A Comparative Framework for Diachronic Lexical Semantic Change in Brazilian Portuguese Political Discourse

João Victor M. Correa[1] **, Oto Araújo Vale**[2] **, Renato Moraes Silva**[1]

[1]Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil

[2]Federal University of São Carlos (UFSCar), São Carlos, São Paulo, Brazil

victormcorrea@usp.br, otovale@ufscar.br, renatoms@icmc.usp.br

Abstract. *Shifts in political discourse often reflect societal changes, but quantifying how lexical usage and semantic representations vary over time remains a challenge for computational linguistics. To compare approaches under different computational constraints, we propose a cost-aware, model- and language-agnostic framework to compare lexical and semantic change signals. We apply it to Brazilian Portuguese political discourse using a shared panel of lemmas combining drift candidates, stable controls, and theory-driven seeds. We compare three families of diachronic drift signals: a TF-IDF lexical-salience baseline, aligned slice-specific Word2Vec models, and a contextual BERT approach. Results indicate disagreement between cheaper baselines, while contextual BERT remains weakly aligned with each method. A filtered contextual panel reduces stable-control leakage and highlights drift candidates such as "bloqueio" and "salário". Rather than treating any detector as ground truth, this study presents an exploratory comparison of complementary drift sensitivities, interpretability, and computational cost in Brazilian Portuguese political discourse.*

1. Introduction

Political language carries historical change in its lexical choices, as shifts in ideological agendas, coalition structures, and policy framing leave traces in word usage over time. Analyzing these large-scale patterns by hand is impractical. Diachronic natural language processing (NLP) methods can capture some of the movement at scale, but different modeling choices foreground different aspects of change [Hamilton et al. 2016, Kutuzov et al. 2018, Tahmasebi et al. 2021]. In political corpora the picture is especially complex, since a term may drift because of a new policy referent, a rhetorical reframing, a shift in collocational profile, or a reorganization of discourse coalitions. No single mechanism accounts for all of them.

This complexity carries over to evaluation, where shared-task benchmarks have improved comparability in lexical semantic change detection [Schlechtweg et al. 2020], but curated external labels remain scarce, especially outside the few languages and corpora covered by standard tasks. Moreover, distributional pipelines can be affected by temporal artifacts, alignment noise, and other confounds when outputs are read too literally [Dubossarsky et al. 2019]. Without externally validated ground truth, an important issue is what different detectors capture and where their rankings agree or diverge on the same political timeline. This question is timely for Brazilian Portuguese because, despite the recent availability of large-scale political corpora such as BrPoliCorpus [Lima-Lopes 2025], comparative studies benchmarking multiple drift detection families on the same Portuguese political corpus remain scarce.

To address this gap, we propose a cost-aware comparative framework for diachronic lexical and semantic change analysis. The framework stages detection methods into lexical-statistical, static-embedding, and contextual tiers, then compares their outputs on a shared

panel rather than letting each method define its own evidence base. We instantiate it on the floor subset of BrPoliCorpus and evaluate TF-IDF, aligned Word2Vec, and BERTimbau Large on the same set of lemmas. The paper addresses the following research questions:

1. How strongly do the drift detection methods agree on the same lemmas?
2. Which lemmas identified as drift candidates are method-specific?
3. Does the higher-cost contextual stage of the proposed framework add enough interpretive value to justify its computational expense?

2. Related Work

Distributional semantic change detection evolved from co-occurrence comparisons to aligned static embeddings and then to contextual representations [Kutuzov et al. 2018, Tahmasebi et al. 2021]. Among static approaches, slice-specific Word2Vec models aligned with orthogonal Procrustes can recover known historical shifts and reveal statistical regularities of change [Hamilton et al. 2016]. Such embeddings remain attractive because they are efficient and produce interpretable neighborhood movements, but compress all word usages into a single vector, blurring cases where older and newer meanings coexist. Moreover, evaluation remains a central challenge. Standardized shared tasks have improved comparability [Schlechtweg et al. 2020], yet many real-world studies still operate without external semantic ground truth, and apparent change can emerge from temporal referencing, alignment noise, or genre shifts rather than from semantic drift alone [Dubossarsky et al. 2019]. These concerns motivate the stable controls and agreement analysis proposed in this study.

Contextual models provide a complementary perspective. BERT-style representations encode token-level usages and can surface variation that static spaces smooth over [Giulianelli et al. 2020, Devlin et al. 2019]. Scalable contextual methods that improve interpretability have also been proposed [Montariol et al. 2021], and pre-trained Portuguese checkpoints such as BERTimbau [Souza et al. 2020] have made contextual analysis feasible for Portuguese NLP.

Diachronic analysis of political language has gained traction through parliamentary corpora. The study of [Hofmann et al. 2020] compared lexical usage in Austrian political discourse across time, but comparable studies of Brazilian Portuguese political speech are still scarce. Available Portuguese diachronic resources have focused primarily on historical and literary material rather than political discourse [Osório and Cardoso 2025]. BrPoliCorpus [Lima-Lopes 2025] now provides the scale needed for controlled comparative work in this domain. This literature motivates a comparison in which no single detector is treated as the ground truth. Our framework makes the comparison explicit because cheaper methods screen broadly, contextual modeling is used selectively, and agreement diagnostics show where the resulting rankings converge or diverge.

3. Comparative Drift Framework

We first describe the framework in general terms, independent of any particular corpus or model. Section 4 then instantiates it with BrPoliCorpus, TF-IDF, Word2Vec, and BERTimbau Large. Figure 1 summarizes the overall flow.

The framework assumes a diachronic corpus partitioned into temporal slices, where each slice must contain enough documents and tokens to support stable frequency estimates and distributional model training. A preprocessing step then normalizes surface forms to a shared lemma representation so that all three tiers operate on the same vocabulary. Not every word form is suitable for drift scoring, so the framework applies three joint filters before any method runs: a minimum per-slice frequency to guarantee sufficient token mass for embedding training, a document-dispersion threshold to prevent a single source

from inflating apparent salience, and a minimum slice-presence rate to restrict attention to genuinely diachronic terms rather than episodic ones. Drift and stable-control candidates may be further restricted to dominant part-of-speech classes.

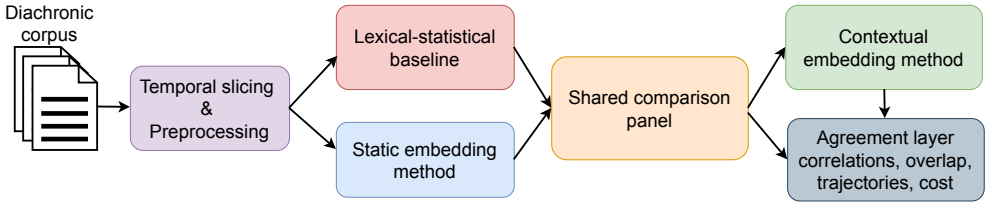

Figure 1. Staged framework for comparing lexical and semantic change signals. Two cheaper tiers screen the vocabulary, whose candidates are combined in a shared panel before the contextual tier and agreement analysis.

The framework then organizes comparative change analysis into three computational tiers arranged by increasing cost. A **lexical-statistical baseline** captures shifts in lexical salience with no model training. A **static-embedding method** aligns one vector space per temporal slice and measures neighborhood displacement. A **contextual-embedding method** represents sampled token usages with a pretrained language model, which can surface usage-level variation that cheaper type-level methods smooth over, but at much higher computational cost.

The key design principle is *staged cost awareness*. The two cheaper tiers first score the full eligible vocabulary, using their own representations to estimate diachronic drift for each lemma. Their top candidates feed into the expensive contextual stage, avoiding costly inference over every eligible lemma. This ordering follows the practical distinction between broad type-based discovery and costlier token-based analysis in lexical semantic change work [Kurtyigit et al. 2021]. Cross-method comparison then requires a fixed evaluation set, so no method is judged only on the terms it would naturally select. The framework therefore prescribes a **shared comparison panel** assembled from four sources: (i) high-drift candidates nominated by the lexical-statistical baseline; (ii) high-drift candidates nominated by the static-embedding method; (iii) low-drift stable controls (lemmas drawn from the low-drift tail of each cheaper tier), which act as negative examples to help detect spurious drift signals rather than only highlighting positive cases [Dubossarsky et al. 2017]; and (iv) theory-driven seed terms drawn from domain expertise or prior work. Because all three tiers score the same candidate set, any disagreement reflects genuine differences in what each method measures rather than divergent vocabularies.

Finally, the framework adds an **agreement analysis layer**, which compares the rankings produced after all tiers score the same shared panel. This layer comprises four complementary analyses. First, pairwise rank correlations measure agreement across the full ranking. Second, top-k overlap curves quantify agreement among high-ranked lemmas as k grows. Third, stable-control leakage rates measure whether stable controls remain low ranked by identifying controls that rise into candidate regions, exposing method-side noise, as emphasized by control-based analyses of semantic change models [Dubossarsky et al. 2017]. Fourth, after leaked controls are removed, the filtered contextual ranking identifies the contextual candidates most suitable for qualitative inspection.

The framework is model- and language-agnostic, allowing any combination of methods respecting the three-tier cost ordering; the next section details our implementation.

4. Methodology

We instantiate the framework on Brazilian Portuguese parliamentary discourse, specifying the corpus, preprocessing choices, and the model used in each tier.

4.1. Corpus and Preprocessing

More diachronic Portuguese resources are now available [Osório and Cardoso 2025], ranging from literary collections such as the Tycho Brahe Parsed Corpus [Galves and Faria 2017] to broad repositories such as the Corpus do Português [Davies and Ferreira 2006]. However, these resources target broad historical periods rather than domain-specific political analysis with continuous yearly slices. BrPoliCorpus [Lima-Lopes 2025] addresses this gap by providing large-scale parliamentary data for Brazilian Portuguese.

We performed experiments on its subset containing Brazilian Chamber floor speeches from 2000 to 2023, organized into 24 yearly slices. This floor subset offers the largest contiguous yearly coverage (428,366 speeches across 24 slices), a uniform parliamentary register, and exact date metadata, making it the most suitable segment for a year-by-year comparative drift study in the political domain. Here, diachronic analysis denotes comparison across time-ordered corpus slices rather than a fixed minimum historical span. Temporal granularity varies with the research question and available data, with prior studies using monthly or yearly slices for socio-cultural change and longer bins for long-term linguistic change [Kutuzov et al. 2018]. Our 24 yearly slices yield 23 adjacent-year transitions and support analyses of annual change in political discourse.

As described in Section 3, not all tokens are eligible for drift scoring. Raw speeches are processed using `spaCy` (`pt_core_news_sm`) [Honnibal et al. 2020]. We retain nouns, verbs, and adjectives, exclude proper nouns, remove stopwords, punctuation, and numeric tokens, and keep only tokens with at least two characters. All text is lowercased while preserving accents. Each token is then mapped to its lemma form, producing the textual representation used by all three tiers of the proposed framework (Figure 1). The experimental environment used spaCy 3.8.11. The parser and named-entity recognizer were disabled, while the model's morphologizer and lemmatizer remained active. A fixed correction list repaired selected lemma errors and malformed Portuguese infinitives. After preprocessing, the corpus retains 428,366 speeches and 63,036,642 tokens across 225,484 unique lemmas. Table 1 summarizes the corpus totals and the quartile structure of yearly variation.

A lemma enters the scored vocabulary only if it appears at least 50 times per slice, in at least 5 documents per slice, and in at least 80% of the 24 slices. These thresholds were fixed as practical corpus-stability defaults rather than tuned to optimize any particular method, and may vary when applied to corpora with different size, genre composition, or temporal density.

The frequency floor ensures sufficient token mass for reliable embedding training in each yearly slice; the document-dispersion requirement prevents a single speech from inflating a lemma's apparent salience; and the slice-presence constraint restricts attention to lemmas that are genuinely diachronic rather than episodic. Drift and stable-control candidates are further restricted to nouns and adjectives, and a small set of procedural high-frequency terms is excluded. These filters reduce the scored vocabulary to 3,221 lemmas for the two cheaper tiers of the framework (Section 3).

Yearly document counts range from 3,041 to 26,868 (median 16,752) and retained token volume from 630K to 4.2M (median 2.6M), which is why the frequency and slice-presence thresholds matter in practice.

4.2. Drift Detection Tiers

With the eligible vocabulary defined, we now describe the well-established model used in each tier of the framework. The lexical-statistical tier uses a TF-IDF lexical-salience baseline, the static-embedding stage uses slice-aligned Word2Vec with Orthogonal Procrustes, and the contextual stage uses BERTimbau Large. These individual techniques are standard

Table 1. Corpus summary for the BrPoliCorpus floor subset. Yearly slice statistics are reported as the 25th percentile (Q1), median, and 75th percentile (Q3) across the 24 yearly files; token and character counts are shown in millions.

Period	Slices	Measure	Corpus total	Yearly slices		
				Q1	Median	Q3
2000–2023	24	Speeches	428,366	14,648	16,752	21,987
		Tokens	63.0M	2.00	2.62	3.45
		Characters	538.5M	16.9	22.3	29.7
		Lemmas	225,484	39,456	50,269	58,637

in the literature [Hamilton et al. 2016, Giulianelli et al. 2020]; the contribution here lies in combining them inside the staged comparison.

For the **lexical-statistical tier**, we treat each yearly slice as one document in the corpus-level TF-IDF computation. Term frequency normalizes a lemma's count by the slice's total token count, while smoothed inverse-document frequency is computed across the 24 slices, so that a lemma's document frequency corresponds to the number of slices in which it appears. For each eligible lemma, we then read its TF-IDF weight from each slice, yielding a 24-entry temporal trajectory. Drift is defined as the mean absolute change in weight between consecutive slices. This provides a continuous ranking of shifts in lexical salience, reflecting how central a term becomes to each year's agenda, without relying on distributional neighborhoods. For lemma w and year t, with $N = 24$ yearly slices, the smoothed weight and change score are $\mathrm{idf}(w) = \log[(1+N)/(1+\mathrm{df}(w))] + 1$, $x_{w,t} = \mathrm{tf}(w,t)\,\mathrm{idf}(w)$, and $s_{\text{TF-IDF}}(w) = \frac{1}{N-1}\sum_{t=2}^{N} |x_{w,t} - x_{w,t-1}|$. The absolute difference prevents increases and decreases from cancelling. Each lemma-year pair has one scalar weight rather than a word vector, so cosine distance would not provide a meaningful comparison for this representation.

For the **static-embedding tier**, we train independent **Word2Vec** (skip-gram) models per yearly slice (300 dimensions, window 5, 5 negative samples, minimum count 5, 5 epochs, and 2 replicates with different seeds) and align them into a common space using Orthogonal Procrustes over the top 20,000 anchor words [Hamilton et al. 2016]. Drift is computed as the mean cosine distance between a lemma's aligned vectors in consecutive slices, averaged over replicates. This captures persistent neighborhood reorganization.

For the **contextual tier**, we embed sampled in-context occurrences using a Portuguese BERT checkpoint (**BERTimbau Large** [Souza et al. 2020]), extracting hidden states from layers -4 (an upper-mid layer) and -1 (the final layer). The choice of these two layers follows systematic studies of contextual lexical-change models, which report that semantic-change signals are often strongest in upper-mid layers, while the final layer is more exposed to surface-form information that can bias change scores [Laicher et al. 2021, Periti and Tahmasebi 2024]. We therefore use layer -4 as a robustness comparison and layer -1 as the primary ranking. For each lemma and slice, up to 64 occurrences are sampled using up to 48 pretokenized words on each side of the target. The contextual drift score is computed as the mean cosine distance between slice-level centroid embeddings across consecutive slices.

4.3. Shared Comparison Panel and Agreement Layer

Following the framework design (Section 3), we build one shared panel before running the contextual stage. The panel contains 55 lemmas: 15 Word2Vec-led drift candidates, 15 TF-IDF-led drift candidates, 20 stable controls, and 5 theory seeds. The values of this composition are fixed a priori as a pragmatic design choice considering three practical

concerns: coverage of each method's top region, enough negative examples for the leakage diagnostics, and BERT inference time over 24 yearly slices. The method-led candidates are the highest-ranked nouns or adjectives from each cheaper tier after the eligibility filters described in Section 4.1. Limiting each tier to 15 items keeps the panel within the informative top region without inflating contextual cost, since each additional lemma requires sampled contexts across all slices. Including 20 stable controls terms ensures that leakage diagnostics have enough negative examples to detect spurious drift signals. The 5 theory seeds are *democracia*, *corrupção*, *reforma*, *economia*, and *liberdade*. We selected them by design as broad political vocabulary anchors, based on domain knowledge rather than a formal selection criterion, and used them as an interpretive reference rather than as an additional set of drift candidates.

All three methods score this same panel. The panel contains 33 nouns and 22 adjectives, which are ranked jointly rather than in separate part-of-speech lists. For display, each method is reranked within the fixed panel and rank r is converted to $p = 1 - (r-1)/(N-1)$ with $N = 55$. Thus, $p = 1$ denotes the highest score and $p = 0$ the lowest. This is an ordinal display scale, not a probability or a comparable effect size. The agreement layer then reports pairwise Spearman rank correlations, top-k overlap curves for $k = 1, \ldots, 15$, stable-control leakage, and a filtered contextual ranking for qualitative inspection. To assess whether methods meaningfully separate drift candidates from stable controls, we also compare their within-panel rank-percentile distributions using one-sided Mann-Whitney tests, where the alternative hypothesis is that drift candidates receive systematically higher ranks.

5. Results

Figure 2 reports the pairwise Spearman rank-correlation diagnostic from the agreement analysis layer. Each point is one of the 55 shared-panel lemmas, and each panel compares how two methods score the same lemma. The four panels compare BERT vs. Word2Vec, BERT vs. TF-IDF, Word2Vec vs. TF-IDF, and BERT layers -4 vs. -1.

The two lighter baselines disagree strongly even after they are constrained to the same candidate universe. As shown in Figure 2, the Word2Vec vs. TF-IDF panel displays a clear negative trend; lemmas ranked highly by one baseline tend to be ranked low by the other. Quantitatively, the two methods reach Spearman $\rho = -0.54$ (p-value $= 2.1 \times 10^{-5}$), showing that they order the shared panel differently.

Figure 2 also shows that the contextual tier partly mediates this split without collapsing onto either baseline. In the BERT vs. Word2Vec and BERT vs. TF-IDF panels, contextual BERT displays weak but positive associations with both methods. On the primary layer (-1), BERT reaches $\rho = 0.21$ (p-value $= 0.128$) with Word2Vec and $\rho = 0.12$ (p-value $= 0.365$) with TF-IDF.

The BERT layer comparison gives a separate robustness check. Layers -4 and -1 agree closely (Spearman $\rho = 0.858$, p-value $< 10^{-15}$), so the contextual conclusions are not driven by a single extraction depth.

Figure 3 reports the remaining diagnostics of the agreement layer. Panel A shows top-k overlap across methods, while Panels B–D use standard boxplots to compare the within-panel rank-percentile distributions of drift-candidate buckets and stable controls for Word2Vec, TF-IDF, and BERT. In the primary BERT layer (-1), the top-15 overlap is 7 terms with Word2Vec and 6 terms with TF-IDF, whereas the baseline-to-baseline overlap remains zero. Thus, contextual BERT recovers candidates favored by both cheaper tiers while preserving an intermediate ranking structure of its own.

The boxplots in Panels B–D further evaluate whether each method separates nominated drift terms from stable controls. Word2Vec and TF-IDF rank drift candidates signifi-

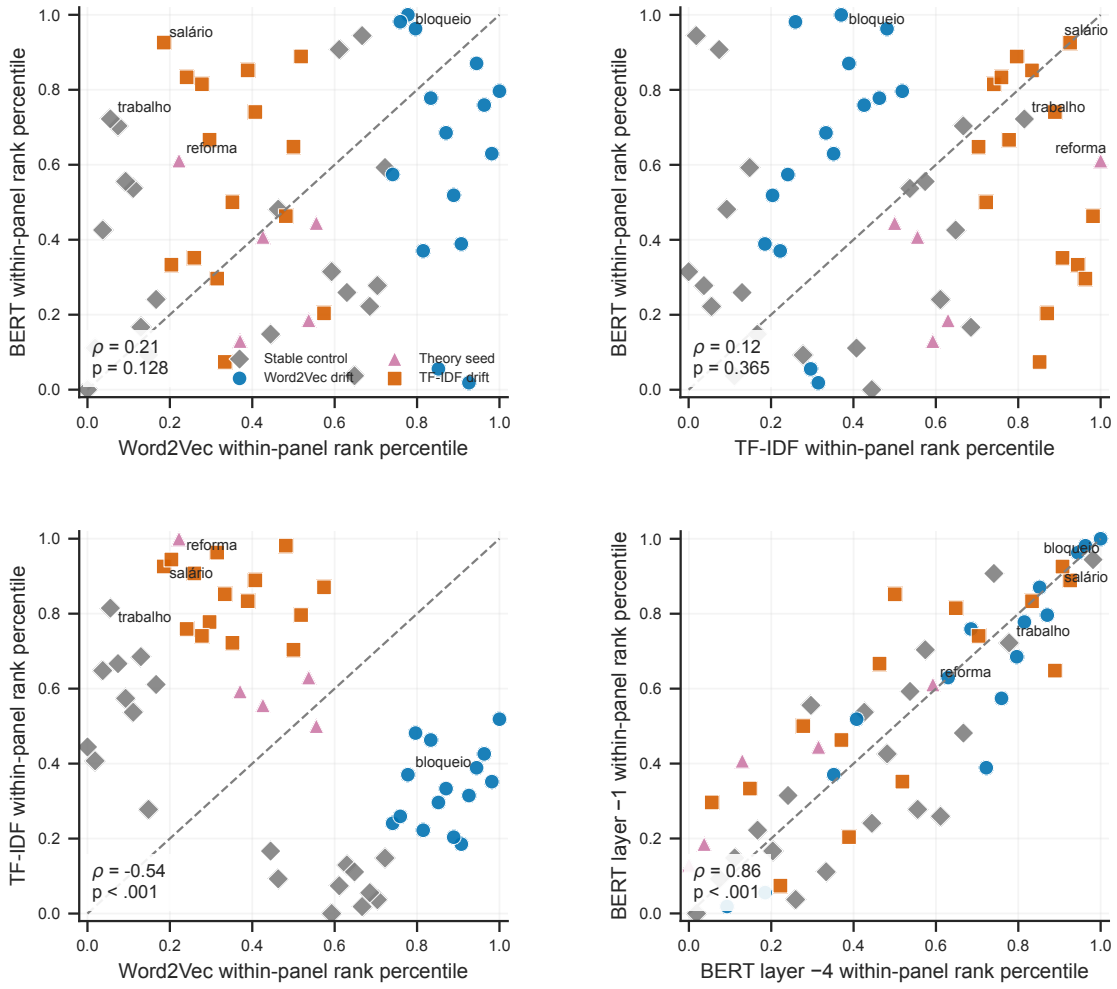

Figure 2. Pairwise within-panel rank agreement from the agreement layer on the 55-lemma shared panel: BERT vs. Word2Vec (top left), BERT vs. TF-IDF (top right), Word2Vec vs. TF-IDF (bottom left), and BERT layers -4 vs. -1 (bottom right). Points are lemmas; annotations report Spearman ρ and two-sided p-values. Within-panel percentiles range from 1 for the highest change score to 0 for the lowest.

cantly above stable controls (p-value < 0.001 in both cases, one-sided Mann–Whitney), showing clean bucket separation. BERT also separates the groups, but with weaker margins (p-value $= 0.008$). For that reason, the contextual ranking is most useful as a selective interpretive layer rather than as an unrestricted final candidate list. After filtering leaked controls, the remaining contextual candidates provide informative arbitration among terms proposed by the cheaper methods.

As a robustness check, we recomputed contextual scores using average pairwise distance (APD) [Giulianelli et al. 2020] instead of the centroid-based metric. On the primary layer, APD and centroid rankings correlate at Spearman $\rho = 0.79$ (p-value $< 10^{-12}$), with 7 of 15 top terms shared. The two metrics foreground different properties. APD promotes high-frequency polysemous terms whose embedding clouds are broad (e.g. *trabalho*, *público*), while the centroid metric favours terms with coherent directional shifts (*bloqueio*, *salário*). Notably, the centroid metric separates method-nominated drift candidates from stable controls more cleanly than APD (p-value $= 0.008$ vs. p-value $=$

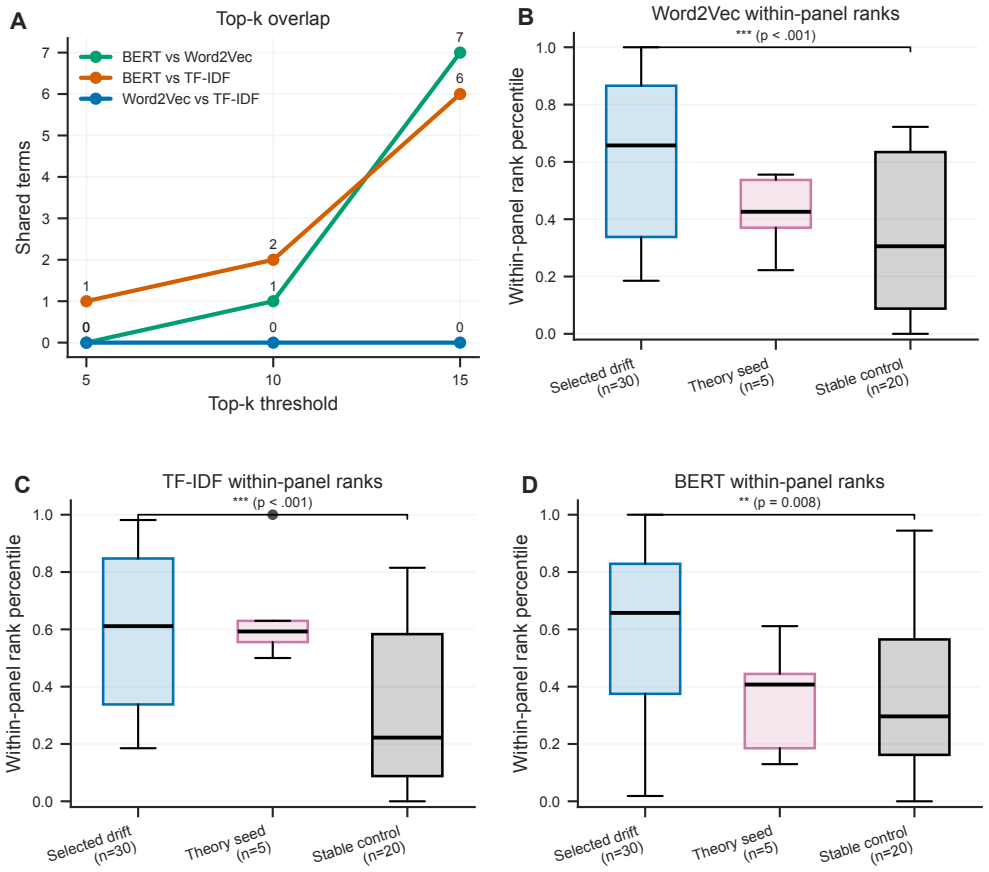

Figure 3. Overlap and rank-distribution summaries on the shared panel. Panel A shows top-k overlap between methods. Panels B–D show boxplots of within-panel rank-percentile distributions by bucket for Word2Vec, TF-IDF, and BERT. Brackets indicate one-sided Mann–Whitney tests for drift terms vs. stable controls.

0.081, Mann–Whitney).

To connect these rankings with observed usage, Table 2 presents paired early and late occurrences for one candidate nominated by each cheaper method and for a stable control. The excerpts support qualitative inspection, but they are not annotated semantic-change labels.

Table 2. Illustrative corpus occurrences for two method-nominated drift candidates and one stable control. BERT ranks refer to the primary layer (-1) within the 55-lemma shared panel. The symbol "…" indicates omitted text.

Term and selection	Early occurrence (year 2000)	Late occurrence (year 2023)
bloqueio Word2Vec-led; BERT 1/55	"… o terrorismo é um apêndice, uma alternativa do bloqueio para golpear a nação cubana …"	"… a falta de água, comida e energia elétrica, e o bloqueio para crianças na Faixa de Gaza."
salário TF–IDF-led; BERT 5/55	"… o Brasil vive, no momento, a grande discussão sobre o salário mínimo."	"… reajustar o valor do salário mínimo acima da inflação; … recuperar o poder de compra …"
trabalho Stable control; BERT 16/55	"… o desemprego oculto pelo trabalho precário …"	"… ilustre amigo que faz muito bem o seu trabalho."

The examples illustrate distinct behaviors captured by the framework and the need to distinguish drift signals from confirmed semantic change. *Bloqueio*, a Word2Vec-led candidate ranked 1/55 by BERT, exhibits a change in referent and geopolitical frame across occurrences, consistent with contextual or neighborhood movement but not sufficient to establish a conventionalized new sense. *Salário*, a TF–IDF-led candidate ranked 5/55 by BERT, remains associated with minimum-wage policy in both excerpts, suggesting increased political salience rather than clear semantic change. Finally, *trabalho*, a stable control ranked 16/55 by BERT, illustrates ordinary contextual variation and why drift scores and selected excerpts should not be interpreted as semantic-change gold labels.

6. Discussion

To examine how the methods differ at the level of individual terms, Figure 4 plots standardized temporal trajectories for four representative lemmas. The selected cases cover four informative situations: a Word2Vec-led term with strong contextual support (*bloqueio*), a TF-IDF-led term with strong contextual support (*salário*), a theory seed whose contextual rank rises despite modest static drift (*reforma*), and a stable-control diagnostic term (*trabalho*) that reveals contextual leakage.

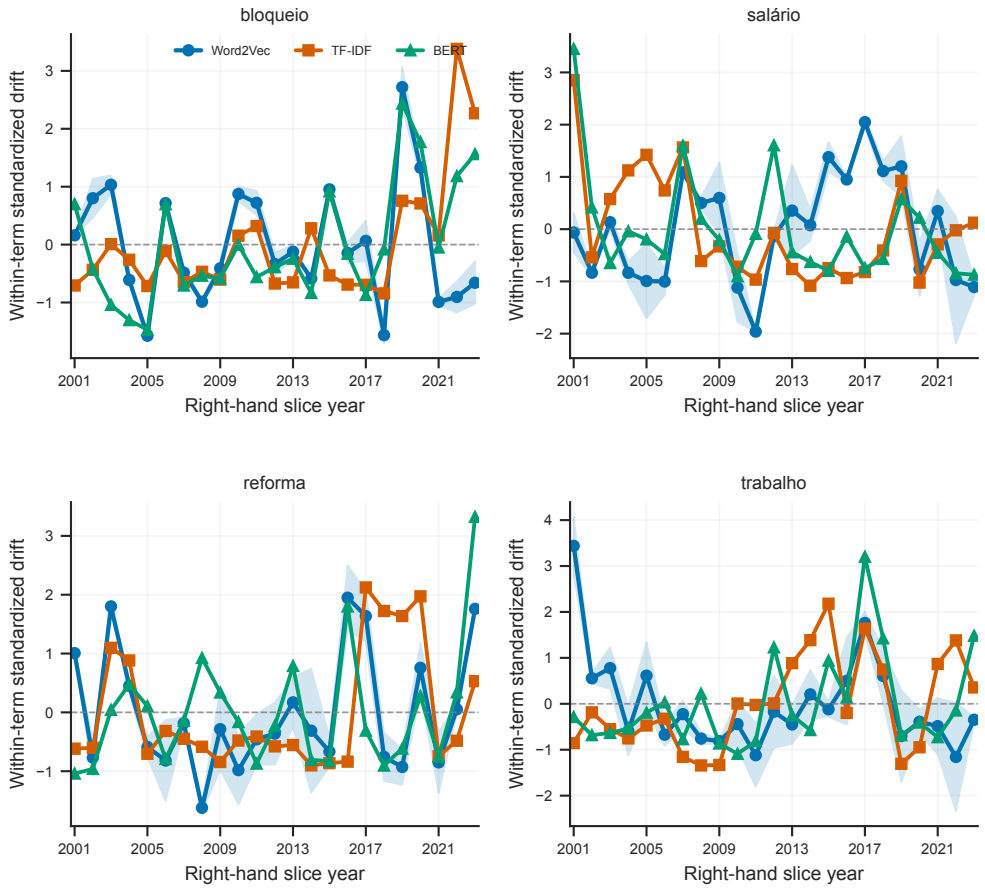

Figure 4. Representative term trajectories for *bloqueio* (top left), *salário* (top right), *reforma* (bottom left), and *trabalho* (bottom right). Scores are standardized within term and method to emphasize temporal shape rather than absolute scale. Word2Vec ribbons denote replicate variability.

The disagreement itself is substantively informative, because each method foregrounds a different facet of lexical change. TF-IDF is most sensitive to lexical-profile

reweightings, so a term whose salience rises in yearly agendas, policy disputes, or formulaic framing can register as drifting even if its local semantic neighborhood remains comparatively stable. Word2Vec, by contrast, captures persistent neighborhood reorganization across slices, highlighting terms whose distributional associations reorganize over time. Contextual BERT occupies an intermediate position and can support either a TF-IDF-led or a Word2Vec-led candidate through its token-level representations, although the stable-control leakage analysis shows that greater context sensitivity does not automatically produce a cleaner unrestricted ranking.

The paired occurrences in Table 2 help constrain the interpretation of these rankings. *Salário* behaves like a lexical-profile case with strong contextual support, but both excerpts retain the minimum-wage sense. Its ranking therefore more directly supports changing political salience than semantic change. In the selected *bloqueio* excerpts, the geopolitical referent shifts from Cuba to Gaza while the core sense of a blockade remains stable. Its ranking is therefore more consistent with contextual and distributional reorganization than with clear sense innovation. *Reforma* illustrates a complementary pattern; it is only modestly ranked by the cheaper tiers, but it is ranked higher by the contextual model, a rank difference that does not establish greater accuracy in detecting semantic change. Finally, the stable-control excerpts for *trabalho* illustrate ordinary polysemy, while its BERT rank shows that context sensitivity can still promote a term expected to remain stable, motivating the leakage diagnostics in the agreement layer.

Table 3. Method scope and computational cost (estimated relative to TF-IDF).

Method	Signal focus	Scope	Relative cost
TF-IDF	Adjacent-slice salience shift	3,221 lemmas	seconds ($1\times$)
Word2Vec	Aligned neighborhood displacement	3,221 lemmas	minutes ($\sim 56\times$)
BERT	Usage divergence from sampled contexts	55 lemmas / 82,461 occurrences	hours ($\sim 600\times$)

Table 3 quantifies the cost gap observed in our single-machine experimental setting. Given the modest rank correlations reported in Section 5, the current evidence does not support replacing the cheaper methods with contextual models. The pattern supports the staged design; cheaper methods screen the vocabulary broadly, while contextual modeling is applied selectively when disagreement is substantively interesting.

7. Conclusion

This paper introduced a shared, cost-aware comparative framework for contrasting three families of lexical and semantic change signals in Brazilian Portuguese political discourse. The research questions posed in the introduction receive three answers from the evidence. First, regarding the strength of inter-method agreement, TF-IDF and Word2Vec disagree strongly on the shared panel and contextual BERT remains only weakly correlated with either baseline. Second, regarding method-specific candidates, the two cheaper methods identify largely non-overlapping sets of drift terms. BERT partially bridges this split by placing candidates from both camps among its top-ranked terms, but it also retains some stable controls in high positions, so its raw ranking requires stable-control filtering before qualitative interpretation.

Third, regarding the cost-benefit of the contextual stage, BERT provides useful adjudication when inter-method disagreement is substantively interesting, as illustrated by terms such as *bloqueio*, *salário*, and *reforma*. At roughly $600\times$ the cost of TF-IDF, though, the current evidence favors using it selectively rather than as a replacement for the cheaper methods. Because the comparative framework specifies roles and interfaces, it can be instantiated with alternative static or contextual embeddings.

8. Limitations

The primary limitation is the lack of external semantic ground truth for the BrPoliCorpus floor subset. Therefore, agreement and disagreement are analyzed as comparative evidence rather than as proof that one detector is correct. A second, methodological limitation is that the raw contextual ranking still admits some stable controls near the top, even though the filtered contextual list is useful for interpretation; contextual sensitivity does not automatically yield a cleaner unrestricted candidate list. The choice of contextual aggregation strategy also merits discussion. We report centroid-based drift as the primary contextual metric and verify it against APD [Giulianelli et al. 2020, Periti and Tahmasebi 2024] (Section 5); the strong correlation and the centroid metric's tighter bucket separation support this choice, though APD captures complementary embedding-cloud variance. Sense-based clustering methods [Montariol et al. 2021] were not attempted here. The TF-IDF tier also treats each annual slice as one aggregate document. Token-normalized term frequency reduces differences in yearly corpus size but does not control speaker composition, verbosity, or topical breadth. Retrieval research on document-length normalization illustrates this broader concern [Lipani et al. 2015]; a speech-level or scope-aware count baseline would use a different document unit and require separate evaluation. Finally, yearly slicing is appropriate for a first comparative baseline but compresses within-year shocks, campaign cycles, and short-lived institutional disputes that may matter for political language.

Future work could address the evaluation gap through external annotation or noisy supervision from labeled parliamentary datasets, though such resources still require careful deduplication before serving as reliable comparison targets. On the modeling side, adopting task-specialized models such as XL-LEXEME [Cassotti et al. 2023], exploring newer divergence metrics [Goworek and Dubossarsky 2026], and applying orthographic-bias mitigation through subword masking [Laicher et al. 2021] would bring the contextual stage closer to current best practices. A symbolic companion analysis using lexical complexity or readability features could also help separate lexical-semantic drift from rhetorical or stylistic change.

9. Ethics Statement

The study analyzes public institutional political speech and reports only aggregate corpus statistics, ranked lemmas, and usage-level comparative signals. We do not profile individual speakers or infer private attributes. Even so, political corpora encode unequal representation and occasionally harmful rhetoric, so drift scores should not be read as normative judgments, evidence of speaker intent, or standalone claims about social meaning. All outputs are exploratory comparative signals whose interpretation depends on corpus composition, preprocessing choices, and the distinction between lexical salience, contextual variation, and broader discourse change.

10. Generative AI Use Disclosure

Generative AI tools were used to suggest alternatives for rewriting some sentences in a concise way and to correct spelling. All suggestions were manually verified by the authors.

11. Code Availability

The source code for the comparative drift framework is available at: `https://github.com/nilc-nlp/brpolicorpus-semantic-drift`.

12. Acknowledgements

We gratefully acknowledge the support provided by the São Paulo Research Foundation (FAPESP; grant No. 2024/17834-6).

References

Cassotti, P., Siciliani, L., DeGemmis, M., Semeraro, G., and Basile, P. (2023). XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Davies, M. and Ferreira, M. J. (2006). Corpus do português. Available at `http://www.corpusdoportugues.org`. Accessed: May 15, 2026.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dubossarsky, H., Grossman, E., and Weinshall, D. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470. Association for Computational Linguistics.

Galves, C. and Faria, P. (2017). Tycho Brahe parsed corpus of historical Portuguese.

Giulianelli, M., Del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Goworek, R. and Dubossarsky, H. (2026). Rethinking metrics for lexical semantic change detection. In Tahmasebi, N., Cassotti, P., Montariol, S., Kutuzov, A., Huebscher, N., Spaziani, E., and Baes, N., editors, *The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26)*, pages 147–161, Rabat, Morocco. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Hofmann, K., Marakasova, A., Baumann, A., Neidhardt, J., and Wissik, T. (2020). Comparing lexical usage in political discourse across diachronic corpora. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 58–65, Marseille, France. European Language Resources Association.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python. DOI: 10.5281/zenodo.1212303.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). Lexical semantic change discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), pages 6971–6981, Online. Association for Computational Linguistics.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuber, J., and Schulte im Walde, S. (2021). Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Lima-Lopes, R. E. d. (2025). BrPoliCorpus: Brazilian political corpus. Repositório de Dados de Pesquisa da Unicamp. DOI: 10.25824/REDU/YCFPIV. Available at: `https://redu.unicamp.br/citation?persistentId=doi:10.25824/redu/YCFPIV`. Accessed: May 15, 2026.

Lipani, A., Lupu, M., Hanbury, A., and Aizawa, A. (2015). Verboseness fission for BM25 document length normalization. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*, pages 385–388, Northampton, Massachusetts, USA. Association for Computing Machinery.

Montariol, S., Martinc, M., and Pivovarova, L. (2021). Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Osório, T. F. and Cardoso, H. L. (2025). Historical Portuguese corpora: a survey. *Language Resources and Evaluation*, 59:1797–1832.

Periti, F. and Tahmasebi, N. (2024). A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., and Shutova, E., editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems – BRACIS 2020*, pages 403–417, Rio Grande, Brazil. Springer.

Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., and Hengchen, S., editors (2021). *Computational Approaches to Semantic Change*. Language Science Press, Berlin.