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

***Abstract*—This study proposes a modular approach for detecting ideological structures in political discourse without relying on predefined ideological axes, external labels or political system specifics'. Building on a sociological understanding of ideology as the narrative construction of political action, the method combines topic modeling, narrative mining, and frequent itemset analysis to inductively reconstruct ideological patterns from large-scale datasets of 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.**

***Index Terms*—Ideology Detection, Content Analysis, Political Discourse Analysis, Natural Language Processing, Computational Grounded Theory.**

I. Introduction

Despite its conceptual elusiveness, ideology persists as a useful analytic construct for linking individual belief systems to broader political currents. Most efforts to formalize ideology have relied on geometric abstractions [1] treating it as a semantic space in which individuals, social movements, and other actors can be situated. While these models offer intuitive appeal and analytical utility, the contractions, expansions and other transformations applied in these characterizations tend to impose rigid coordinates on phenomena that, by flattening ideological nuance into a single spectrum, obscure alignment in multipolar or unstable systems [2]. When generalized globally, it distorts political genealogies leading to the continued misrepresentation of the political thought of understudied settings in scholarly work.

Computational methods have advanced the study of ideology while somewhat solving the issues of low-dimensional representations by scaling classification tasks and adding precision, proving useful for sorting politicians [3], [4], predicting votes [5], [6], and exploring ideological phenomena [7]–[10]. Yet, most continue to reproduce assumptions of traditional models. Gravely, party affiliation is often used as a proxy for ideology [11], a shortcut that works in contexts with ideologically coherent and stable parties, and assumption that in many regions, such as Latin America, does not hold [12]. The application of inappropriate, low-dimensional models then obscures ideological structures in these contexts.

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 and propose a methodological pipeline capable of representing ideology more accurately.

II. Literature Review

A. Ideological Models

Most theoretical models that represent the ideological positions of political actors rely on Geertzian interpretation of ideology as a symbolic system: a coherent set of cultural meanings that provides individuals and groups with interpretive frameworks for navigating their social world [13]. From this framework, ideology functions as a structured set of heuristics guiding interpretation. The classic approach builds on this by framing ideology as an *a priori* thesis: an internally consistent model of society's goals, principles, and methods [14]. Within this formulation, ideology expands by generating new stances from preexisting principles [15].

The most common representation in this space remains to be the left–right continuum. Originating in the 1789 French National Assembly, with radicals to the left and monarchy supporters to the right [1], it has become a general framework for distinguishing ideologies: equality and redistribution on the left, hierarchy and tradition on the right. Despite shifts in the issues mapped onto it, the model has endured, especially in Western democracies [16]. Philosophically, it aligns with the classic approach, treating "left" and "right" as stable packages of normative ideas that generate new stances. Yet the spectrum's simplicity is also its weakness. A single axis collapses ideological diversity, obscuring combinations such as cultural conservatism with economic progressivism [2], [17]. Alternatives have added dimensions, such as Eysenck's authoritarian–libertarian axis [18] or Thurstone's nationalism–internationalism divide [19]. While richer, these frameworks never displaced the convenience and resonance of left–right. Its persistence, despite acknowledged shortcomings, highlights the need for alternatives.

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 [20]. 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.

Marxist traditions defining ideology as highest level of abstraction societal beliefs [21], provide a useful lens. While this resembles the *a priori* model, Marxism saw ideology as the generalization of lived social relations rather than imposed principles. Political ideology could then be interpreted as the formalization of political action [22]. Historical analysis [23], and the breadth and impact of text based political analysis [24] 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 [22], 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 [20], and, second, it must avoid using unreliable proxies for ideology, an effort that could be achieved via inference based approaches and grounded theory [25]. 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.

B. Computational Approaches to the Representation of Ideology

In recent years, computational methods have increasingly been harnessed to tackle the classification and measurement of ideological constructs in political sociology and cultural analysis [26] maturing from being novelties to serving as rigorous tools for theoretically driven research questions [11], [27]. Previous work has 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 [28].

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 to aid close readings with large-scale mining of corpora of political speeches [29], [30], manifestos [11], and social media [3], [31], among many others, algorithmically detecting ideological leanings and frames embedded in language. Among several techniques, topic models have proven to be useful tools for systematically extracting ideological dimensions within explicitly political semantic spaces; Barron et al. [8], for instance, utilized topic modeling to reveal political alignments emerging from consistent thematic groupings within parliamentary speech during the French Revolution. Similarly, Nelson [7] 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 by traditionally relying, either implicitly or explicitly, on recognizable thematic coherence or stable ideological structures [32]. 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 [2], [17]. 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 [33] 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 [7] 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 dimensions 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 [34]. 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 [12], [35], [36], 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 so, I propose a computationally based pipeline whose results are meant to be interpreted through inference and domain knowledge, allowing for the study of ideology to be expanded theoretically, methodologically, and thematically.

C. The Colombian Case

As a proof of concept for the method developed in this study, I examine a legislative setting in Colombia from 2000 to 2024. This case is particularly useful, as it shares features common to many political systems in the Global South: low levels of citizen identification with parties, fluid relationships between elite partisanship and ideological standing, and a proliferation of movements and discourses that do not map neatly onto conventional left–right categories [12], [37].

In such contexts, ideological commitments are not adequately captured by party labels, which often serve as transient vehicles for shifting coalitions. Political actors regularly reconfigure affiliations, while citizens engage with ideological currents that operate independently of formal partisan structures. As a result, party membership is a poor proxy for deeper commitments, and relying on it risks misclassifying or obscuring meaningful ideological patterns [36].

Because ideology itself often shapes party affiliation rather than the reverse, this case provides an ideal setting to test a pipeline designed to decouple ideological analysis from inherited categories. By focusing on discourse, the project seeks to capture ideology in its own right, demonstrating the value of a system-agnostic approach.

III. 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 session. 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 [38]. 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.

IV. 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 [39] abstractions of the original text (Fig. 1). 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 interpreted using the three phases of the computational grounded theory (CGT) framework [33]: 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, representing each speaker as the set of topic narratives present in their intervention, a frequent itemset analysis algorithm was applied to detect frequently co-occurring narratives, interpreting these results as signals of political ideology.

A. BERTopic Model

Topic models are unsupervised statistical learning methods that use word co-occurrence to infer topical categories from a corpus [40]. These have become commonplace in social science research, applied to, for example, legislative proceedings, court rulings, news coverage, and social media content. For the present analysis, the BERTopic topic model [41] was used to extract the main substantive themes of congressional discussion in Colombia. BERTopic overperforms traditional topic modeling approaches, such as Latent Dirichlet Allocation, 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. This first instace of the topic model constituted CGT's pattern identification phase.

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

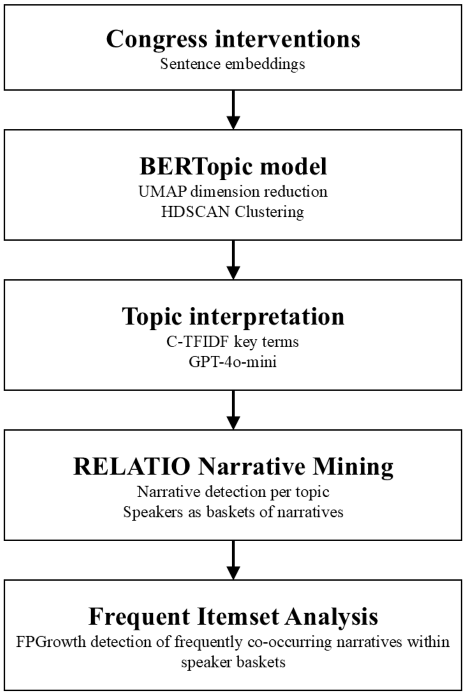

Fig. 1. Methodological framework

paraphrase-multilingual-MiniLM-L12-v2 [42] sentence embedding model, a multilingual BERT model based on siamese BERT networks. For dimensionality reduction, UMAP (Uniform Manifold Approximation and Projection) [43] 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 [44]. 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 [45]. 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 accounting for and encapsulating the semantic variety.

B. Narrative Mining

To extract meaningful narratives from the congressional speeches, the narrative mining model RELATIO [46] 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 [38] 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.

C. frequent itemset analysis

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 analysis algorithm to detect frequently co-occurring topic narratives.

FP-growth [47] 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. 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.

For the analysis, the minimum support threshold was set at 0.5, meaning that an itemset must appear in at least half of the baskets to be considered frequent. The minimal length of the itemsets was set to 3 narratives. 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 various 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 of ideology common enough to be considered a relevant 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.

V. Results

A. Topic Model

The initial BERTopic model produced 168 distinct topics from the semantic content of congressional interventions. Manual inspection revealed that several clusters consisted primarily of procedural speech—for instance, *31_morning_hour_afternoon_session*, centered on session scheduling. While such utterances are inherent to legislative proceedings, they add little to ideological analysis. Their prominence underscores a key challenge of congressional speech: procedural scaffolding often dominates the record, threatening to obscure substantive content. Topic modeling, however, effectively filters this noise and isolates semantically meaningful interventions.

After removing procedural clusters and other noise, the refined dataset contained 83,797 interventions organized into thematically coherent topics. Among these, the largest cluster was *0_colombians_colombian_colombia*, revolving around national identity and competing narratives about the country's political challenges, with recurring references to inequality and conflict as enduring issues. Other prominent topics included *3_senator_president_minister*, encapsulating appeals for collaboration and coalition-building, and *4_budget_budgetary_minister*, which captured budgetary debates often directed at the executive. These themes, frequently invoked and substantively rich, signal the core arenas of legislative discourse.

Interpreted through an understanding of ideology as the narrative abstraction of political action, these results identify the subject domains that structure legislative speech. Political action in this setting is topically framed around diagnosing challenges, building coalitions, and contesting resource allocation. While these findings warrant further exploration on their own, their significance lies in establishing the semantic foundations for subsequent steps in the pipeline, where narrative analysis will connect these topics into broader patterns of ideological signals.

TABLE I
Most frequent topic narratives

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

TABLE II
Most supported topic narratives

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

B. Narrative Mining

The python implementation of the RELATIO narrative mining algorithm initially identified 6,977 unique narratives. These were linked to individual interventions and aggregated into narrative baskets, representing the complete set of narratives used by each speaker.

To improve interpretability and efficiency, narratives were filtered by frequency. Those appearing in more than 80% of baskets—such as the generic "I - Thank - Presidency"—were excluded, as they lacked ideological value. Narratives occurring in fewer than 20% of baskets were also removed, ensuring inclusion only of patterns shared broadly enough (at least 120 speakers) to capture ideologically significant discourse while excluding overly idiosyncratic forms.

After filtering, 303 narratives remained. The reduction was largely driven by narratives that appeared in only a few baskets. The most frequent and supported narratives are presented in table I and table II. Support measures the proportion of baskets containing a given narrative, reflecting its prevalence across speakers.

Many high-frequency narratives aligned with the main topics identified by the BERTopic model, such as topic 0, 3 and 4, confirming the interpretive consistency of the pipeline. These narratives, however, provided more fine-grained ideological insight. For instance, within debates on national challenges, the narratives *0_20: WE-SUPPORT-PEACE* and *0_21: THEY-IMPOSE-ACCORDS* captured opposing stances on peace-building, illustrating the pipeline's capacity to recover politically salient narratives. By identifying frequent and well-supported narratives, this stage highlights the ideological content embedded within topics.

C. Frequently co-occuring topic narrative sets

Finally, understanding ideology as the narrative constructed to rationalize and explain political action, which for legislators can be captured in their political speech, a higher-level representation can be built by identifying recurring sets of narratives and interpreting the unifying logic within the sets. Following, I implemented frequent itemset analysis using FP-growth, yielding 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, two categories of itemsets emerged: within-topic itemsets, coherent ideological narratives around specific substantive domains; and mixed-topic itemsets, bridging narratives across different subject areas. These are interpreted as signaling ideological positions consistent with politically relevant issues, and guiding discursive strategies characteristic of Colombian politics, respectively. See Table III for more detail.

1) Within-topic narrative sets: A first set of emerging topic–narrative itemsets shows how frequently co-occurring narratives within the same substantive domain form coherent ideological stances. These combinations illustrate how legislators construct frameworks of meaning that remain internal to particular debates. Because BERTopic already identified the domains most central to legislative discourse, examining narrative co-occurrences within each topic allows us to see how ideological orientations are articulated and sustained in context. These within-topic itemsets do not merely reflect repetition but structure interpretive logics through which political actors frame challenges, attribute responsibility, and justify responses.

Several broad patterns emerge across the corpus. In discussions around national identity, for instance, narratives such as *we-support-peace*, *reform-solves-deficit*, and *president-fixes-safety* frequently appear together, forming a stance that ties peacebuilding and reform to strong executive leadership. In contrast, narratives like *they-imposed-accords*, *groups-is-terrorists*, and *investment-falls* co-occur to construct an opposing stance that challenges peace processes by linking them to threats against security and economic stability. These clusters illustrate how ideological differentiation emerges not from isolated claims but from structured packages that bind political, institutional, and economic arguments together.

A similar logic appears in fiscal debates, where itemsets combine narratives of accountability—such as *congress-is-accountable* and *minister-fail*—with structural concerns like *pension-need-contribution* and *budget-execute-bad*. The frequent co-occurrence of these narratives suggests that budgetary discourse simultaneously identifies actors to blame for fiscal shortcomings while diagnosing systemic vulnerabilities in the state's economic model. This layered discourse demonstrates how ideological framing operates at both the level of political responsibility and the level of structural critique.

Governance-related itemsets likewise emphasize the moral dimension of political action. Narratives such as *fight-corruption*, *collaboration-is-pleasure*, and *president-implicates-accomplices* regularly appear together, revealing how collaboration is framed as legitimate only when free from corruption or complicity. The recurrent presence of these moralized narratives across ideological camps underscores their role as a shared language of legitimacy through which legislators position themselves and attack opponents.

Taken together, these examples demonstrate how within-topic itemsets crystallize the interpretive structures that underpin ideological differentiation. By surfacing the recurrent ways in which narratives are bundled, the analysis provides a first map of ideological orientations in legislative speech: grounded not in external categories but in the patterns of rationalization articulated by political actors themselves.

2) Between-topic narrative sets: A second set of emerging itemsets draws together narratives from different topics. While within-topic sets capture ideological positions internal to a particular domain, mixed-topic sets reveal how those stances are bridged across issue areas to form broader ideological configurations. It is in these cross-topic connections that ideology, understood here as the narrative interpretation of political action, becomes visible as a system of rationalizations that integrate disparate claims into coherent structures. By linking narratives from peace, governance, and fiscal domains, legislators articulate not just isolated positions but interpretive

TABLE III
MOST SALIENT WITHIN-TOPIC AND MIXED-TOPIC FREQUENT ITEMSETS

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]

frameworks that weave together multiple dimensions of political life.

One salient mixed-topic itemset combines *government-finance-war*, *government-kill-young*, *we-support-peace*, and *president-support-no one*. This set constructs a pro-peace stance that simultaneously acknowledges the state's role in perpetuating violence. Unlike pro-peace itemsets centered on institutional reform, this grouping frames peace as a critical project tied to recognition of state failures. The inclusion of *president-support-no one* reinforces this critical distance, suggesting that peace-supporting actors often reject personalistic alignments, emphasizing structural transformation rather than loyalty to specific leaders.

Another important itemset combines *they-impose-accords*, *we-believe-government*, *we-demand-information*, and *president-support-no one*. Here, opposition to peace accords is linked to an appeal to institutional trust and demands for transparency. While governments may lead peace processes, institutionalism itself becomes contested terrain: critics attack particular administrations but simultaneously use the language of institutional legitimacy to justify their positions. The result is an ideological stance that blends skepticism toward political figures with appeals to broader principles of accountability.

Mixed-topic itemsets also reveal how claims about social justice are constructed. The combination of *budget-execute-bad*, *pension-need-contribution*, *system-boost-rural*, and *we-support-peace* ties demands for peace directly to concerns about fiscal management and rural inequality. This set frames peacebuilding as inseparable from the promise of economic improvement and social inclusion, particularly for rural populations marginalized by conflict. By contrast, an itemset like *citizenship-bet-science*, *pension-need-contribution*, *we-believe-government*, and *minister-fails-government* articulates a more technocratic vision, where social betterment is tied to governance reform and rational administration. Juxtaposing these two clusters highlights how peace-centered and technocratic frameworks represent distinct pathways through which social justice is narratively imagined.

Together, these mixed-topic itemsets show how ideology emerges not from single debates but from the interpretive linkages across domains. By weaving together stances on peace, governance, fiscal policy, and social justice, legislators produce multi-dimensional ideological frameworks that explain and justify political action. These patterns underscore the flexible and strategic nature of ideology, capturing how political actors bridge issues into broader visions of order and change.

VI. 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 adopt a sociological framing of ideology as the rationalization and narrative construction of political action [22]. Finding incompatibility between this definition and traditional representation models—whose limits are clearest in fragmented and fluid political systems—the aim was to propose a method that captures ideology inductively at scale, through the organization and connection of speech rather than the imposition of inherited coordinates. The pipeline operationalized this goal through topic modeling (to detect discourse axes), narrative modeling (to extract stances within topics), and frequent itemset analysis (to detect patterns of co-occurrence between individual narrative sets), enabling the bottom-up reconstruction of ideological structures.

Computational approaches to ideology detection have offered powerful tools for large-scale text analysis but remain constrained by key assumptions. Most rely on party affiliation as a proxy for ideology [11] and assume stable structures inherited from Western bipartite systems [32]. In contexts where partisanship is weak and alignments fluid [12], [37], these approaches misclassify or obscure ideological structures. They risk imposing exogenous categories, neglecting the possibility that dimensions emerge differently across settings.

The method proposed here responds to these limitations: by grounding ideological representations in political speech itself, allowing dimensions to emerge inductively rather than being predefined, and using a modular pipeline of established computational methods reinterpreted through grounded theory [33].

Applied to a large corpus of congressional interventions, the method revealed coherent and context-specific ideological structures. Results show that ideological organization is strongly shaped by attitudes toward peace processes, economic anxieties, and concerns about governance integrity. Three attitudinal groups toward peacebuilding were detected: a pro-peace stance aligned with reform and presidential authority; a pro-peace stance critical of government responsibility for violence and mismanagement; and an anti-peace stance portraying accords as externally imposed and threatening both security and economic stability. These patterns cannot be adequately mapped onto left–right models or dimensional schemes; they are specific to context, emerging directly from the text. Effectively demonstrating the unique effectiveness of the pipeline.

The bridging of narratives across topics, as shown in mixed-topic frequent itemsets, illustrates most clearly the definition of ideology as the rationalization of political action through connected narratives. Actors link stances across domains—such as peacebuilding, fiscal governance, and institutional trust—into integrated frameworks. For example, narrative sets connecting rural development with critiques of budget execution and support for peace articulate a vision of post-conflict reform, while others link skepticism toward peace accords with calls for transparency, weaving opposition into procedural appeals. These connections demonstrate how disparate concerns are narratively rationalized into ideological systems.

More broadly, these results show that political stances, once systematically linked through grounded narrative structures, can serve as culturally validated proxies for ideology. This expands the methodological toolkit by providing a grounded basis for supervised approaches such as stance detection or classification without external labels. It also opens possibilities for culturally anchored categories adaptable to new contexts, facilitating comparative research. Because the pipeline draws on modular tools like BERTopic, RELATIO, and FP-growth, it can integrate newer techniques as they emerge, ensuring scalability and long-term relevance.

By focusing on recurring narrative combinations rather than static dimensions, the method captures ideological "incongruence," where seemingly contradictory stances coexist. A speaker may, for example, support rural reform while emphasizing strict security measures. Traditional supervised models could easily misinterpret such patterns, but this approach preserves the complexity of ideological articulation. It also enables new substantive questions to emerge—for instance, the finding that economic uncertainty is attributed both to the executive and the legislature, a nuance traditional models would likely miss. Capturing such dynamics offers a richer understanding of how discourse builds up ideology.

Some limitations remain. Congressional speech is a valuable window into elite discourse but cannot be assumed to represent broader society, constraining generalization. The application of thresholds, while necessary for interpretability, excludes low-frequency narratives that may still carry substantive importance. Exploring alternative measures for extracting itemsets could strengthen robustness.

Future work could extend the method to citizen discourse, petitions, or digital platforms, where it may reveal different formations. Comparative applications are also possible: cautiously standardizing emergent narrative structures could allow between-system comparisons. Additionally, consolidating the pipeline into a unified software package would further improve accessibility and encourage broader use.

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

VII. 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 related to the composition of a political system. 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 making.

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.

References

[1] N. Bobbio, *Left and Right: The Significance of a Political Distinction*. University of Chicago Press, 1996.

[2] P. C. Bauer, P. Barberá, K. Ackermann, and A. Venetz, "Is the Left-Right Scale a Valid Measure of Ideology?" *Political Behavior*, vol. 39, no. 3, pp. 553–583, Sep. 2017.

[3] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar, "Beyond Binary Labels: Political Ideology Prediction of Twitter Users," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 729–740.

[4] M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik, "Political Ideology Detection Using Recursive Neural Networks," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Toutanova and H. Wu, Eds. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 1113–1122.

[5] A. Budhwar, T. Kuboi, A. Dekhtyar, and F. Khosmood, "Predicting the Vote Using Legislative Speech," in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. Delft The Netherlands: ACM, May 2018, pp. 1–10.

[6] B. J. Dietrich, R. D. Enos, and M. Sen, "Emotional Arousal Predicts Voting on the U.S. Supreme Court," *Political Analysis*, vol. 27, no. 2, pp. 237–243, Apr. 2019.

[7] L. K. Nelson, "Cycles of Conflict, a Century of Continuity: The Impact of Persistent Place-Based Political Logics on Social Movement Strategy," *American Journal of Sociology*, vol. 127, no. 1, pp. 1–59, Jul. 2021.

[8] A. T. J. Barron, J. Huang, R. L. Spang, and S. DeDeo, "Individuals, institutions, and innovation in the debates of the French Revolution," *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, pp. 4607–4612, May 2018.

[9] P. DiMaggio, M. Nag, and D. Blei, "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding," *Poetics*, vol. 41, no. 6, pp. 570–606, Dec. 2013.

[10] B. Bonikowski, Y. Luo, and O. Stuhler, "Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models," *Sociological Methods & Research*, vol. 51, no. 4, pp. 1721–1787, Nov. 2022.

[11] L. Rheault and C. Cochrane, "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora," *Political Analysis*, vol. 28, no. 1, pp. 112–133, Jan. 2020.

[12] N. Lupu, *Party Brands in Crisis: Partisanship, Brand Dilution, and the Breakdown of Political Parties in Latin America*. Cambridge University Press, 2016.

[13] C. Geertz, "Ideology as a Cultural System," in *Ideology*. Routledge, 1994.

[14] A. Downs, "An Economic Theory of Political Action in a Democracy," *Journal of Political Economy*, vol. 65, no. 2, pp. 135–150, Apr. 1957.

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

[16] L. M. Imbeau, F. Pétry, and M. Lamari, "Left-right party ideology and government policies: A meta-analysis," *European Journal of Political Research*, vol. 40, no. 1, pp. 1–29, Aug. 2001.

[17] M. Jankowski, S. H. Schneider, and M. Tepe, "How stable are 'left' and 'right'? A morphological analysis using open-ended survey responses of parliamentary candidates," *Party Politics*, vol. 29, no. 1, pp. 26–39, Jan. 2023.

[18] H. J. Eysenck, *The Psychology of Politics*. New York: Routledge, 1954.

[19] L. L. Thurstone, "The measurement of social attitudes," *The Journal of Abnormal and Social Psychology*, vol. 26, no. 3, pp. 249–269, 1931.

[20] S. Vaisey, "Motivation and Justification: A Dual-Process Model of Culture in Action1," *American Journal of Sociology*, May 2009.

[21] R. C. Tucker, *The Marx-Engels Reader*. Norton New York, 1978.

[22] J. L. Martin, "What is ideology?" *Sociologia, Problemas e Práticas*, no. 77, 2015.

[23] H. Arendt, *The Human Condition: Second Edition*, M. Canovan and a. N. F. b. D. Allen, Eds. Chicago, IL: University of Chicago Press, 1958.

[24] J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, vol. 21, no. 3, pp. 267–297, Jul. 2013.

[25] I. Tavory and S. Timmermans, "Abductive analysis and grounded theory," *The SAGE handbook of current developments in grounded theory*, pp. 532–546, 2019.

[26] J. W. Mohr, C. A. Bail, M. Frye, J. C. Lena, O. Lizardo, T. E. McDonnel, A. Mische, I. Tavory, and F. F. Wherry, *Measuring Culture*. Columbia University Press, 2020.

[27] T. König, M. Marbach, and M. Osnabrügge, "Estimating Party Positions across Countries and Time—A Dynamic Latent Variable Model for Manifesto Data," *Political Analysis*, vol. 21, no. 4, pp. 468–491, Jan. 2017.

[28] B. Bonikowski and L. K. Nelson, "From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research," *Sociological Methods & Research*, vol. 51, no. 4, pp. 1469–1483, Nov. 2022.

[29] J. Fuhse, O. Stuhler, J. Riebling, and J. L. Martin, "Relating social and symbolic relations in quantitative text analysis. A study of parliamentary discourse in the Weimar Republic," *Poetics*, vol. 78, p. 101363, Feb. 2020.

[30] B. E. Lauderdale and A. Herzog, "Measuring Political Positions from Legislative Speech," *Political Analysis*, vol. 24, no. 3, pp. 374–394, Jul. 2016.

[31] H. Borja-Orozco, "Deslegitimación del adversario y orientación ideológica: análisis de publicaciones de dos líderes políticos colombianos en Twitter," *Acta Colombiana de Psicología*, vol. 27, no. 1, pp. 17–36, Feb. 2024.

[32] R. Németh, "A scoping review on the use of natural language processing in research on political polarization: Trends and research prospects," *Journal of Computational Social Science*, vol. 6, no. 1, pp. 289–313, Apr. 2023.

[33] L. K. Nelson, "Computational Grounded Theory: A Methodological Framework," *Sociological Methods & Research*, vol. 49, no. 1, pp. 3–42, Feb. 2020.

[34] C. M. Federico and A. Malka, "The Psychological and Social Foundations of Ideological Belief Systems," in *The Oxford Handbook of Political Psychology*, L. Huddy, D. O. Sears, J. S. Levy, and J. Jerit, Eds. Oxford University Press, Sep. 2023, p. 0.

[35] P. A. Baisotti and F. Lagos Rojas, Eds., *Ideology, Post-Ideology and Anti-Ideology in Latin America: Reflections from the Last Decade*. London ; New York, NY: Bloomsbury Academic, 2025.

[36] S. Mainwaring, *Party Systems in Latin America*. Cambridge University Press, Feb. 2018.

[37] C. Meléndez, "The Post-Partisans: Anti-Partisans, Anti-Establishment Identifiers, and Apartisans in Latin America," *Elements in Politics and Society in Latin America*, Aug. 2022.

[38] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.

[39] B. Latour, *Pandora's Hope: Essays on the Reality of Science Studies*. Harvard University Press, Jun. 1999.

[40] J. Grimmer, M. E. Roberts, and B. M. Stewart, *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press, Jan. 2022.

[41] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022.

[42] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019.

[43] L. McInnes and J. Healy, "Accelerated Hierarchical Density Clustering," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 33–42.

[44] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Berlin, Heidelberg: Springer, 2013, pp. 160–172.

[45] OpenAI, "GPT-4o-mini," 2025.

[46] E. Ash, G. Gauthier, and P. Widmer, "Relatio: Text Semantics Capture Political and Economic Narratives," *Political Analysis*, vol. 32, no. 1, pp. 115–132, Jan. 2024.

[47] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *SIGMOD Rec.*, vol. 29, no. 2, pp. 1–12, May 2000.