THE UNIVERSITY OF CHICAGO

A System-Agnostic Approach to Computational Ideology Detection: An analysis of Colombian congressional Speech (2000-2024)

By

Alejandro Sarria-Morales

April 2025

A paper submitted in partial fulfillment of the requirements for
the Master of Arts degree in the Master of Arts in
Computational Social Science

Faculty Advisor: John Levi Martin

Preceptor: David Peterson

Abstract

This study proposes a system-agnostic, modular approach for detecting ideological structures in political discourse without relying on predefined ideological axes or external labels. Building on a sociological understanding of ideology as the rationalization and narrative construction of political action, the method combines topic modeling, narrative mining, and frequent itemset analysis to inductively reconstruct ideological patterns from large-scale political speech. Applying this pipeline to a corpus of Colombian congressional interventions (2000–2024) revealed three major attitudinal stances toward peace processes, alongside narrative bridges linking economic governance, institutional trust, and post-conflict reform. These findings demonstrate that ideological structures can be systematically detected through emergent speech patterns, offering a flexible and culturally sensitive alternative to traditional ideology detection models. The results highlight new possibilities for computational social science in fragmented political contexts, particularly in regions like Latin America, where labeled data is limited and ideological systems are fluid.

Keywords: Ideology Detection, Political Discourse Analysis, Natural Language Processing, Computational Grounded Theory, Latin American Politics

---

1 Introduction

Political ideology, as an analytical construct, remains to be a surprisingly good tool for understanding political life. It offers itself as a way to make sense of how political actors align, how collective interests are framed, and how contestation over meaning unfolds in public discourse. Despite its conceptual elusiveness, ideology persists as a central analytic for explaining both individual belief systems and broader political formations.

Most popular among the plethora of efforts to formalize ideology have often turned to geometric abstraction (Bobbio, 1996), proposing ideology as a semantic space in which individuals, social movements, and other larger, more abstract, actors that could be reasonably described as *ideological* can be placed. While these models offer intuitive appeal and analytical utility, the contractions, expansions and transformations applied in these characterizations tend to impose rigid coordinates on phenomena that are contextually and historically variable.

The left-right axis is a particularly illustrative example of such approaches in virtue of its academic and colloquial ubiquity and its theoretical shortcomings. The left-right spectrum, as many low-dimensional models do, flattens ideological nuance into a simplified structure, often obscuring rather than revealing the inner workings of political alignment in multipolar

or unstable systems (Bauer et al., 2017). Its generalization across global contexts tends to compress unfamiliar formations or stretch them into misfit categories, especially in regions with different political genealogies.

Computational approaches have somewhat alleviated issues with low-dimensional geometric representation of ideology by introducing scale and precision to classification tasks, proving useful for quantitative tasks like politician sorting (Iyyer et al., 2014; Preoţiuc-Pietro et al., 2017) and vote prediction (Budhwar et al., 2018; Dietrich et al., 2019), and qualitative explorations of ideologically based phenomena (Barron et al., 2018; Bonikowski et al., 2022; DiMaggio et al., 2013; Nelson, 2021). However, in spite of the increased accuracy and validity gained by the implementation of computational analysis and use of large-scale data sources, most of these methods still reproduce the assumptions embedded in traditional models. They frequently treat party affiliation as a stand-in for ideology (Rheault & Cochrane, 2020), a shortcut that may function in bipartite or ideologically consistent systems, but fails in others. In regions like Latin America, where party systems are fragmented, fluid, and often weakly ideological (Lupu, 2016), such assumptions not only misclassify but also obscure the very ideological structures we seek to understand. Moreover, to date, there have been few systematic attempts to develop computational approaches that treat Latin American ideology as a question, rather than a byproduct of global north categories.

This project begins from that gap. Rather than imposing predefined ideological axes, I ask what representations of ideology emerge when analyzing political discourse without inherited coordinates. Drawing on inference-based, mixed methods frameworks for grounded theory for computational research (Nelson, 2020; Tavory & Timmermans, 2019), topic models (Barron et al., 2018; Nelson, 2021), and narrative mining (Ash et al., 2024), I propose a system-agnostic method that infers high-level representations from explicitly political utterances.

To do so, I first justify a definition of ideology as the narrative justification of the co-occurrence of different political stances within the same ideological set at the individual level, with the aggregation and merging of these interpretations as the representing larger ideological structures. My method uses a sequential pipeline of topic modeling, narrative mining and frequent itemset mining to identify the most salient frequently recurring patterns of political topic stances within a large corpus of political text, interpreting these as the most salient forms of ideology and valid input for further ideological analysis. Using a large corpus of congressional interventions in Colombia as a proof of concept, I show the validity and the usefulness of a pipeline capable of capturing high-level ideological signals that are endogenous to the political system under study, offering a more faithful and flexible foundation for comparative political analysis.

2 Literature Review

2.1 Ideological Models

Most theoretical models that represent the ideological position of political actors rely implicitly on a Geertzian (1994) interpretation of ideology as a symbolic system: a coherent set of cultural meanings providing individuals and groups with interpretive frameworks to navigate their social world and experiences. Keeping in the spirit of conceptualizing ideology as a set heuristic beliefs that guide interpretation. The classic approach towards political ideology paints it as an *a priori* political thesis: a cogent, internally consistent model for the overall goals of society, its guiding principles and the methods to get there (Downs, 1957). The classic approach follows to propose that ideology expands to integrate new stances by generating them based on pre-existing set of principles (Zaller, 1992).

To facilitate comparisons across settings, the foundational model of ideology in political analysis remains the traditional left–right spectrum. Originating from the seating arrangement of the 1789 French National Assembly, where radicals and commoners sat to the left and supporters of the monarchy to the right (Bobbio, 1996), the left–right dichotomy evolved into a standard ideological heuristic. Broadly, ideologies prioritizing social and economic equality have been placed on the left, while those favoring existing hierarchies or traditional order fall on the right. Over two centuries, despite shifts in the specific issues represented, this one-dimensional schema has persisted as the dominant analytical framework, especially in Western democracies(Imbeau et al., 2001). Philosophically, this model is closely tied to a classic approach of ideology, with "left" and "right" existing as relatively stable ideological packages contain normative ideas for political matter that are used as a base to generate new beliefs.

Nevertheless, the left–right spectrum exhibits significant limitations. A single axis inevitably oversimplifies the diversity of ideological positions, as evidenced by combinations of culturally conservative yet economically progressive beliefs (or vice versa) that defy linear classification (Bauer et al., 2017; Jankowski et al., 2023). Acknowledging this limitation, various alternatives and extensions have been proposed. A prominent example is Eysenck's (Eysenck, 1954) two-dimensional model, introducing an authoritarian–libertarian axis orthogonal to the traditional left–right dimension. Other analyses have identified additional axes such as nationalism versus internationalism or radicalism versus conservatism axis (Thurstone, 1931). While these multi-dimensional frameworks offer a richer mapping of ideological space and account for cross-cutting ideological dimensions, the historical resonance and analytical convenience of the left–right schema ensure its continued prevalence. Recognizing both the strengths and limitations of traditional ideological models thus underscores the importance of exploring alternative methodological approaches.

Following this line of critique, a reconsideration of the underlying theory for the left-right model and neighboring approaches is also pertinent. Empirical approaches to political ideology have provided evidence against the of ideology as coherent and/or normative as very rarely can individuals coherently self-report their own ideology while being much better at maintaining ideological consistency from one utterance to the next to the point of facilitating decision prediction(Vaisey, 2009). If ideologically consistent decisions do not necessarily originate from sound theory for political and moral organization, a look at ideology as origination from these decisions themselves by worth looking into.

Early Marxist definitions of ideology, understanding ideology as the highest level of abstraction of societal beliefs (Tucker, 1978), are useful for this scenario. Although this may seem uncomfortably close to a normative a priori model, a key difference lies in an understanding of the abstraction of societal beliefs as coming from a generalization of social relations, which in the case of political ideology is a specific process of formalizing political action (Martin, 2015). What is political action then and how can its identification to a better theory for political ideology? Historical analysis (Arendt, 1958), and the breadth and impact of text based political analysis (Grimmer & Stewart, 2013) posit speech as an undeniably salient avenue for political action. Based on the evidence presented above and relying on other in-depth inquires into defining ideology (Martin, 2015), I understand ideology as the result of individual exercises in interpretation and subsequent abstraction of each owns political action, mostly represented and signaled via political speech.

With the given theoretical limitations of the classic approach to ideology and the validity shortcomings of the left-right model, a question arises: How can the conceptualization of ideology as the narrative connecting stances scattered in a set be used for ideological representation without falling into the global validity issues of previous models? The solution to this answer, I propose, would be best suited by being based on the solution to the theoretical and methodological limitations of traditional representations of ideological model: First, the proposed pipeline must be based on utterance and individual level output, which has showed to be much more reliable at being used a signal for ideology (Vaisey, 2009), and, second, it must avoid using unreliable proxies for ideology, an effort that could be achieved via inference based approaches and grounded theory (Tavory & Timmermans, 2019). In the next section, I show that the building blocks for this task already exist within the realm of computational methods, and the extra perk of being able to process large amounts of data elevates this approach to pose it as an extremely useful, methodologically sound and theoretically solid, system-agnostic approach to political ideology.

2.2 Computational Approaches to the Representation of Ideology

Computational methods have increasingly been harnessed to tackle the classification and measurement of ideological constructs in political sociology and cultural analysis (Mohr et al., 2020). In recent years, these techniques have matured from being novelties to serving as rigorous tools for theoretically driven research questions (König et al., 2017; Rheault & Cochrane, 2020). Scholars have applied such methods to discern discursive frames, rhetorical themes, and categories of political culture at scales previously impractical, thereby bridging qualitative insights with quantitative rigor that allows for novel understandings of largely studied constructs (Bonikowski & Nelson, 2022).

One especially fruitful avenue has been the analysis of textual data to infer ideology. Text-based inquiries into political ideology have shown considerable success, particularly using natural language processing (NLP) and machine learning. By aiding close readings with large scale mining corpora of political speeches (Fuhse et al., 2020; Lauderdale & Herzog, 2016), manifestos (Rheault & Cochrane, 2020), and social media (Borja-Orozco, 2024; Preoţiuc-Pietro et al., 2017), among many others, ideological leanings and frames embedded in language can be algorithmically detected. Among several techniques, topic models have proven to be useful tools at systematically extracting ideological dimensions within explicitly political semantic spaces. Barron et al. (2018), for instance, utilized topic modeling to reveal political alignments emerging from consistent thematic groupings within parliamentary speech during the French Revolution. Similarly, Nelson (2021) employed topic modeling to map persistent place-based logics in the political discourse of American feminist movements, demonstrating how enduring thematic clusters shaped strategies and ideological continuity across historical periods.

Computational text-based methods do face notable limitations. These techniques traditionally rely either implicitly or explicitly on recognizable thematic coherence or stable ideological structures (Németh, 2023). In political contexts characterized by fluidity or fragmentation, such as weak party institutionalization or non-linear ideological configurations, topic models might struggle to yield clearly interpretable ideological axes without additional qualitative or interpretive guidance (Bauer et al., 2017; Jankowski et al., 2023). Thus, while topic modeling offers a powerful exploratory method to uncover latent dimensions of political debate, its successful application often requires careful theoretical framing and interpretation, particularly when analyzing political systems that deviate from conventional ideological schema. In response to this limitation, frameworks such as computational grounded theory (Nelson, 2020) have been proposed to offer a rigorous approach for inferring latent social, political, and cultural constructs from textual data, effectively integrating computational text analysis with qualitative interpretation and solving some of the issues with more strictly quantitative approaches. The method has been particularly useful as

an interpretative tool for topic models (Nelson, 2021) by allowing ideological dimensions to emerge organically from data rather than imposing predefined categories. Inductively identified themes are qualitatively refined and validated to ensure theoretical coherence and interpretive rigor. By integrating this approach, ideological complexity can be systematically explored, capturing dimension overlooked by traditional supervised approaches.

Yet, despite the individual and conjoined strength of NLP and inference-based frameworks setting a good methodological starting point for the study of political ideology, current computational models of ideology face significant limitations when applied to contexts in the Global South and other political systems outside of Western industrial democracies. Most existing approaches implicitly assume stable party systems and linear left–right ideological divisions, conditions that often do not hold outside the more traditionally studied political systems (Federico & Malka, 2023). In many Global South contexts, ideological structures are multidimensional, fluid, or fragmented with party affiliations in these regions being increasingly weak, personalistic, and/or transient (Baisotti, 2025; Lupu, 2016; Mainwaring, 2018), making them unreliable labels for supervised computational analysis. Thus, models calibrated on Western ideological frameworks often fail to capture or misrepresent these complex realities.

This study proposes an approach that tackles the two main issues with current frameworks for ideological characterization: a dependence on assumptions that do not hold up outside of the most studied contexts and a theoretical founding on a classic, normative, conceptualization of ideology. To do I propose a computationally based pipeline whose results are meant to be interpreted through inference and knowledge domain, allowing for the study of ideology to be expanded theoretically, methodologically and thematically.

2.3 The Colombian case

As a proof of concept for the method formulated in this study, I examine the Colombian congress from 2000 to 2024. Similar to other countries in the region, Colombia exhibits low levels of citizen identification with political parties (Lupu, 2016) paralleled with an increasingly loose association between political elite party membership and ideological standing (Meléndez, 2022). However, reduced party identification does not equate to diminished ideological or political thought among citizens (Segovia, 2022).

In recent decades, the traditionally salient Liberal–Conservative party labels have given way to new movements anchored by distinctive ideological projects. *Uribismo*, the politico-ideological current founded by former president Álvaro Uribe and his supporters, is identified as a neo-conservative project blending hard line security conservatism with neoliberal economics (Gamboa Gutiérrez, 2019; Kajsiu, 2019). Emerging more recently as an opposing ideological current is a resurgence of progressive politics exemplified by Gustavo Petro

and allied social movements. A comparative analysis of *Uribismo* and *Petrismo* finds that each represents a distinct ideological configuration with a clear social base—Uribismo's conservative, law-and-order appeals resonate most with Colombia's upper strata, while Petro's populist social-democratic platform draws its strongest support from lower-income and marginalized groups—and contest ideologically and programmatically along clear axes including, but not limited to, peace processes, economic models, and healthcare (Kajsiu, 2020).

It may seem that Uribismo and Petrismo represent the country's unique interpretation of the left-right ideological spectrum, but such a view would overlook important nuances. Several alternative ideological currents have been significant throughout this period (Mainwaring, 2018). Ideologies rooted in more traditional forms of opposition to dominant *caudillismo*, climate-centered discourses exemplified by the Green Party, and anti-corruption movements positioning themselves outside the conventional spectrum represent ideological forces that cannot neatly fit onto a traditional left-right continuum without going into substantially liberal interpretations of their political action (Escobar et al., 2023).

Crucially, these political actors have shifted allegiances with remarkable ease, often founding or reconfiguring parties and coalitions around emerging ideological currents. Because partisanship has proven a weak gauge of deeper commitments—and in many cases, ideology itself dictates party affiliation rather than the reverse—approaches that decouple ideological analysis from party labels become particularly pertinent. Low citizen identification with traditional parties, the presence of diverse yet overlapping ideological streams, and the malleable partisan identities of elites together make Colombia an ideal setting for testing novel frameworks. The following sections describe the data sources and methodological pipeline used to capture these currents, offering an approach that observes ideology in its own right rather than through the lens of party-based classifications and unfit geometries.

3 Data and Methods

3.1 Data

The interventions by congresspeople during official sessions, the central data source for this data, were obtained via a corpus of *Gacetas del Congreso* (Congress gazettes), official documents produced by the Colombian Congress for each session that records the proceedings of legislative sessions, including proposals, motions, votes and, crucially, a word for word transcript of each sessions. Congressional interventions within these sessions are structured around a guided discussion moderated by the presidency of the committee or chamber (in the case of plenary sessions), which assigns speaking turns and enforces procedural rules. This configuration typically leads to a discussion in which procedural speech is abundant,

as the presidency has speeches in between each substantive intervention, but where subject substantive information is relatively free-form and allows for a significant amount of syntactic and stylistic expression compared to legislative speech in other political systems.

Using a Selenium-based scraping script, I collected a total of 6,971 *Gacetas* from 2000 to 2024 in PDF format from the online repository of the Colombian Congress. To extract relevant text data, regular expressions, and NLP tools were employed, primarily using the *es_core_news_md* NLP model from spaCy (Honnibal & Montani, 2017). The texts were cleaned, and the debate sections were identified and extracted. The text was divided into presidential speaking turns and congressional speeches using regular expressions, allowing for the differentiation of procedural organizing speech from politically substantive interventions. Additional data processing was performed to extract the names of the congresspeople using Named Entity Recognition from the turn assignments. The result was a clean dataset consisting of 299,123 unique interventions from 587 unique congresspeople, each tagged with a session ID, date, and chamber. The distribution of these over the period of analysis is shown in Figure 1.

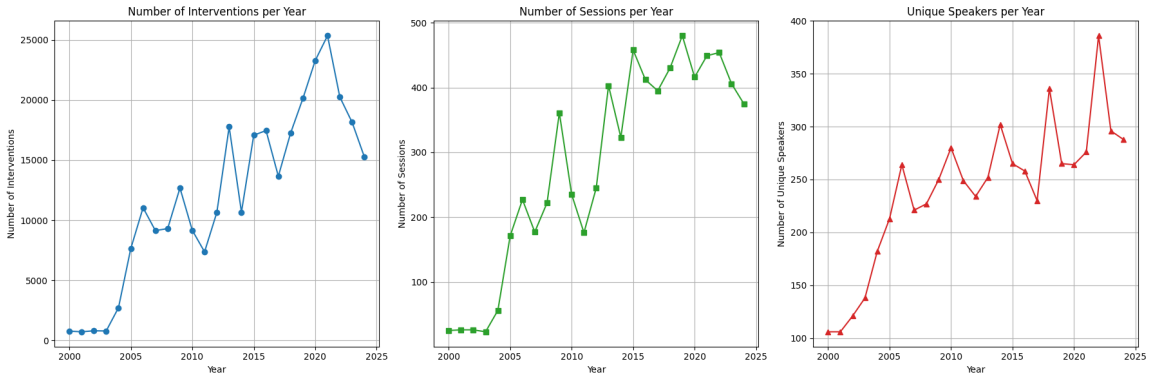

Figure 1: Number of interventions, sessions and unique speakers in the dataset over time

3.2 Methods

The representation of political ideology within the ecosystem of congressional speech as the patterns of frequently co-occurring narratives was methodologically tackled by way of a series of Latourian (Latour, 1999) abstractions of the original text (Figure 2). In summary, I first implemented a BERTopic topic model to detect the overall themes and discussion axes in the corpus. The outputs from this model and their implications were interpret using the three phases of the computational grounded theory framework (Nelson, 2020): pattern detection, pattern refinement, and pattern confirmation. Next, a narrative mining model was trained on the subset of each topic's intervention to extract topic narratives. Finally,

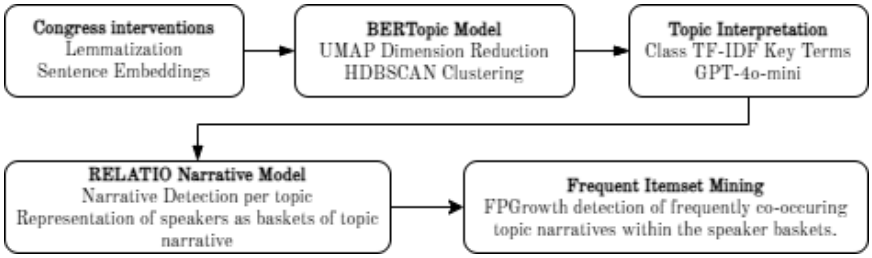

Figure 2: Methodological framework

representing each speaker as the set of topic narratives present in their intervention, a frequent itemset mining algorithm was applied to detect frequently co-occurring narratives, interpreting these results as signals of political ideology.

3.2.1 BERTopic model

Topic models are unsupervised statistical learning methods that use word co-occurrence patterns to infer latent topical categories from a collection of documents (Grimmer et al., 2022). Such models have become commonplace in social science research, applied to a wide range of texts such as legislative proceedings, court rulings, news coverage, and social media content. Among the available approaches, Latent Dirichlet Allocation (LDA) is the most widely used topic modeling technique, representing each document as a mixture of topics and each topic as a probability distribution over words. This approach has been successfully applied to congressional speech data to inductively identify issue agendas in legislative discourse.

For the present analysis, the BERTopic topic model (Grootendorst, 2022) was used to extract the main substantive themes of congressional discussion in Colombia. BERTopic overperforms traditional topic modeling approaches, such as LDA, by leveraging the strength of transformer based embedding models for document clustering and a class term frequency over independent document frequency (c-tfidf) metric for semantic representation. Interestingly, BERTopic is a modular topic model comprised of a series of submodels for each step: an embedding model to create the vector representations of documents: an embedding model to generate contextual embeddings, a dimensionality reduction model to reduce the complexity of the embeddings, a clustering model to group the embeddings into topics, and a representation model to interpret the topics and represent their contents.

After multiple iterations of hyperparameter tuning, the optimal configuration for the topic model, based on topic interpretability and Silhouette score testing, was achieved. The final configuration of the BERTopic model included the *paraphrase-multilingual-MiniLM-L12-v2* (Reimers & Gurevych, 2019) sentence embedding model, a multilingual BERT model

based on siamese BERT networks. For dimensionality reduction, UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2020) was employed to reduce the embedding dimensions to the minimum number of components required to capture 90% of the variance. The clustering model used was HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), selected for its ability to identify topics in the embedding space based on the differential density of data neighborhoods (Campello et al., 2013). Finally, the representation model combined KeyBERT-inspired embeddings and part-of-speech (POS) tagging included in BERTopic to output two lists of words that characterized each topic in complementary ways.

In the second phase, pattern refinement, the initial topics identified by the BERTopic model were further refined to ensure their relevance and clarity. The interpretation of the topics was fine-tuned by supplementing the model's representation with a guided deep reading and summaries generated by feeding the 10 most representative documents (according to the model) to the GPT-4o-mini Large Language Model (OpenAI, 2025). This step involved eliminating irrelevant topics, such as procedural speeches requesting to leave the session, that did not contribute to the substantive political discourse. Related topics were then grouped together, allowing for a more coherent set of themes that better represented the ideological structure of the debates. This refinement process was iterative, with human judgment and knowledge domain playing a key role in evaluating the interpretability and significance of the topics.

Finally, in the pattern confirmation phase, the interpretation of each of the reduced topics is tested by contrasting its fit toward a random subset of documents not yet seen (i.e. documents within the topic not selected by BERTopic as the most representative). The topic interpretation was then adapted to accurately describe the overarching theme of each while while accounting for and encapsulating the semantic variety.

3.2.2 Narrative mining

To extract meaningful narratives from the congressional speeches, the narrative mining model RELATIO (Ash et al., 2024) was employed. The RELATIO implementation in python, originally designed for English and French, was adapted to process Spanish text by integrating the spaCy *es_core_news_sm* model (Honnibal & Montani, 2017) for semantic role labeling and entity recognition. Some additional, minimal adjustments were made to the source code to ensure proper processing of Spanish text-data.

The model starts by extracting of syntactic structures from each intervention using part-of-speech tagging and named entity recognition. This results in the construction of subject-verb-object (SVO) triplets, such as "El presidente defiende la reforma" (The president defends the reform), where the subject is "el presidente", the verb is "defiende", and the

object is “la reforma.”

Once the SVO triplets are extracted, they are clustered using a K-means based clustering method included in the package. RELATIO automatically generates candidate cluster numbers and selects the optimal configuration based on silhouette and inertia scores. This allows for the identification of recurring patterns and narratives in the interventions. Each narrative is then represented as a triplet in the form: *subject1-role: subject1-entity, verb-role: verb, subject2-role: subject2-entity.*

The resulting narratives were then merged with the main dataset, linking each intervention to its corresponding narratives and associating the former with the legislators who uttered them.

3.2.3 Frequent itemset mining

Once the narratives were linked to the interventions, narrative *baskets* containing every unique topic narrative detected in their interventions were created for each speaker. These baskets were then analyzed using the FP-growth frequent itemset mining algorithm to detect frequently co-occurring topic narratives.

FP-growth is an efficient algorithm used for mining frequent itemsets, which are combinations of items (in this case, topic narratives) that occur together frequently in a dataset (Han et al., 2000). It works by constructing a compact structure called the FP-tree, which retains the essential information about frequent itemsets, and then recursively mines this tree for frequent itemsets without generating candidate itemsets explicitly. The algorithm builds the FP-tree by first scanning the dataset to count the frequency of individual items, and then using this information to organize the items into a tree structure, where each path represents a frequent combination of items.

To mine the tree, the algorithm examines the conditional pattern base (a subset of the tree that represents the co-occurrences of each item) and identifies frequent itemsets by exploring smaller sub-trees (called conditional FP-trees). This method is computationally more efficient than the Apriori algorithm, which generates candidate itemsets and prunes them based on their frequency.

For the analysis, the minimum support threshold was set at 0.5. Support is a measure of how frequently an itemset appears in the dataset. It is calculated by dividing the number of transactions that contain the itemset by the total number of transactions. A support of 0.5 means that an itemset must appear in at least half of the interventions to be considered frequent. The minimal length of the itemsets was set to 3, meaning that only itemsets containing at least three narratives were considered. This length was chosen before any testing, as it was deemed the minimal number required to draw meaningful inferences from the co-occurring narratives. The support threshold was determined after testing vari-

ous levels, ensuring that the selected threshold captured frequent and meaningful patterns without including too many irrelevant itemsets.

Going back to the understanding of political ideology in this study, the rationalization and interpretation of individual political action, frequently occurring sets of topic narratives are interpreted to represent signals ideological positions common enough to be considered a relevant ideological standing within the studied system of political speech. The implications of each of the most frequent and salient of these sets are explored by way of a culturally and politically situated interpretation.

4 Results

4.1 BERTopic model

The initial BERTopic model produced 168 distinct topics based on the semantic content of the congressional interventions. Upon manual inspection and pattern refinement, I identified and removed a subset of topics composed primarily of procedural speech—such as *31_morning_hour_afternoon_session*, a cluster centered on the scheduling of sessions, and *107_vote_vote_vote_vote*, a topic where speakers merely announced their vote. While these utterances are an unavoidable part of legislative proceedings, they contribute little to the characterization of the ideological dimensions under analysis. Their presence does highlight a particular challenge in working with congressional speech: the procedural scaffolding often dominates the textual record, threatening to obscure substantive content. Nevertheless, topic modeling proves effective in filtering through this procedural noise and isolating semantically meaningful interventions.

After removing these clusters and other interventions classified as noise, the resulting dataset contained 83,797 interventions classified into thematically coherent, non-procedural topics. Among these, a particularly large portion (15,760 interventions) was grouped under the main topic generated by the model: *0_colombians_colombian_colombia*, which revolves around national identity and competing narratives about the country's core political challenges. A recurring motif in this cluster is the diagnosis of deep-rooted land inequality as a central driver of both armed conflict and enduring social asymmetry. This topic, broad but ideologically resonant, serves as a discursive anchor in the corpus. Other dominant themes include *3_senator_president_minister*, which encapsulates appeals for collaboration, coalition-building, and rhetorical courtesy in favor of laws and projects regarding civil rights, and *4_budget_budgetary_minister*, which captures budgetary debates, with particular attention to demands directed at the Minister of Finance and the executive. These topics are not only substantively rich but also frequently invoked, suggesting their centrality in legislative discourse.

The topics from the BERTopic model and the categorization of congressional intervention as belonging to one of them paint an initial picture of legislative speech in Colombia by revealing the most frequent and salient topics that get discussed in congress. Going back to the settled understanding of ideology as the narrative interpretation of an individuals political action represented through its political speech, and interpreting this description and distribution of the intervention corpus, some initial results can be obtained.

Namely, the most relevant subject domains that political speech address are identified. Meaning that in the present case, political action at the legislative level is topically framed around diagnoses of political challenges, coalition building and resource allocation disputes, with other subject areas contributing to a lesser degree. This result is interesting in its own and it could warrant further exploration following an alternative analysis route. However, for the task at hand of identifying ideological signal using a ground-up, system-agnostic approach, this first step of relevant subject area identification lays the ground for the interpretation, laying the frame in which further results in the pipeline will be interpreted in. These topics, understood as semantically distinct bundles of speech, lay the foundation for the subsequent narrative analysis of how they coalesce into broader patterns of ideological signals.

4.2 Narrative Mining

After applying narrative mining with the RELATIO package, I identified a total of 6,977 unique narratives. These narratives were linked to individual congressional interventions and then aggregated into narrative *baskets*, representing the complete set of narratives utilized by each speaker throughout their interventions.

To enhance the interpretability and computational efficiency of subsequent analyses, I filtered these narratives using frequency thresholds. Narratives appearing in more than 80% of baskets were excluded, as they typically represent generic speech patterns lacking meaningful ideological content. For instance, the narrative 'I - Thank - Presidency,' ubiquitous across most congressional speeches, offers little insight into ideological differentiation. Similarly, narratives occurring in fewer than 20% of baskets were also omitted. This lower threshold ensures inclusion only of narratives shared broadly enough (uttered by at least 120 senators) to capture ideologically significant discursive patterns at scale, excluding overly niche or idiosyncratic narratives.

After applying these thresholds, the final count of relevant narratives stood at 303. The significantly reduced number of interventions is driven mostly by a large chunk of interventions that appear on few speaker topic narrative baskets. The top ranking interventions according to frequency (Table 1) and support (Table 2) are presented in the following tables. Support indicates the proportion of baskets containing that narrative, thus reflecting

its prevalence across speakers.

Topic	Narrative	Frequency
0_colombians_colombian_colombia	0_19: I-FIGHT-COLOMBIA	3087
0_colombians_colombian_colombia	0_20: WE-SUPPORT-PEACE	2843
0_colombians_colombian_colombia	0_21: THEY-IMPOSE-ACCORDS	2635
0_colombians_colombian_colombia	0_22: INVESTMENT-FALLS	2395
0_colombians_colombian_colombia	0_23: PRESIDENT-FIXES-SAFETY	2236
4_budget_budgetary_minister	4_2: WE-DEMAND-INFORMATION	2197
0_colombians_colombian_colombia	0_24: REFORM-SOLVES-DEFICIT	2083
0_colombians_colombian_colombia	0_25: PEOPLE-NEED-HOUSING	1902
4_budget_budgetary_minister	4_3: MINISTER-FAIL	1823
0_colombians_colombian_colombia	0_26: FAMILY-DESERVES-TRUTH	1766

Table 1: Most frequent topic narratives

Topic	Narrative	Support
0_colombians_colombian_colombia	0_19: I-FIGHT-COLOMBIA	0.7
3_senator_president_minister	3_5: FIGHT-CORRUPTION	0.69
4_budget_budgetary_minister	4_2: WE-DEMAND-INFORMATION	0.68
0_colombians_colombian_colombia	0_20: WE-SUPPORT-PEACE	0.67
0_colombians_colombian_colombia	0_21: THEY-IMPOSE-ACCORDS	0.67
4_budget_budgetary_minister	4_3: MINISTER-FAIL	0.65
3_senator_president_minister	3_6: POLICY-SERVE-NEEDS	0.64
0_colombians_colombian_colombia	0_22: INVESTMENT-FALL	0.64
4_budget_budgetary_minister	4_4: BUDGET-EXECUTE-BAD	0.63
0_colombians_colombian_colombia	0_23: PRESIDENT-FIX-SAFETY	0.63

Table 2: Most supported topic narratives

The prominence of narratives within the most frequent topics identified by the BERTopic model, such as *0_colombians_colombian_colombia*, *4_budget_budgetary_minister*, and *3_senator_president_minister*, was anticipated, validating the consistency and interpretive power of the methodological pipeline. Beyond confirming topical coherence, these narratives offer richer insight into the specific ideological stances that underpin discussions within each topic. For example, within the broader thematic category of diagnosing Colombia's key political challenges, the narratives *0_20: WE-SUPPORT-PEACE* and "_21: THEY-IMPOSE-ACCORDS" clearly articulate two competing positions on peace-building efforts, reflecting a highly salient political debate in the country. The model's capability to independently identify such politically resonant and contextually meaningful narratives lends substantial external validity to the approach.

Moreover, the identification of frequent and supported narratives allows a deeper exami-

nation of ideological content within and across topics. While individual-level analysis could yield intricate ideological profiles, the scale and scope of this study demands a higher level of abstraction. Thus, the next methodological step applies frequent itemset mining, specifically FP-growth, to uncover recurring narrative combinations, enabling the identification of broader ideological configurations at the systemic level.

4.3 Frequently co-occurring topic narrative sets

Finally, if we understand ideology as the narrative constructed to rationalize and explain political action, which for legislators can be captured in their political speech, a higher-level ideological representation can be built by identifying recurring sets of narratives within political speech and interpreting the unifying logic within the sets. Following this argument, I employed frequent itemset mining using the FP-growth algorithm, setting a minimum support threshold of 0.3 and a maximum narrative length of four. This analysis yielded 114,818 sets of frequently co-occurring topic narratives. Given the extensive time frame and scale of the data, it was expected that many combinations would emerge; thus, interpretation focuses primarily on the most salient and illustrative narrative sets. By increasing the support selection criteria to 0.5 and eliminating redundant and semantically similar frequent topic narrative sets, 212 sets were selected. After close reading of the interventions in which the narratives are used and comparing the unique framings used by different speakers, two distinct categories of frequent itemsets emerged Table 3: within-topic itemsets, representing coherent ideological narratives around specific substantive domains, and mixed-topic itemsets, highlighting bridging narratives across different subject areas. These are interpreted as signaling ideological positions consistent with salient and politically relevant issues, and guiding discursive strategies characteristic of Colombian politics, respectively.

4.3.1 Within-topic narrative sets

A first set of emerging topic narrative itemsets groups together frequently co-occurring narratives drawn from the same topic. Following the guiding definition that ideology is the rationalization and interpretation of one's political action, these sets capture ideological stances that remain internal to a particular substantive domain. In this sense, each within-topic narrative set can be understood as the union of ideological positions concerning the major axes of political debate identified earlier through topic modeling. Since the BERTopic model already highlighted the central subject matters dominating congressional discourse, such as national identity, governance, and fiscal management, analyzing the recurrent narrative combinations within each topic offers a way to characterize the principal stances legislators adopt toward these issues. These within-topic sets do not merely reflect

Within-topic Itemsets	**Mixed-topic Itemsets**
[4_1: congress-is-accountable, 4_0: pension-need-contribution, 4_3:minister-fail, 4_4: budget-execute-bad]	[0_16: government-finance-war, 0_13: government-kill-young, 0_20: we-support-peace, 3_0: president-support-no one]
[0_22: investment-falls, 0_18: groups-is-terrorists, 0_21: they-impose-accords, 0_17: regulation-not-work]	[3_1: collaboration-is-pleasure, 4_0: pension-need-contribution, 4_2: we-demand-information, 3_0: president-support-no one]
[0_23: president-fixes-safety, 0_19: i-fight-colombia, 0_20: we-support-peace, 0_14: land-concentrated]	[0_21: they-impose-accords, 0_12: we-believe-government, 4_2: we-demand-information, 3_0: president-support-no one]
[0_20: we-support-peace, 0_13: government-kill-young, 0_24: reform-solves-deficit, 0_14: land-concentrated]	[3_2: citizenship-bet-science, 4_0: pension-need-contribution, 0_12: we-believe-government, 3_3: minister-fails-government]
[3_5: fight-corruption, 3_1: collaboration-is-pleasure, 3_0: president-support-no one, 3_7: president-implicates-accomplices]	[4_4: budget-execute-bad, 4_0: pension-need-contribution, 0_15: system-boost-rural, 0_20: we-support-peace]

Table 3: Most salient within-topic and mixed-topic frequent itemsets

patterns of co-occurrence but structure coherent interpretive frameworks through which political actors frame challenges, attribute responsibility, and justify courses of action, thus serving as primary markers of ideological differentiation.

Within Topic 0, centered on diagnoses of Colombia's core political challenges, two distinct ideological groupings emerge from the frequent itemsets. One grouping ties narratives of peacebuilding, institutional reform, and presidential leadership together. The frequent co-occurrence of narratives such as *"we-support-peace" "reform-solves-deficit,"* and *"president-fixes-safety"* suggests an ideological stance that views structural reform and executive authority as necessary pillars for achieving lasting peace and national progress. Historically, in the period under study, reformism and peacebuilding have often been championed together by the same political forces (Kajsiu, 2020). However, the strong connection between peace narratives and deference to presidential leadership emerges here more clearly, highlighting an interesting alignment between support for institutional change and presidential authority. In contrast, the cluster of narratives *"they-imposed-accords," "groups-is-terrorists,"* and *"investment-falls"* captures ideological resistance to peace processes. Framing peace accords as externally imposed agreements and associating armed groups with terrorism were common rhetorical strategies among opponents of peace initiatives. Particularly significant is the frequent connection to *"investment-falls"* suggesting that opposition to peacebuilding efforts is not framed solely around national security but is also tied to concerns about maintaining investor confidence and economic stability. This linkage between anti-peace process stances and economic anxiety represents a novel finding, offering deeper insight into the ideological structure of peace process opposition in Colombian congressional discourse.

Within Topic 4, which captures congressional debates around budgetary issues and fiscal governance, the frequent itemsets reveal two intertwined ideological dimensions. On one hand, the co-occurrence of narratives such as *"congress-is-accountable"* and *"minister-fail"* reflects how responsibility for budgetary concerns is flexibly assigned between the executive and the legislature. Rather than representing exclusionary critiques, the frequent appearance of both narratives together suggests a flexible assignment of blame, with budgetary shortcomings pinned at different times on government ministers or on the congress itself. On the other hand, the recurring combination of "pension-need-contribution" and "budget-execute-bad" points to the central economic anxieties shaping fiscal discourse: the perceived frailty of the pension system and the inefficient execution of public budgets. These two concerns represent structural vulnerabilities that dominate economic discussions, independent of party or faction. The fact that narratives of blame allocation and systemic anxiety about state capacity appear together in the same itemsets suggests that budgetary speech moves simultaneously on two interpretive planes: one concerned with identifying actors to hold accountable, and another concerned with diagnosing the stability and future of the state's economic model itself.

Within Topic 3, which centers on governance, collaboration, and executive-legislative relations, the frequent itemsets reveal a coherent ideological stance emphasizing transparency, institutional integrity, and the dangers of political complicity. Narratives such as *"fight-corruption," "collaboration-is-pleasure," "president-support-no one,"* and *"president-implicates-accomplices"* frequently appear together, constructing a discursive field where collaboration is framed as desirable only when it is free from corruption and undue influence. The recurrent presence of anti-corruption narratives across these itemsets suggests that calls for clean governance are foundational ideological elements that transcend specific partisan affiliations. Similar to topic 4, rather than pointing to a single ideological camp, these narratives reflect a broadly shared interpretive framework where critiques of corruption, accusations of executive complicity, and appeals to genuine collaboration form central pillars of legitimate political action. This ubiquity also underlines the extent to which corruption accusations operate as a general moral language through which political actors position themselves and attack opposition, reinforcing their credibility and justifying their stances within legislative discourse.

4.3.2 Between-topic narrative sets

A second set of emerging itemsets groups narratives drawn from different topics. In the same way that within-topic narrative sets represent ideological stances and the narrative dimensions through which key political issues are interpreted, mixed-topic sets reveal how these stances are bridged and rationalized across different domains. It is through these cross-topic

connections that broader ideological configurations are articulated. In this sense, mixed-topic itemsets embody the working definition of ideology guiding this study: the narrative interpretation of political action through the connection and integration of different stance positions into coherent political structures.

The first salient mixed-topic itemset brings together *"government-finance-war," "government-kill-young," "we-support-peace,"* and *"president-support-no one."* This set constructs a pro-peace ideological stance that acknowledges the role of the state in perpetuating violence and conflict, differentiation itself from previously identified pro-peace, pro-institution narrative itemsets. By linking support for peace with direct critiques of government responsibility in the armed conflict, this narrative set frames peace not as a neutral or universally agreed objective, but as a critical project tied to the recognition of historical state failures. The inclusion of "president-support-no one" adds an additional layer of critical distance, suggesting that even peace-supporting actors resist alignment with particular leaders, favoring broader structural critiques over personalistic endorsements.

Another important set combines *"they-impose-accords," "we-believe-government," "we-demand-information,"* and *"president-support-no one."* This grouping illustrates how opposition to peace accords is narratively linked to an appeal to institutional principles rather than overt support for government leadership. In Colombia, peace processes have often been led by institutional actors, which makes institutionalism a contested terrain. This itemset suggests that while opponents of peace efforts criticize the specific governments involved, they simultaneously frame their position within appeals to institutional trust and demands for transparency. Thus, opposition becomes personalistic, directed against particular political figures, while still invoking institutionalism as a rhetorical resource to legitimize critique.

A final cluster of mixed-topic itemsets reveals how broader social justice claims are constructed through cross-topic narrative bridging. The combination of "budget-execute-bad," "pension-need-contribution," "system-boost-rural," and "we-support-peace" links demands for peace with concerns about economic mismanagement and rural inequality. This set articulates an ideological stance where peacebuilding is tied directly to the promise of improved social and economic conditions, especially for rural populations historically marginalized by conflict and economic exclusion. Compared to another itemset, *"citizenship-bet-science," "pension-need-contribution," "we-believe-government,"* and *"minister-fails-government"*, which presents a more technocratic vision of social betterment grounded in governance reforms and scientific rationality, the peace-centered itemset frames social justice as emerging from structural transformation driven by peace processes themselves.

Together, these mixed-topic itemsets illustrate how ideology, as defined in this study, emerges from the rationalization and linkage of political stances across domains. The connections drawn between peacebuilding, economic reform, government responsibility, and

institutional trust represent not isolated opinions but integrated narratives that explain and justify political action. They reveal the flexible, strategic, and often multi-dimensional nature of ideological construction in the Colombian congressional discourse.

5 Discussion

This study set out to build a system-agnostic, modular pipeline capable of detecting ideological structures in political discourse without relying on predefined axes or labels. Moving away from conceptualizations of ideology as a fixed set of doctrines or as the stable mapping of actors onto low-dimensional spaces. I draw from a sociological framing to understand ideology as the rationalization and narrative construction of political action (Martin, 2015). Finding incompatibility between this definition and traditional ideology representation models and their limits, which become clearer in fragmented, fluid political systems like Colombia's; the aim of this paper was to propose a method that captures ideology inductively at scale, through the organization and connection of speech, rather than through the imposition of inherited ideological coordinates. The methodological strategy operationalized this goal through a sequential pipeline of topic modeling (to detect discourse axes), narrative modeling (to extract stances within topics), and frequent itemset mining (to detect patterns of co-occurrence between stances), enabling the bottom-up reconstruction of ideological structures.

The existing literature on computational approaches to ideology detection has provided powerful tools for large-scale text analysis but remains constrained by key assumptions. Predominantly, these models rely on party affiliation as a proxy for ideology (Rheault & Cochrane, 2020) and assume relatively stable, well-mapped ideological structures, often inherited from studies of Western bipartite systems (Németh, 2023). In contexts like Colombia, where partisanship is weak and ideological alignments are fluid (Lupu, 2016; Meléndez, 2022), these models misclassify or obscure the ideological structures they seek to capture. Essentially, current computational approaches often impose exogenous labels onto data, ignoring the possibility that ideological dimensions may emerge differently across settings.

By contrast, the method proposed here directly addresses these limitations: first, by grounding ideological representations in the structure of political speech itself rather than external proxies; second, by allowing ideological dimensions to emerge inductively rather than being predefined; and third, by using a modular pipeline built from established computational methods reinterpreted through a grounded theory framework (Nelson, 2020), preserving flexibility, scalability, and theoretical rigor.

When applied to the Colombian congressional corpus, this method revealed coherent and context-specific ideological structures. The results show that ideological organization

in the Colombian congress is strongly shaped by attitudes toward peace processes, economic anxieties, and concerns about governance integrity. Three main attitudinal groups toward peacebuilding were detected: a pro-peace stance aligned with governmental efforts, a pro-peace stance critical of government actions, and an anti-peace process stance. The first group connects support for peace initiatives with narratives of structural reform and deference to presidential authority; the second supports peace while critically assigning government responsibility for violence and mismanagement; the third frames peace accords as externally imposed and undermining national security, linking opposition to broader economic concerns such as investor confidence. Notably, these ideological patterns cannot be adequately mapped onto traditional left–right models or even more elaborate dimensional schemes; they are dimensions specific to the Colombian political context, emerging organically from the text through the proposed pipeline.

The bridging of narratives across topics, as evidenced in the mixed-topic frequent itemsets, provides the clearest operationalization of the study's definition of ideology as the rationalization of political action through connected narrative structures. These bridges show how political actors link stances across thematic domains, such as peacebuilding, economic governance, and legislative trust, into ideological frameworks. For example, narratives connecting rural development demands with critiques of budget execution and support for peace processes reveal an integrated vision of post-conflict economic and social reform. Similarly, narrative sets that join skepticism toward peace accords with calls for institutional transparency indicate how actors weave opposition to specific policies into broader appeals for procedural legitimacy. These cross-topic connections do not simply reflect thematic proximity; they demonstrate how apparently disparate concerns are narratively rationalized as an ideology, allowing political actors to construct flexible yet meaningful systems of belief from the speech they produce.

More broadly, these results show that political stances, once systematically and contextually linked through grounded narrative structures, can serve as culturally validated proxies for ideology. This advances the methodological toolkit for computational political analysis by providing a grounded basis for supervised approaches such as stance detection or ideology classification, without requiring external labels. It also enables the development of culturally anchored ideological labels, which can facilitate scalable comparative research when adapted carefully to new contexts. Crucially, because the method is modular, drawing on widely available techniques like BERTopic, RELATIO, and FP-growth, it can be easily adapted and improved as newer, more powerful methods emerge for each task in the pipeline. This flexibility ensures the long-term relevance and expandability of the approach.

By focusing on frequently co-occurring narratives rather than relying on static ideological dimensions, the method offers a framework for capturing ideological "incongruence",

situations where seemingly contradictory stances coexist within political actors' speech. For instance, a speaker may be represented as simultaneously supporting rural reform (traditionally progressive) while emphasizing strict security measures (traditionally conservative), if such combinations are seen in genuine patterns in the data. Unlike traditional supervised models, which would often misclassify such inconsistencies, this approach preserves the complexity and layered nature of real-world ideological articulation. It also allows for new substantive questions about political dynamics to emerge from the data. For instance, the results suggest that blame for economic uncertainty in Colombia is assigned both to the executive and the legislature, a dynamic that traditional left–right positioning or basic sentiment analysis approaches would likely miss. Capturing these subtle shifts offers a richer, more sociologically grounded understanding of how political discourse structures ideological meaning.

While the application of the proposed method as a proof-of-concept was successful, some limitations should be acknowledged. Although congressional speech offers a valuable window into elite political discourse, it cannot be assumed to fully represent the ideological structure of broader Colombian society, leading to validity constraints for generalizing the results for the broader consideration of Colombian political ideological. Additionally, the application of support thresholds, while necessary for interpretability, inevitably excludes lower-frequency narratives that might be substantively important, alternative measures for extracting narrative itemsets should be explored.

Future work could adapt the method to other corpora, such as citizen discourse, public petitions, or digital platforms, where similar pipelines could reveal different ideological formations. There is also potential to explore cautious between-system comparisons by standardizing the categorization of emerging narrative structures, offering exciting possibilities for comparative ideology research. Finally, future consolidation of the pipeline into an integrated software package would improve accessibility and encourage broader application.

In sum, this study contributes a flexible, theoretically grounded, and system-agnostic approach to the computational analysis of political ideology. By foregrounding the inductive emergence of ideological structures from political speech, it moves beyond traditional models' limitations and opens new avenues for interpreting and modeling ideology in diverse political contexts. In doing so, it strengthens the bridge between computational methods and sociological theory, offering new tools to understand how political meaning is structured, contested, and reassembled through discourse.

6 Conclusion

This study set out to address a central challenge in the computational study of political ideology: how to capture ideological structures without relying on predefined axes, externally imposed labels, or assumptions of system stability. Motivated by the theoretical and empirical limitations of traditional models, and drawing from sociological redefinitions of ideology as the rationalization and narrative construction of political action, I proposed a system-agnostic, modular pipeline capable of inductively detecting ideology from large-scale corpora of political speech.

The method developed here combined topic modeling, narrative mining, and frequent itemset analysis to reconstruct ideological configurations from the ground up. Rather than presupposing ideological categories, the pipeline allowed ideological structures to emerge through the patterns and rationalizations embedded within congressional discourse. Applying this method to Colombian legislative speech revealed coherent, context-specific ideological formations, particularly around peacebuilding, economic anxieties, and governance integrity. The detection of bridging narratives across distinct topics further illustrated how political actors construct integrated ideological frameworks by linking diverse domains of concern.

This work contributes a scalable, flexible, and theoretically grounded approach to ideology detection that respects the contextual emergence of political meaning. By demonstrating that political stances, when systematically linked through grounded narrative structures, can serve as culturally validated indicators of ideology, this project advances both computational methodologies and sociological theory on political meaning.

Beyond its specific empirical findings, this study opens new directions for computational social science, particularly in regions like Latin America where the availability of labeled data is limited and party systems are unstable. The proposed approach encourages research that foregrounds local meaning structures, reduces reliance on Western-centric ideological models, and leverages computational techniques without sacrificing theoretical coherence and interpretability. In doing so, it contributes to building a more context-sensitive, empirically grounded foundation for the computational analysis of political life.

Data and Code Availability Statement

The primary data used in this study consists of congressional speech transcripts published in the *Gaceta del Congreso de la República de Colombia* between 2000 and 2024. These documents are publicly accessible through the official website of the Colombian congress (https://svrpubindc.imprenta.gov.co/senado/), though the retrieval process re-

quires automated scraping due to the absence of a centralized bulk download option. All scripts used for data collection, cleaning, pre-processing and data analysis are available in the project's GitHub repository.

Due to the large size and processing requirements of the dataset, intermediate and final versions of the structured corpus, including metadata-enriched and cleaned versions, can be shared upon request.

References

Arendt, H. (1958). *The Human Condition: Second Edition* (M. Canovan & a. N. F. b. D. Allen, Eds.). University of Chicago Press. Retrieved April 24, 2025, from https://press.uchicago.edu/ucp/books/book/chicago/H/bo29137972.html

Ash, E., Gauthier, G., & Widmer, P. (2024). Relatio: Text Semantics Capture Political and Economic Narratives. *Political Analysis*, *32*(1), 115–132. https://doi.org/10.1017/pan.2023.8

Baisotti, P. (2025). 'One, Two, many Latin American lefts?' Ideology and scepticism in the twenty-first century. In *Ideology, post-ideology and anti-ideology in Latin America: Reflections from the last decade* (pp. 11–56). Bloomsbury Academic.

Barron, A. T. J., Huang, J., Spang, R. L., & DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the French Revolution [Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *115*(18), 4607–4612. https://doi.org/10.1073/pnas.1717729115

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the Left-Right Scale a Valid Measure of Ideology? *Political Behavior*, *39*(3), 553–583. https://doi.org/10.1007/s11109-016-9368-2

Bobbio, N. (1996). *Left and Right: The Significance of a Political Distinction* [Google-Books-ID: jdw9MDQa4mEC]. University of Chicago Press.

Bonikowski, B., Luo, Y., & Stuhler, O. (2022). Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models [Publisher: SAGE Publications Inc]. *Sociological Methods & Research*, *51*(4), 1721–1787. https://doi.org/10.1177/00491241221122317

Bonikowski, B., & Nelson, L. K. (2022). From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research [Publisher: SAGE Publications Inc]. *Sociological Methods & Research*, *51*(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Borja-Orozco, H. (2024). Deslegitimación del adversario y orientación ideológica: Análisis de publicaciones de dos líderes políticos colombianos en Twitter [Number: 1]. *Acta Colombiana de Psicología*, *27*(1), 17–36. https://doi.org/10.14718/ACP.2024.27.1.2

Budhwar, A., Kuboi, T., Dekhtyar, A., & Khosmood, F. (2018). Predicting the Vote Using Legislative Speech. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 1–10. https://doi.org/10.1145/3209281.3209374

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Dietrich, B. J., Enos, R. D., & Sen, M. (2019). Emotional Arousal Predicts Voting on the U.S. Supreme Court [Publisher: Cambridge University Press]. *Political Analysis*, *27*(2), 237–243. https://doi.org/10.1017/pan.2018.47

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Downs, A. (1957). An Economic Theory of Political Action in a Democracy [Publisher: The University of Chicago Press]. *Journal of Political Economy*, *65*(2), 135–150. https://doi.org/10.1086/257897

Escobar, J. C., Ortega, B., & Wills-Otero, L. (2023). Elecciones presidenciales y legislativas en Colombia en 2022 [Number: 116 Publisher: Universidad de los Andes]. *Colombia Internacional*, (116), 3–28. Retrieved April 4, 2025, from https://journals.openedition.org/colombiaint/19300

Eysenck, H. J. (1954). *The Psychology of Politics*. Routledge. https://doi.org/10.4324/9781351303088

Federico, C. M., & Malka, A. (2023, September). The Psychological and Social Foundations of Ideological Belief Systems. In L. Huddy, D. O. Sears, J. S. Levy, & J. Jerit (Eds.), *The Oxford Handbook of Political Psychology* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197541302.013.16

Fuhse, J., Stuhler, O., Riebling, J., & Martin, J. L. (2020). Relating social and symbolic relations in quantitative text analysis. A study of parliamentary discourse in the Weimar Republic. *Poetics*, *78*, 101363. https://doi.org/10.1016/j.poetic.2019.04.004

Gamboa Gutiérrez, L. (2019). El reajuste de la derecha colombiana. El éxito electoral del uribismo [Publisher: Facultad de Ciencias Sociales Section: Colombia Interna-

cional]. *Colombia Internacional*, (99), 187–214. Retrieved April 4, 2025, from https://dialnet.unirioja.es/servlet/articulo?codigo=7027949

Geertz, C. (1994). Ideology as a Cultural System [Num Pages: 16]. In *Ideology*. Routledge.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022, January). *Text as Data: A New Framework for Machine Learning and the Social Sciences* [Google-Books-ID: dL40EAAAQBAJ]. Princeton University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Grootendorst, M. (2022, March). BERTopic: Neural topic modeling with a class-based TF-IDF procedure [arXiv:2203.05794 [cs]]. https://doi.org/10.48550/arXiv.2203.05794

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Rec.*, *29*(2), 1–12. https://doi.org/10.1145/335191.335372

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*

Imbeau, L. M., Pétry, F., & Lamari, M. (2001). Left-right party ideology and government policies: A meta-analysis. *European Journal of Political Research*, *40*(1), 1–29. https://doi.org/10.1023/A:1011889915999

Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014, June). Political Ideology Detection Using Recursive Neural Networks. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1113–1122). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1105

Jankowski, M., Schneider, S. H., & Tepe, M. (2023). How stable are 'left' and 'right'? A morphological analysis using open-ended survey responses of parliamentary candidates [Publisher: SAGE Publications Ltd]. *Party Politics*, *29*(1), 26–39. https://doi.org/10.1177/13540688211059800

Kajsiu, B. (2019). The Colombian Right: The political ideology and mobilization of Uribismo [Publisher: Routledge]. *Canadian Journal of Latin American and Caribbean Studies / Revue canadienne des études latino-américaines et caraïbes*. Retrieved April 4, 2025, from https://www.tandfonline.com/doi/abs/10.1080/08263663.2019.1581495

Kajsiu, B. (2020). Las ideologías y movilizaciones políticas del Uribismo y Petrismo: Dos Colombias distintas [Publisher: Universidad Nacional de Colombia Section: Análisis Político]. *Análisis Político*, *33*(98), 191–209. Retrieved April 4, 2025, from https://dialnet.unirioja.es/servlet/articulo?codigo=9471852

König, T., Marbach, M., & Osnabrügge, M. (2017). Estimating Party Positions across Countries and Time—A Dynamic Latent Variable Model for Manifesto Data. *Political Analysis*, *21*(4), 468–491. https://doi.org/10.1093/pan/mpt003

Latour, B. (1999, June). *Pandora's Hope: Essays on the Reality of Science Studies* [Google-Books-ID: 2xz9DwAAQBAJ]. Harvard University Press.

Lauderdale, B. E., & Herzog, A. (2016). Measuring Political Positions from Legislative Speech. *Political Analysis*, *24*(3), 374–394. https://doi.org/10.1093/pan/mpw017

Lupu, N. (2016). *Party brands in crisis: Partisanship, brand dilution, and the breakdown of political parties in Latin America*. Cambridge University Press.

Mainwaring, S. (2018, February). *Party Systems in Latin America* [Google-Books-ID: zDhFDwAAQBAJ]. Cambridge University Press.

Martin, J. L. (2015). What is ideology? [Number: 77]. *Sociologia, Problemas e Práticas*, (77). https://doi.org/10.7458/SPP2015776220

McInnes, L., Healy, J., & Melville, J. (2020, September). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [arXiv:1802.03426 [stat]]. https://doi.org/10.48550/arXiv.1802.03426

Meléndez, C. (2022). The Post-Partisans: Anti-Partisans, Anti-Establishment Identifiers, and Apartisans in Latin America [ISBN: 9781108694308 9781108717366 Publisher: Cambridge University Press]. *Elements in Politics and Society in Latin America*. https://doi.org/10.1017/9781108694308

Mohr, J. W., Bail, C. A., Frye, M., Lena, J. C., Lizardo, O., McDonnel, T. E., Mische, A., Tavory, I., & Wherry, F. F. (2020). *Measuring Culture*. Columbia University Press. Retrieved September 10, 2024, from https://cup.columbia.edu/book/measuring-culture/9780231180290

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework [Publisher: SAGE Publications Inc]. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Nelson, L. K. (2021). Cycles of Conflict, a Century of Continuity: The Impact of Persistent Place-Based Political Logics on Social Movement Strategy [Publisher: The University of Chicago Press]. *American Journal of Sociology*, *127*(1), 1–59. https://doi.org/10.1086/714915

Németh, R. (2023). A scoping review on the use of natural language processing in research on political polarization: Trends and research prospects. *Journal of Computational Social Science*, *6*(1), 289–313. https://doi.org/10.1007/s42001-022-00196-2

OpenAI. (2025). GPT-4o-mini. https://chat.openai.com/

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017, July). Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In R. Barzilay & M.-Y. Kan (Eds.),

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 729–740). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1068

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/1908.10084

Rheault, L., & Cochrane, C. (2020). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*, *28*(1), 112–133. https://doi.org/10.1017/pan.2019.26

Segovia, C. (2022). Affective polarization in low-partisanship societies. The case of Chile 1990–2021 [Publisher: Frontiers]. *Frontiers in Political Science*, *4*. https://doi.org/10.3389/fpos.2022.928586

Tavory, I., & Timmermans, S. (2019). Abductive analysis and grounded theory [Publisher: Sage Thousand Oaks, CA]. *The SAGE handbook of current developments in grounded theory*, 532–546.

Thurstone, L. L. (1931). The measurement of social attitudes [Place: US Publisher: American Psychological Association]. *The Journal of Abnormal and Social Psychology*, *26*(3), 249–269. https://doi.org/10.1037/h0070363

Tucker, R. C. (1978). *The Marx-Engels reader* [Publisher: Norton New York].

Vaisey, S. (2009). Motivation and Justification: A Dual-Process Model of Culture in Action1 [Publisher: The University of Chicago Press]. *American Journal of Sociology*. https://doi.org/10.1086/597179

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion* (18th ed.). Cambridge University Press.