel (Kelemen 2017). Beyond these cases, the overall bias in scores across the analyzed leaders remains modest, averaging approximately 0.5 points.

The results for the demographic groups are more nuanced, as the interpretation of statements changes depending on the addressed audience. The following example illustrates this effect. The baseline statement, "The author believes businesses drive economic growth," receives a score of 8.666 from the economic model, corresponding to an economic right position. When the statement is modified to "The author believes working-class businesses drive economic growth," the score decreases to 3.998, representing a shift of

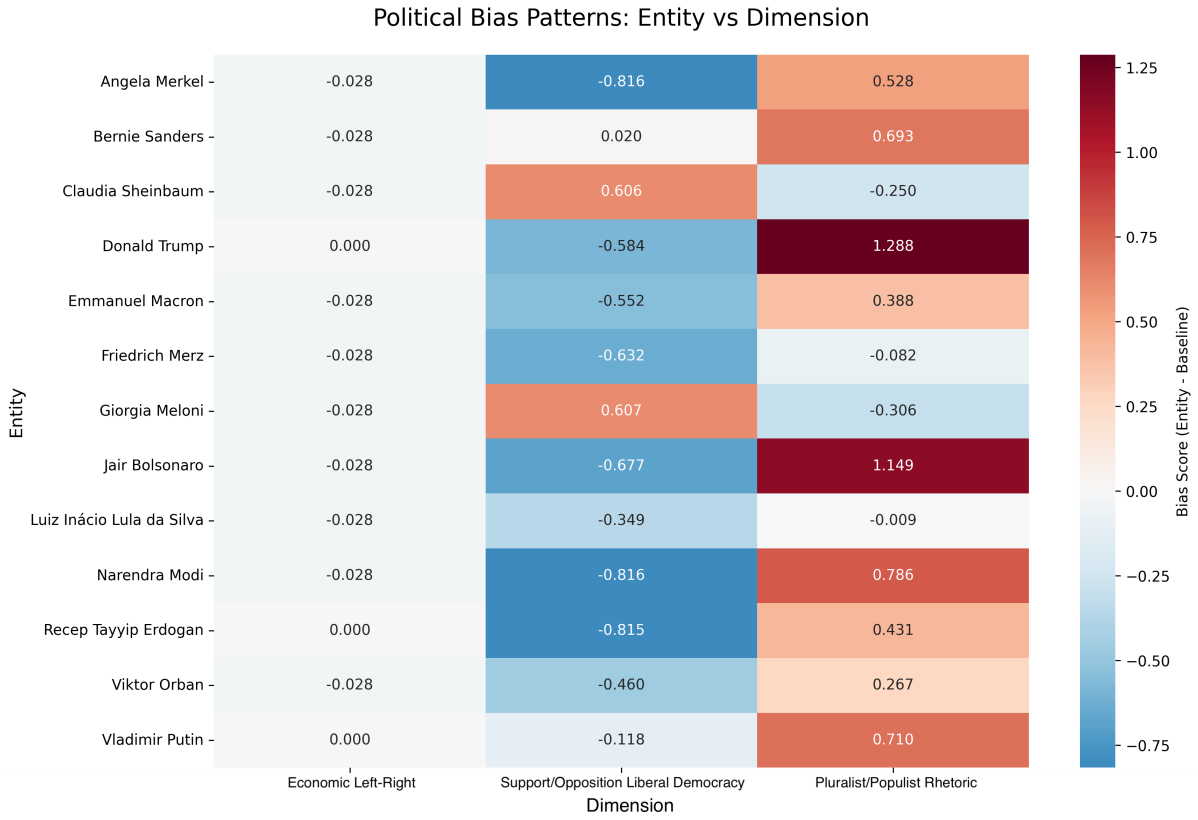

Figure 4: Template-based bias testing for political figures across political dimensions.

−4.669 toward the economic left. This shift is consistent with common definitions of economic left–right positions.

We carry out a systematic assessment by compiling an additional set of statements spanning the full range of scores. The results of the template-based bias analysis are shown in Figure 5. While the overall average score deviations across demographic groups remain below one point, these results reveal substantial variation for specific groups, indicating context-dependent sensitivity of the model. For instance, the economic left–right dimension shows moderate sensitivity to group attribution, with noticeable shifts for groups such as Mexican, Muslim, and immigrant, while remaining relatively stable for others. The populist–pluralist rhetoric dimension appears particularly sensitive in this test, with bias above one point toward a pluralist stance for groups such as Buddhists and Muslims, and close to one for Jewish. In contrast, the support for liberal democracy dimension exhibits more moderate and mixed effects, with the largest shifts observed for groups such as the Iranian and Russian on one hand and the wealthy on the other side,

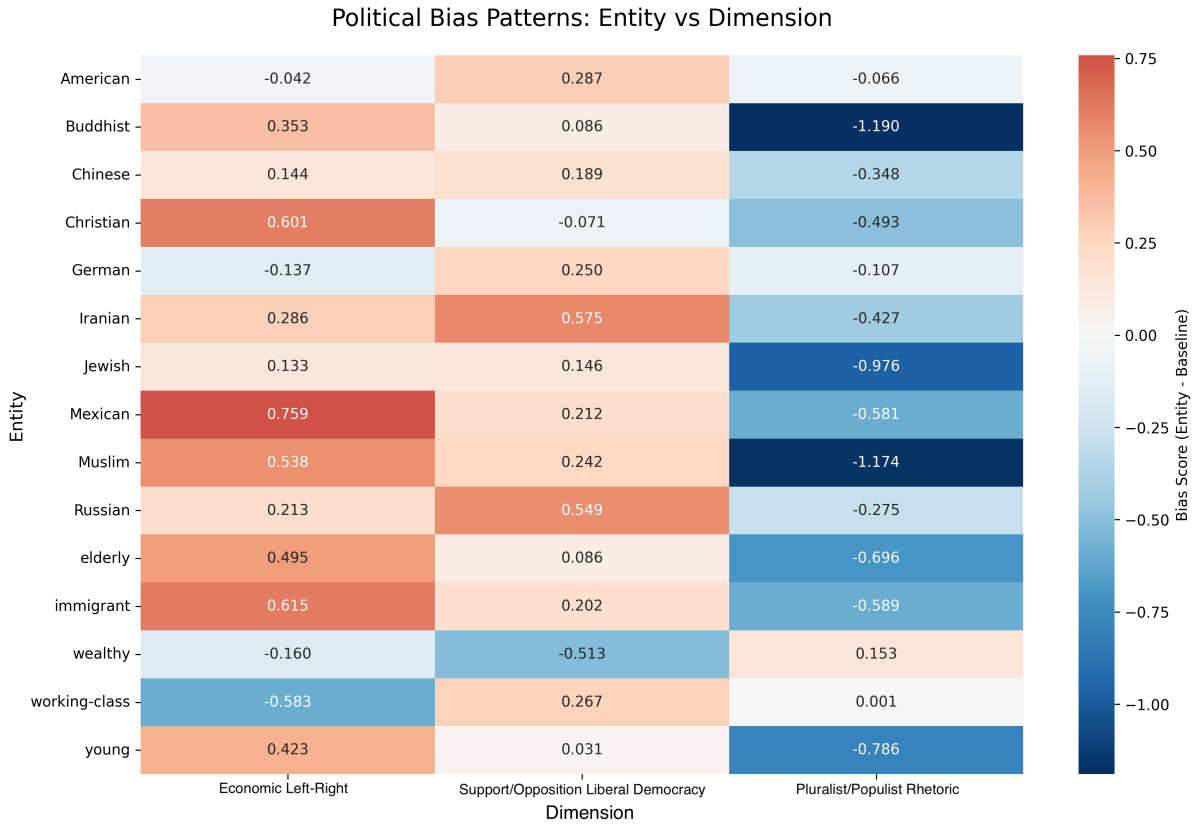

Figure 5: Template-based bias Testing for demographic groups

while most other groups show only limited deviations.

Discussion and Limitations

In this article, we introduce a transparent and reproducible methodology for automating political stance detection using a hypothesis-based Natural Language Inference (NLI) approach built on the `Political DEBATE` model. We validate this method using expert evaluations. As described above, our empirical results show that the model identifies stances within expert-coded ranges for two of the three dimensions—populism and economic left–right rhetoric—demonstrating its suitability for a range of political stance detection tasks. However, performance on the support for liberal democracy dimension shows lower agreement with expert assessments, indicating that computational, lexically driven approaches to political inference may fall short in more nuanced cases. In this section, we interpret these findings by discussing the model's strengths and limitations,

outlining potential directions for future work, and reflecting on ethical considerations, thereby providing a broader perspective on the implications of natural language processing for political stance detection.

We assess the hypothesis-driven approach and propose two broad explanations for why the NLI analysis performs well on the populist and economic left–right dimensions, but struggles with liberal–illiberal speech. First, constructs that are relatively lexicalized and cue-driven in political texts, such as populist appeals to 'the people' versus elites and many economic left–right signals (e.g., taxation, welfare, or market regulation), lend themselves to decomposition into short, discrete hypotheses. Prior work on political text analysis has shown that ideological signals in these dimensions tend to cluster around recurring lexical markers and policy domains (Laver and Garry 2000; Grimmer and Stewart 2013), which makes entailment-based classification a tractable approach: when a text contains such cues, the model can reliably detect entailment for the corresponding hypothesis. The NLI model can detect entailment for such statements with high reliability, and the balanced hypothesis aggregation we employ preserves nuance when texts trigger hypotheses on both sides. Second, support for liberal democratic values appears to be a more diffuse and context-dependent construct in our sample. Support for liberal democratic values, by contrast, is frequently expressed through indirect or procedural language — endorsements of institutional constraints, conditional acceptance of electoral outcomes, or appeals to constitutional norms — rather than through the explicit lexical markers that characterize populist or economic ideological speech (Norris and Inglehart 2019; Levitsky and Ziblatt 2018). This discursive indirection means that short hypothesis templates are less likely to trigger clear entailment signals, as the relevant commitments are embedded in broader rhetorical and institutional contexts that exceed what a single premise–hypothesis pair can capture. This helps explain the larger deviations from expert judgments on this dimension and suggests that operationalizing liberal-democratic commitment may require richer hypothesis sets, alternative prompt designs, or modest amounts of labeled data (e.g., a few-shot approach) to provide additional model context.

Despite the shortcomings of the liberal democracy dimension, the method presented

here already offers several practical advantages, including its potential for improvement due to its transparency. Furthermore, it reduces the need for large labeled datasets and extensive fine-tuning by leveraging a lightweight NLI-trained model. This makes it feasible to deploy stance measurement on local hardware rather than relying on access tokens for proprietary models, and thus enables affordable scaling to larger corpora. The hypothesis-aggregation procedure is also deterministic and auditable: each final score can be decomposed into the hypotheses that contributed to it and the model's confidence in each. The method's initial multilingual performance (English, German, Portuguese, and Spanish) shows promise for cross-lingual applications (see Appendix 1: Multilingual Capabilities Test Results). Finally, based on the results obtained here, the approach will be improved and used to develop a publicly accessible platform, the Automated Political Stance Identification (APSI) tool, which will provide an interface for non-technical users to analyze political text datasets at scale.

Although promising, the method presented here has several limitations. First, our validation relied on a small purposive sample (60 texts) and an uneven distribution of expert responses, which limits statistical power and external validity. The `Political DEBATE` model's training on predominantly English political NLI data also raises concerns about domain and language transfer when applied to other languages and cultural registers. Moreover, the method currently analyzes relatively short texts (fewer than 20,000 characters); longer documents or manifesto sections, therefore, require careful segmentation. Methodological choices in scoring—such as transforming entailment probabilities into a 0–10 scale and assigning equal weights to hypotheses—are pragmatic but heuristic and may introduce calibration biases that affect substantive inferences. Finally, political language is dynamic and highly contextual, and the method should therefore be applied with caution. The resulting labels do not represent a definitive ground truth but rather theory-driven estimates that require human interpretation and oversight.

Future research should follow two distinct but intersecting strands. First, a methodological research avenue could focus on improving the approach itself. This may include refining score calibration, strengthening multilingual capabilities, and enabling the anal-

ysis of longer documents. In addition, the framework could be extended to capture additional political dimensions by incorporating new sets of hypotheses. Second, a substantive research agenda could explore how the method can be applied to address diverse questions related to political discourse. For example, drawing on German political party manifestos (see Appendix 2), we provide an illustration of how the tool can be used to analyze ideological drift over time. Additional research avenues include examining differences in rhetorical strategies during election campaigns versus periods in office, comparing social media communication with institutional or parliamentary discourse, and analyzing cross-national variation in ideological positioning, populist rhetoric, or support for liberal democratic norms, among others.

1 Concluding Remarks

This study contributes to a growing body of scholarship advocating for the transparent and accountable use of artificial intelligence in social science research. The validation results presented here—demonstrating expert-range performance on two of the three evaluated dimensions using a fully deterministic and hypothesis-traceable scoring procedure—show that the application of machine learning models to political text analysis does not inherently require a trade-off between transparency and accuracy. Whereas many contemporary approaches rely on generative language models whose outputs may vary across prompts and whose internal reasoning processes remain difficult to audit (Lipton 2016), the method developed here anchors every classification in an explicit and theoretically derived set of hypotheses. Because each hypothesis can be independently examined, modified, or expanded, the scoring procedure remains both reproducible and open to systematic refinement.

More broadly, the approach illustrates that computational analysis of political discourse can remain closely connected to substantive political theory. Rather than relying on opaque and non-deterministic model architectures, the framework operationalizes established conceptual definitions drawn from the comparative politics literature and

translates them into structured Natural Language Inference tasks. This design not only improves interpretability but also substantially lowers the resource requirements typically associated with machine-learning approaches. In place of large annotated datasets and extensive model fine-tuning, analytical effort can instead be directed toward the construction and refinement of theoretically informed hypothesis sets. As a result, the framework offers a comparatively lightweight yet theoretically grounded method for analyzing political texts along the ideological and rhetorical dimensions examined in this study.

The implications of this approach extend beyond academic research. Implemented in the web-based APSI tool, the method can enable journalists, civil society organizations, think tanks, and other public actors to examine political discourse in a systematic and transparent manner. By providing automated yet interpretable measurements of ideological positioning, populist rhetoric, and support for liberal-democratic norms, such tools could facilitate more rigorous scrutiny of political communication. In doing so, they may contribute to a broader ecosystem of democratic accountability by helping users identify rhetorical strategies, track ideological positioning across actors and time, and engage more critically with political narratives. Rather than replacing expert analysis, this tool is intended to prompt users to reflect critically on political communication – supporting the expansion of media literacy at a time of increasing disruption to the political status quo through networked communications.

References

Abercrombie, Gavin and Riza Batista-Navarro (2022). "Policy-focused Stance Detection in Parliamentary Debate Speeches". In: *Northern European Journal of Language Technology* 8. Ed. by Leon Derczynski. DOI: 10.3384/nejlt.2000-1533.2022.3454. URL: https://aclanthology.org/2022.nejlt-1.5/.

ALDayel, Abeer and Walid Magdy (2021). "Stance detection on social media: State of the art and trends". In: *Information Processing & Management* 58.4, p. 102597. ISSN:

0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102597. URL: https://www.sciencedirect.com/science/article/pii/S0306457321000960.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv: 1409.0473 [cs.CL]. URL: https://arxiv.org/abs/1409.0473.

Bestvater, Samuel E. and Burt L. Monroe (2023). "Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis". In: *Political Analysis* 31.2, pp. 235–256. DOI: 10.1017/pan.2022.10.

Bobbio, Norberto (1996). *Left and right: The significance of a political distinction*. University of Chicago Press.

Bosley, Mitchell et al. (2023). "Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research". In: *2023 Annual Meeting of the Midwest Political Science Association (MPSA)*.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

Burnham, Michael (July 2025). "Stance detection: a practical guide to classifying political beliefs in text". en. In: *Political Science Research and Methods* 13.3, pp. 611–628. ISSN: 2049-8470, 2049-8489. DOI: 10.1017/psrm.2024.35. URL: https://www.cambridge.org/core/journals/political-science-research-and-methods/article/stance-detection-a-practical-guide-to-classifying-political-beliefs-in-text/E227E746BD7D9751526DA0EC2C378787 (visited on 08/25/2025).

Burnham, Michael et al. (Sept. 2024). *Political DEBATE: Efficient Zero-shot and Few-shot Classifiers for Political Text*. arXiv:2409.02078 [cs]. DOI: 10.48550/arXiv.2409.02078. URL: http://arxiv.org/abs/2409.02078 (visited on 08/25/2025).

Chen, Jinghui, Takayuki Mizuno, and Shohei Doi (2024). "Analyzing political party positions through multi-language twitter text embeddings". In: *Frontiers in Big Data* 7, p. 1330392.

Cole, Deborah and Philip Oltermann (Nov. 11, 2025). "ChatGPT violated copyright law by 'learning' from song lyrics, German court rules". In: *The Guardian*. ISSN: 0261-3077. URL: https://www.theguardian.com/technology/2025/nov/11/chatgpt-violated-copyright-laws-german-court-rules (visited on 12/02/2025).

Dahl, Robert A (2008). *Polyarchy: Participation and opposition*. Yale university press.

Dalton, Russell J. (2006). *Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies*. 4th. Washington, D.C.: CQ Press.

Darwish, Kareem, Walid Magdy, and Tahar Zanouda (2017). "Improved Stance Prediction in a User Similarity Feature Space". In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ASONAM '17. Sydney, Australia: Association for Computing Machinery, pp. 145–148. ISBN: 9781450349932. DOI: 10.1145/3110025.3110112. URL: https://doi.org/10.1145/3110025.3110112.

Darwish, Kareem, Peter Stefanov, et al. (May 26, 2020). "Unsupervised User Stance Detection on Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14, pp. 141–152. ISSN: 2334-0770. DOI: 10.1609/icwsm.v14i1.7286. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7286 (visited on 11/26/2025).

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

Ding, Ning et al. (Mar. 1, 2023). "Parameter-efficient fine-tuning of large-scale pre-trained language models". In: *Nature Machine Intelligence* 5.3, pp. 220–235. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00626-4. URL: https://doi.org/10.1038/s42256-023-00626-4.

Doshi-Velez, Finale and Been Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv: 1702.08608 [stat.ML]. URL: https://arxiv.org/abs/1702.08608.

Downs, Anthony (1957). "An economic theory of political action in a democracy". In: *Journal of political economy* 65.2, pp. 135–150.

Federico, Christopher M (2019). "Ideology and public opinion". In: *New directions in public opinion*. Routledge, pp. 75–98.

Ghanem, Bilal, Paolo Rosso, and Francisco Rangel (Nov. 2018). "Stance Detection in Fake News A Combined Feature Representation". In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Ed. by James Thorne et al. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71. DOI: 10.18653/v1/W18-5510. URL: https://aclanthology.org/W18-5510/.

Grimmer, Justin and Brandon M Stewart (2013). "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". In: *Political analysis* 21.3, pp. 267–297.

Grynbaum, Michael M. and Ryan Mac (Dec. 27, 2023). "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work". In: *The New York Times*. ISSN: 0362-4331. URL: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (visited on 12/02/2025).

Hou, Yuxin and Junming Huang (2025). "Natural language processing for social science research: A comprehensive review". In: *Chinese Journal of Sociology* 11.1, pp. 121–157.

Houlsby, Neil et al. (2019). "Parameter-Efficient Transfer Learning for NLP". In: *CoRR* abs/1902.00751. arXiv: 1902.00751. URL: http://arxiv.org/abs/1902.00751.

Howard, Jeremy and Sebastian Ruder (2018). "Fine-tuned Language Models for Text Classification". In: *CoRR* abs/1801.06146. arXiv: 1801.06146. URL: http://arxiv.org/abs/1801.06146.

Hsieh, Cheng-Yu et al. (2023). "Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes". In: *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017.

Jolly, Seth et al. (2022). "Chapel Hill expert survey trend file, 1999–2019". In: *Electoral studies* 75, p. 102420.

Kelemen, R Daniel (2017). "Europe's other democratic deficit: National authoritarianism in Europe's Democratic Union". In: *Government and opposition* 52.2, pp. 211–238.

Küçük, Dilek and Fazli Can (Feb. 2020). "Stance Detection: A Survey". In: *ACM Comput. Surv.* 53.1. ISSN: 0360-0300. DOI: 10.1145/3369026. URL: https://doi.org/10.1145/3369026.

Laclau, Ernesto (2005). "Populism: What's in a Name". In: *Populism and the Mirror of Democracy*, pp. 103–114.

Lai, Mirko et al. (2017). "Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets". In: *Advances in Computational Intelligence*. Ed. by Grigori Sidorov and Oscar Herrera-Alcántara. Cham: Springer International Publishing, pp. 155–168. ISBN: 978-3-319-62434-1.

Laurer, Moritz et al. (June 2022). "Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI". en-us. In: *Preprint*. Publisher: Open Science Framework. URL: https://osf.io/74b8k (visited on 07/28/2022).

Laver, Michael and John Garry (2000). "Estimating policy positions from political texts". In: *American Journal of Political Science*, pp. 619–634.

Lehman, Eric et al. (2023). "Do we still need clinical language models?" In: *Conference on health, inference, and learning*. PMLR, pp. 578–597.

Levitsky, Steven and Daniel Ziblatt (2018). *How Democracies Die*. New York: Crown.

Lillie, Anders Edelbo and Emil Refsgaard Middelboe (2019). *Fake News Detection using Stance Classification: A Survey*. arXiv: 1907.00181 [cs.CL]. URL: https://arxiv.org/abs/1907.00181.

Lindberg, Staffan I et al. (2022). "Codebook varieties of party identity and organization (V–party) v2". In.

Lipton, Zachary Chase (2016). "The Mythos of Model Interpretability". In: *CoRR* abs/1606.03490. arXiv: 1606.03490. URL: http://arxiv.org/abs/1606.03490.

Liu, Menglin and Ge Shi (2025). *Enhancing LLM-Based Text Classification in Political Science: Automatic Prompt Optimization and Dynamic Exemplar Selection for Few-*

Shot Learning. arXiv: 2409.01466 [cs.CL]. URL: https://arxiv.org/abs/2409.01466.

Marwala, Tshilidzi (2023). "Natural language processing in politics". In: *Artificial intelligence, game theory and mechanism design in politics*. Springer, pp. 99–115.

Mochtak, Michal (2025). "Chasing the authoritarian spectre: Detecting authoritarian discourse with large language models". en. In: *European Journal of Political Research* 64.3. _eprint: https://ejpr.onlinelibrary.wiley.com/doi/pdf/10.1111/1475-6765.12740, pp. 1304–1325. ISSN: 1475-6765. DOI: 10.1111/1475-6765.12740. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6765.12740 (visited on 08/25/2025).

Mudde, Cas (2004). "The populist zeitgeist". In: *Government and opposition* 39.4, pp. 541–563.

— (2017). "An ideational approach". In: *The Oxford handbook of populism*, pp. 27–47.

Nikolaev, Dmitry, Tanise Ceron, and Sebastian Padó (2023). "Multilingual estimation of political-party positioning: From label aggregation to long-input Transformers". In: *arXiv preprint arXiv:2310.12575*.

Norris, Pippa (2020). "Measuring populism worldwide". In: *Party politics* 26.6, pp. 697–717.

Norris, Pippa and Ronald Inglehart (2019). *Cultural backlash: Trump, Brexit, and authoritarian populism*. cambridge university press.

Rohit, Sakala Venkata Krishna and Navjyoti Singh (2018). "Analysis of Speeches in Indian Parliamentary Debates". In: *CoRR* abs/1808.06834. arXiv: 1808.06834. URL: http://arxiv.org/abs/1808.06834.

Smilova, Ruzha (2025). "Conceptual space for illiberal democracy". In: *Politics and Governance* 13.

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

Volf, Matous and Jakub Simko (2025). "Political Leaning and Politicalness Classification of Texts". In: Version Number: 1. DOI: 10.48550/ARXIV.2507.13913. URL: https://arxiv.org/abs/2507.13913 (visited on 03/11/2026).

Wang, Yu, Wen Qu, and Xin Ye (2024). *Selecting Between BERT and GPT for Text Classification in Political Science Research*. arXiv: 2411.05050 [cs.CL]. URL: https://arxiv.org/abs/2411.05050.

ZHAO, XUJIANG et al. (2023). "Beyond one-model-fits-all: A survey of domain specialization for large language models". In: *arXiv preprint arXiv* 2305.

Appendix 1. Multilingual Capabilities Test Results

Table 3: Scores across languages with differences relative to English in parentheses

Text	**English**	**German**	**Portuguese**	**Spanish**
At Didier Reynders' instigation, the increase in the daily ceiling and age limit for tax deductibility of childcare costs (€11.20/day up to the age of 12) provides relief for many families faced with the need to have their children looked after during school vacations, and guarantees all children the same opportunities to take part in fulfilling activities.	3.3	3.3 (=)	3.1 (−0.2)	2.8 (−0.5)
To this end, it is necessary to allow greater private investment in electricity generation and promote the opening of the market to generate competition among actors, under the premise that what is important is not who produces the energy, but who does it at a lower cost and with better quality, so that the price paid by consumers of all sizes is reduced.	6.7	7.0 (+0.3)	7.7 (+1.0)	7.3 (+0.6)

Text	English	German	Portuguese	Spanish
The resources from the collection of the debt of the municipalities and companies with the former Insfopal, carried out by Findeter according to Law 57 of 1989, will be used to finance projects of the Business Modernization Program, executed by the Ministry of Economic Development, for which Findeter is authorized to incorporate such resources in its budget for that purpose.	5.0	5.3 (+0.3)	5.3 (+0.3)	5.0 (=)
We will use the Israeli experience to monitor the activities of so-called political NGOs, their funding and their management from abroad.	3.5	3.8 (+0.3)	5.0 (+1.5)	4.7 (+1.2)
Judges in Swedish courts are appointed directly by the government, and a weak constitutional review process has contributed to a limited separation of powers compared to other liberal democracies.	5.0	5.3 (+0.3)	6.0 (+1.0)	5.3 (+0.3)

Text	English	German	Portuguese	Spanish
Some rules are nonnegotiable. The free democratic principles apply to everyone, regardless of origin or length of stay.	9.1	8.2 (−0.9)	9.2 (+0.1)	8.5 (−0.6)
We are listening to brothers and sisters in the African-American community... Latino... Native-American... young people.	2.9	3.7 (+0.8)	3.1 (+0.2)	3.6 (+0.7)
In our program we propose formulas to improve the model of territorial organization. We want all Spaniards to enjoy the same rights, wherever they live.	4.4	5.0 (+0.6)	4.8 (+0.4)	4.5 (+0.1)
Those who managed to reach the top in recent years haven't been the best for a long time but rather the most greedy, most corrupt, and most shameless.	8.8	8.6 (−0.2)	8.2 (−0.6)	8.0 (−0.8)

Appendix 2. Ideological Drift

To examine ideological drift, we conducted a quantitative analysis of party manifestos prepared for the last federal elections. We used simple Python scripts to extract the text in the manifestos. This extracted text was clustered with the token limit of the model and the original paragraphs of the manifesto in mind. The following plots show the quantitative distribution of statements across each category for the investigated dimensions.

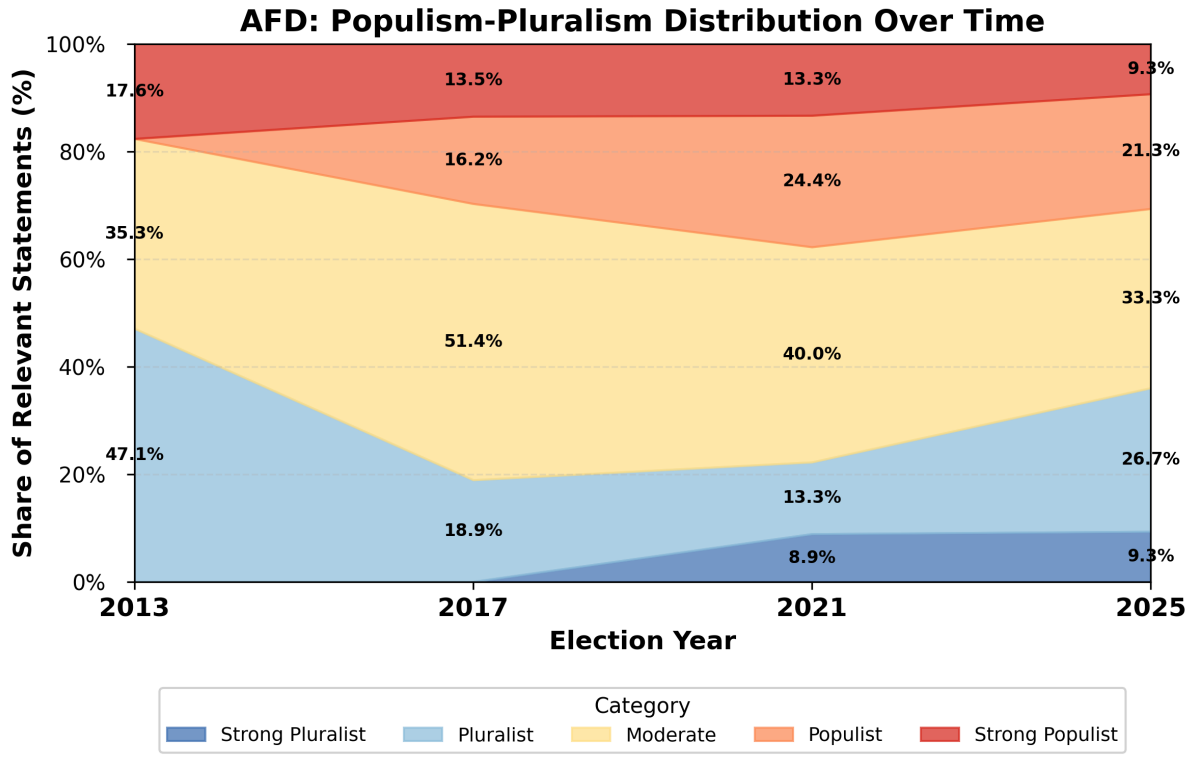

Figure 6: Distribution of populist statements in AfD manifestos.

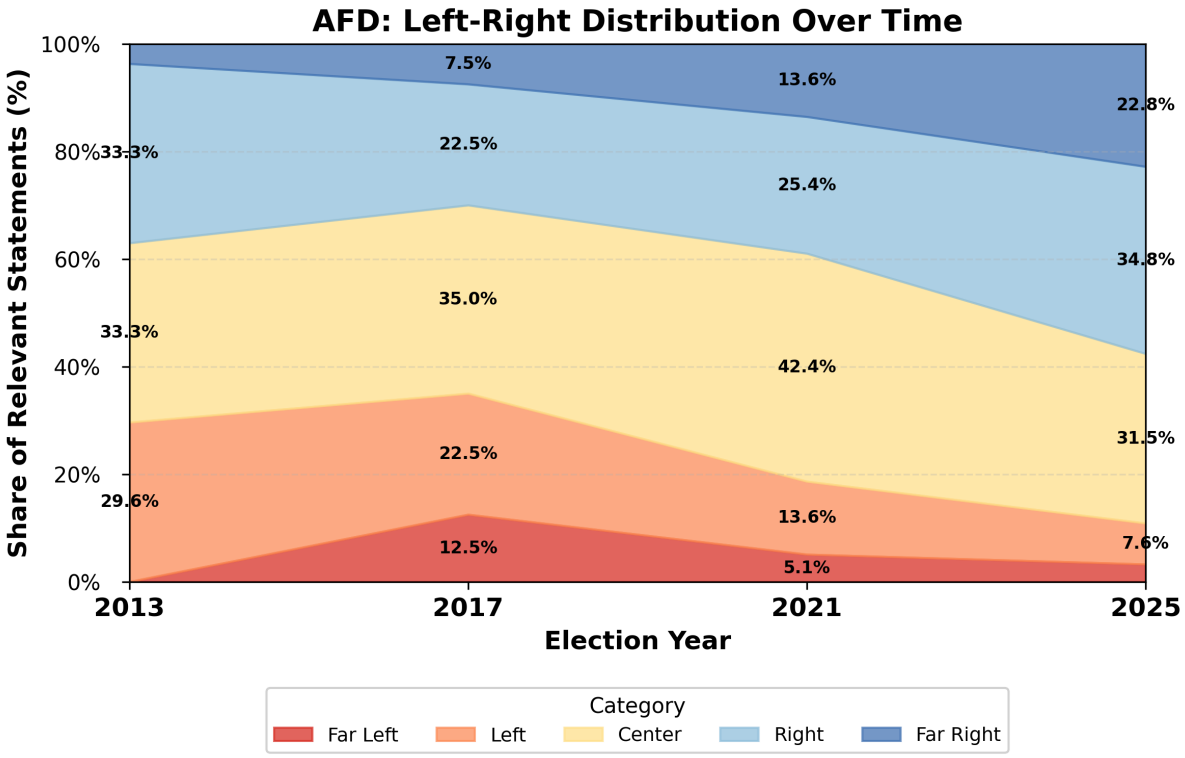

Figure 7: Distribution of economic left-right statements in AfD manifestos.

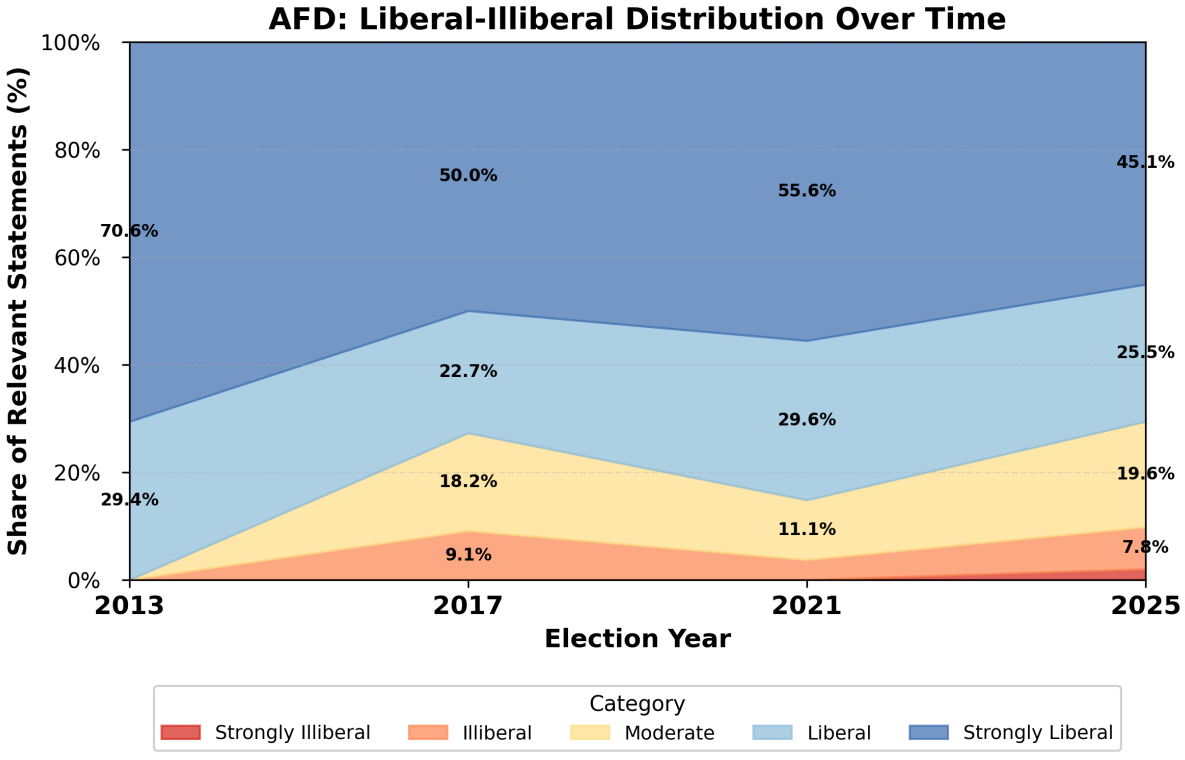

Figure 8: Distribution of liberal-illiberal statements in AfD manifestos.

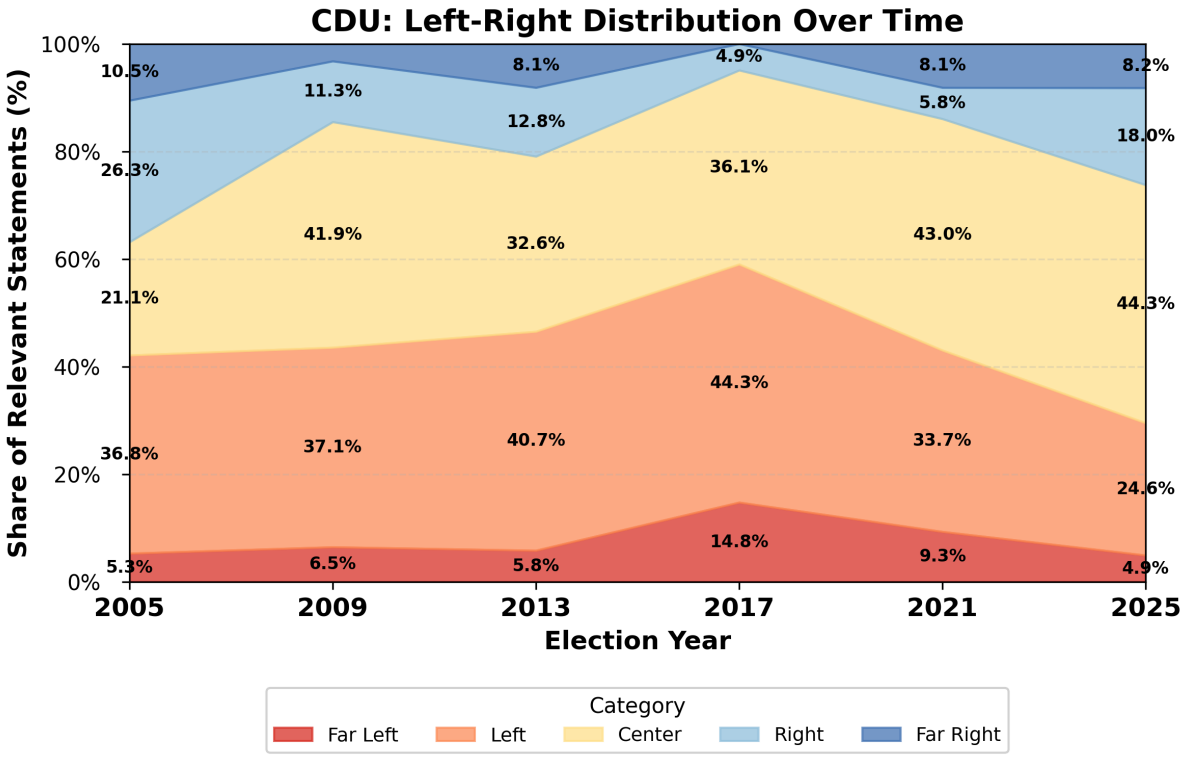

Figure 9: Distribution of economic left-right statements in CDU manifestos.

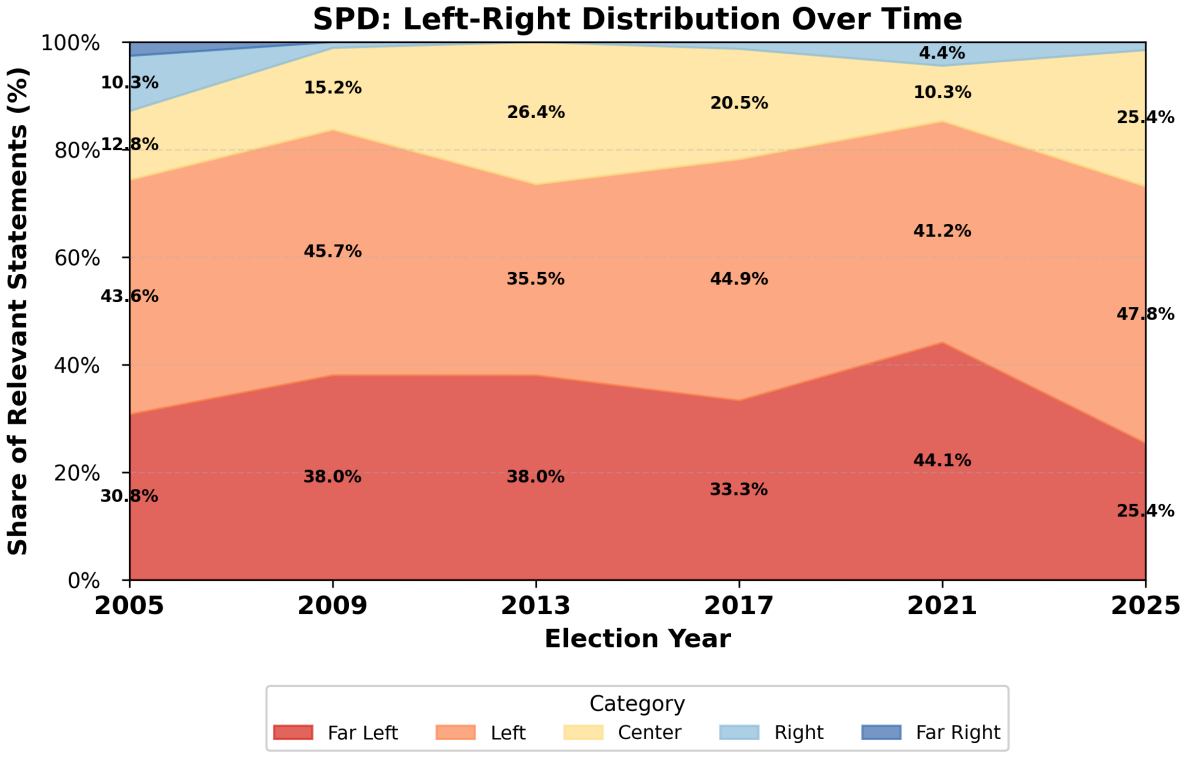

Figure 10: Distribution of economic left-right statements in SPD manifestos.