A Comparative Evaluation of Embedding Models for Retrieval and Clustering of Brazilian Legislative Amendments

João Robson Santos Martins[1,2]

[1] Senado Federal, Brasília, Brazil
joao.robson@senado.leg.br
[2] Universidade de Brasília, Brasília, Brazil
joaorobsonmartins@gmail.com

Abstract. The large volume of amendments submitted in the Brazilian legislative process makes semantic analysis and consolidation labor-intensive. We present a comparative evaluation of 14 distinct base representation models, including BM25L and embedding models spanning open-source multilingual, proprietary, Portuguese-adapted, and legislative-specialized architectures, for retrieving and clustering legislative amendments. Across three corpora totaling 2,462 documents, we assess retrieval, clustering, text segmentation, and thematic granularity. Results show that large-scale open-source multilingual models constitute the best-performing group overall, with *Octen-Embedding-8B* exhibiting the most consistent performance. *gemini-embedding-2* is highly competitive in sentence-similarity mode, while preprocessed BM25L remains a strong baseline, outperforming specialized encoders. Finally, removing structural metadata improves performance, whereas removing justifications degrades global structure; thematic aggregation improves local accuracy at the expense of global consistency.

CCS Concepts: • **Information systems** → **Clustering**; **Document representation**; **Evaluation of retrieval results**; • **Computing methodologies** → *Natural language processing*.

Keywords: Amendments Clustering, Amendments Retrieval, Embeddings, Legislative Process

1. INTRODUCTION

During the processing of legislative proposals in the Brazilian National Congress, hundreds of amendments (formal modifications to the original text, such as additions or deletions) can be submitted for a single bill (e.g., [Presidência da República 2020]). In these scenarios, legislative advisors must manually group semantically similar texts to enable joint analysis and efficient deliberation. To support this task, the Brazilian Federal Senate developed a tool [Senado Federal 2026] based on proprietary embeddings and neighborhood-based clustering, which was initially deployed to process over 2,200 amendments from the 2024 Tax Reform. However, despite the feasibility of this and other recent solutions [Andrade et al. 2025; Pressato et al. 2024], the literature still lacks systematic and comprehensive evaluations of semantic representation models tailored for legislative amendment similarity.

To address these limitations, this article makes three contributions: (i) a multi-task framework integrating retrieval and clustering to evaluate both semantic search and cluster cohesion; (ii) a comparative evaluation of 14 representation models, encompassing BM25L as a lexical baseline and embedding architectures including open-source multilingual, proprietary, Portuguese-specific, and domain-specialized variants; and (iii) an empirical analysis of model scale, linguistic specialization, and text segmentation across three legislative corpora. All source code and artifacts are available on GitHub[1].

[1] https://github.com/joaorobson/br-leg-amendments-embeddings-eval

2. RELATED WORK

In the Brazilian context, [dos Santos et al. 2024] proposes a hybrid system (BM25 and fine-tuned SBERT [Reimers and Gurevych 2019]) for legislative proposal retrieval in the Chamber of Deputies. Due to the 512-token context limit, standalone embeddings underperform BM25, making the hybrid approach superior in terms of recall. Extending this work, [Vitório et al. 2025] compares 12 SBERT variants against lexical approaches (BM25 and BM25L) across three legislative corpora. Except for the fine-tuned model from [dos Santos et al. 2024], lexical methods consistently outperform dense representations, indicating the need for fine-tuning and chunking strategies to mitigate context window limitations. Regarding legislative amendments, [Pressato et al. 2024] evaluates topic retrieval across 269 documents from PEC 6/2019 mapped onto 28 themes, comparing BM25L and SBERT (LegalBERT and BERTimbau). BM25L consistently outperforms dense approaches in Recall, with LegalBERT (a model adapted to the legal domain) showing superior performance to the base BERTimbau. In parallel, [Andrade et al. 2025] investigates the unsupervised clustering of amendments using GPT-4o, varying temperature and text preprocessing techniques. The authors report better performance with low temperatures, noting that results degrade as randomness increases and when preprocessing is applied. They also highlight prompt sensitivity, cost constraints, hallucinations, and absence of a robust ground truth, which restricted their evaluation to proxy metrics of semantic similarity.

Internationally, [Agnoloni et al. 2022] clusters Italian Senate amendments using bag-of-words and hierarchical clustering, mitigating structural ambiguities by normalizing legislative references. Expanding on this scope, [Sajeva et al. 2024] adopts SBERT, achieving gains in Adjusted Mutual Information (AMI) and Adjusted Rand Index (ARI) along with higher semantic consistency, despite reiterating context window constraints. In the Polish context, [Górski 2019] applies TF-IDF, word2vec, and doc2vec to cluster amendments to the Civil Code, demonstrating that while preprocessing optimizes lexical approaches, mean-aggregated word2vec embeddings outperform TF-IDF in cluster coherence.

Despite these advancements, which indicate that lexical approaches remain competitive under proper preprocessing and that embedding models rely heavily on context windows and fine-tuning for consistent performance, existing studies still focus on isolated evaluations (either retrieval or clustering) over limited corpora, particularly within the Brazilian landscape. Consequently, empirical gaps persist regarding the combined impact of model scale, linguistic and domain specialization, and text segmentation strategies on the performance of contemporary embedding models applied to legislative amendment similarity.

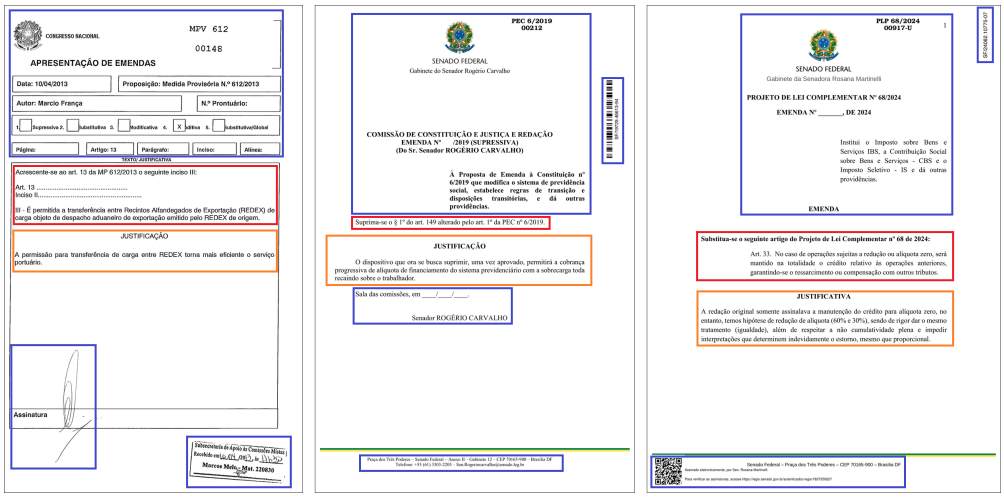

Fig. 1: Examples of amendments (MPV 612/2013, PEC 6/2019, and PLP 68/2024). The bounding boxes delimit the regions considered during segmentation: proposed changes (red), justification (orange), and metadata (blue).

Table I: Statistics for the number of samples per theme

Proposal	Thematic Level	Mean	Median	Std. Dev.	Min	Max	Distinct
MPV 612/2013	Original theme	3.79	1.00	9.82	1	60	58
PEC 6/2019	Original theme	9.24	8.00	7.34	1	28	29
PLP 68/2024	Original theme	7.56	3.00	11.31	1	85	261
	Intermediate theme	19.35	7.00	25.54	1	118	102
	General theme	56.40	17.00	112.43	1	475	35

Table II: Examples of theme hierarchy in PLP 68/2024

Original Theme	Intermediate Theme	General Theme
IBS e CBS (Não Cumulatividade) - Compensação	IBS e CBS (Não Cumulatividade)	IBS e CBS
Bens de Capital - Construção Civil	Bens de Capital	Bens de Capital
Simples Nacional	Simples Nacional	Simples Nacional

3. METHODOLOGY

3.1 Dataset and Text Extraction

The dataset comprises 2,462 parliamentary amendments of varying lengths, distributed across three legislative proposals: 220 documents from MPV 612/2013 (346 ± 382 words), concerning customs areas and payroll tax exemptions; 268 from PEC 6/2019 (499 ± 468 words), regarding the pension reform; and 1,974 from PLP 68/2024 (620 ± 547 words), related to the tax reform. Each amendment consists of a PDF containing the **proposed changes**, the **justification** (or rationale), and institutional **metadata**, such as headers, footers, signatures, and identifiers (Figure 1). The amendments are labeled by theme, presenting a highly imbalanced class distribution (Table I). For PLP 68/2024, the labels follow a three-level hierarchy (specific, intermediate, and general), as shown in Table II. Aggregating the more specific levels reduces the number of themes from 261 to 35 but increases class imbalance, allowing us to measure the impact of thematic granularity on the evaluated tasks.

The PDFs were collected from the Federal Senate's open data API[2]. Out of the total volume, 2,241 documents contained embedded text and were extracted directly, while the remaining 221 (220 from MPV 612/2013 and 1 from PEC 6/2019) were processed via OCR using Tesseract [Smith 2007]. To analyze the impact of structural segmentation within the documents, three versions of each amendment are generated: **raw text** (the full, original PDF content), **proposed change with justification** (raw text excluding metadata), and **proposed change only** (raw text excluding both metadata and justification). For the PEC and PLP corpora, which follow a consistent document structure (Figure 1), segmentation was performed using regular expressions. For the MPV corpus, due to the heterogeneity of amendment formats alongside the inconsistency and noise in the OCR-extracted text, the segmented versions were obtained from manually annotated ROIs using *Label Studio*[3].

3.2 Text Representation

This work adopts two forms of text representation: lexical (BM25L) and semantic (embeddings). Given its competitiveness in the legislative domain (Section 2), the BM25L representations serve as the baseline for this article. They are constructed from the three text versions introduced in Section 3.1 under two distinct configurations: lowercase raw text and preprocessed text. The latter follows the protocol established by [Vitório et al. 2025], which includes punctuation and stopword removal, accent normalization, stemming (Savoy), and the generation of unigrams and bigrams.

For semantic representations, multilingual models (Table III) were ranked via a Borda count over the Clustering (CLT) and Retrieval (RET) tasks of the MMTEB benchmark [Enevoldsen et al. 2025] as of August 22, 2026[7]. Aligning with the official leaderboard [Chung et al. 2025], this criterion unified 159 candidates (all models with available results for both tasks, plus *gemini-embedding-2*[4]) into a single ranking. From this set, the top 8 open-source models from distinct architectural families were selected,

[2]Federal Senate's open data API: `https://legis.senado.leg.br/dadosabertos/api-docs/swagger-ui/index.html`
[3]Label Studio: `https://github.com/HumanSignal/label-studio`

Table III: Selected multilingual embedding models based on the adapted Borda ranking.

Position	Model Identifier	Context	CLT	RET	Borda
1	microsoft/harrier-oss-v1-27b	32k	58.93	78.27	315
2	Qwen/Qwen3-Embedding-8B	32k	57.66	70.88	310
3	codefuse-ai/F2LLM-v2-14B	32k	60.91	66.50	309
4	tencent/KaLM-Embedding-Gemma3-12B-2511	32k	55.77	75.66	308
5	Octen/Octen-Embedding-8B	32k	55.68	71.61	306
8	gemini-embedding-2†	8k	55.30[4]	70.00[4]	302
13	nvidia/llama-embed-nemotron-8b	32k	54.35	68.69	297
15	jinaai/jina-embeddings-v5-text-small	32k	53.41	64.88	287
18	ICT-TIME-and-Querit/ICT-TIME-and-Querit-embedding-v1	32k	53.04	63.99	280
38	text-embedding-3-large†	8k	47.05	59.32	243

Note: Open-source models are available on the Hugging Face Hub at `https://hf.co/<Model-Identifier>`. Proprietary models (†) are accessed via their official APIs [5,6]

thereby mitigating scale-based correlation and filtering out redundant variants of *Qwen3*, *harrier-oss-v1*, *F2LLM-v2*, and *jina-embeddings*. Additionally, two proprietary solutions, *gemini-embedding-2* and *text-embedding-3-large* (ranking 1st and 6th among proprietary models, respectively), were included due to their institutional adoption and relevance within the Federal Senate [Senado Federal 2026]. Furthermore, two specialized open-source models based on BERTimbau [Souza et al. 2020] were selected: **Serafim 335M** [Gomes et al. 2024], from a family of sentence embeddings adapted for Portuguese, evaluating both its Base[8] and Information Retrieval[9] variants; and **BERTimbau-FT**[10], fine-tuned on legislative proposal pairs from the Chamber of Deputies [dos Santos et al. 2024] and identified by [Vitório et al. 2025] as the top-performing SBERT model for document retrieval.

Embeddings were generated from the unprocessed raw text of each version presented in Section 3.1. Two models support task-specific configurations: *jina-embeddings-v5-text-small* was evaluated in *text-matching* (*tm*) and *clustering* (*cls*) modes, while *gemini-embedding-2* was evaluated with its default (*def*), *sentence similarity* (*ss*), and *clustering* (*cls*) configurations[5]. For Serafim and BERTimbau-FT, limited to 128- and 512-token context windows, respectively, we tested truncation (*trunc*) and chunking with mean pooling (*mp*). The remaining models used no text splitting due to their longer context windows; marginal truncation ($< 1\%$ of documents) occurred only for proprietary APIs.

3.3 Experiments and Evaluation

Experiments were conducted independently for each legislative proposal (Section 3.1), using ℓ_2-normalized embeddings ($||\mathbf{x}||_2 = 1$) and BM25L ($k_1 = 1.5$, $b = 0.75$, $\delta = 0.5$). For BM25L, the standard ranking score is used for retrieval, whereas sparse vectors of BM25L weights in the vocabulary space, normalized under the ℓ_2 norm, are employed for clustering. The experimental protocol comprises three steps: evaluating the impact of text segmentation, comparing representation models, and analyzing thematic granularity in PLP 68/2024.

For retrieval evaluation, each amendment serves as a query to rank the remaining ones from the same proposal, discarding single-occurrence themes. This filtering yielded 179 samples for MPV 612/2013, 264 for PEC 6/2019, and 1,966, 1,949, and 1,893 instances for the general, intermediate, and original themes of PLP 68/2024, respectively. Given corpus imbalance (Table I), performance is measured via Precision@1 (P@1), R-Precision, and Mean Average Precision (MAP) [Manning et al. 2008]. Additionally, for PLP 68/2024, we report graded NDCG@all (gains of 3, 2, and 1 for original, intermediate, and general themes; 0 otherwise) to evaluate ranking fidelity across taxonomic levels.

[4]Results reported in [DeepMind 2026]. The model is not yet on the official MMTEB leaderboard, but its inclusion is under discussion in the repository (`https://github.com/embeddings-benchmark/mteb/issues/4260`).
[5]Google Gemini: `https://ai.google.dev/gemini-api/docs/embeddings`
[6]OpenAI: `https://platform.openai.com/docs/guides/embeddings`
[7]`https://huggingface.co/spaces/mteb/leaderboard`
[8]`https://hf.co/PORTULAN/serafim-335m-portuguese-pt-sentence-encoder`
[9]`https://hf.co/PORTULAN/serafim-335m-portuguese-pt-sentence-encoder-ir`
[10]`https://hf.co/josedossantos/bertimbau-tuned`

For clustering, we use scikit-learn's Agglomerative Clustering [Pedregosa et al. 2011] with average linkage and cosine distance. External evaluation relies on Adjusted Rand Index (ARI) [Manning et al. 2008] and V-measure [Rosenberg and Hirschberg 2007], with the number of clusters (K) set to the ground-truth number of themes. Although assuming prior knowledge of K is unfeasible in production environments, it enables a controlled comparison of intrinsic representation quality under identical conditions, mirroring the standard MTEB benchmark protocol [Enevoldsen et al. 2025].

4. RESULTS AND DISCUSSION

Structural segmentation analysis (Table IV) shows that excluding metadata consistently outperforms raw text across all tasks and corpora (one-tailed paired Wilcoxon signed-rank test [Wilcoxon 1945]; 19 retrieval and 21 clustering configurations, $p < 0.05$). Disaggregating gains by model family within each corpus further confirms consistent improvements, except for a marginal 0.1% drop in P@1 for multilingual models on MPV, with MAP gains ranging from +6.9% (BM25L) to +17.2% (specialized encoders) and ARI gains from +13.7% (multilingual models) to +42.2% (BM25L). In contrast, further removing justifications severely degrades global structure (multilingual MAP and ARI fall by 15.9% and 31.0%, respectively), confirming that legislative rationales provide indispensable thematic context and motivating the adoption of the *Changes + Justification* representation for the subsequent analysis.

Based on these textual representations, the results reveal a clear performance landscape across the evaluated model families (Tables V and VII), in which open-weight multilingual models constitute the strongest overall group. Several models consistently rank among the top performers across retrieval and clustering. Among them, *Octen-Embedding-8B* provides the most robust and consistent general performance, attaining the highest MAP on MPV, dominating all retrieval metrics on PEC, and reaching the highest ARI on PLP (0.396). For local top-1 decisions (P@1), however, other architectures achieve isolated peaks, such as *KaLM-Embedding-Gemma3-12B* leading MPV and *F2LLM-v2-14B* on PLP. For clustering, the shifting leadership, with *harrier-oss-v1-27b* and *ICT-TIME-and-Querit-embedding-v1* leading on MPV, and *F2LLM-v2-14B* on PEC, highlights the sensitivity of text representations to sample density and the specific thematic fragmentation of each legislative proposal.

Among proprietary models, *gemini-embedding-2* set for sentence similarity (*ss*) dominates, highlighting the impact of task conditioning. While its default mode (*def*) underperforms, *ss* yields large gains in retrieval (average +43.4% in MAP) and clustering (ARI up to +174.7% in PEC), becoming the best PLP retrieval model and reaching top-tier clustering on PEC and PLP. Interestingly, explicit clustering modes (*cls*) in both *gemini-embedding-2* and *jina-v5-text-small* consistently underperform their similarity modes. This suggests that in our setting, where amendments within a proposal share a broad thematic context, sentence similarity may better capture the finer-grained semantic relationships than explicit clustering. Meanwhile, *text-embedding-3-large* beats the default *gemini-embedding-2* but consistently falls behind its *ss* mode and the leading open multilingual models.

By contrast, preprocessed BM25L remained highly competitive in retrieval, outperforming all

Table IV: Retrieval and Clustering Average Performance across Document Segments (Mean ± Standard Deviation)

		Retrieval			Clustering	
Prop.	**Version**	**P@1**	**R-Precision**	**MAP**	**ARI**	**V-Meas.**
MPV	Raw	0.689 ± 0.132	0.476 ± 0.107	0.508 ± 0.117	0.142 ± 0.055	0.624 ± 0.057
	Chg. + Just.	$\mathbf{0.744 \pm 0.043}^*$	$\mathbf{0.531 \pm 0.071}^*$	$\mathbf{0.567 \pm 0.079}^*$	$\mathbf{0.175 \pm 0.057}^*$	$\mathbf{0.669 \pm 0.042}^*$
	Chg.	0.704 ± 0.071	0.451 ± 0.054	0.488 ± 0.059	0.143 ± 0.055	$\mathbf{0.655 \pm 0.045}^*$
PEC	Raw	0.695 ± 0.115	0.451 ± 0.133	0.490 ± 0.138	0.246 ± 0.120	0.588 ± 0.111
	Chg. + Just.	$\mathbf{0.773 \pm 0.068}^*$	$\mathbf{0.517 \pm 0.110}^*$	$\mathbf{0.559 \pm 0.114}^*$	$\mathbf{0.306 \pm 0.125}^*$	$\mathbf{0.654 \pm 0.093}^*$
	Chg.	0.696 ± 0.069	0.373 ± 0.075	0.404 ± 0.077	0.132 ± 0.037	0.511 ± 0.057
PLP	Raw	0.728 ± 0.096	0.441 ± 0.105	0.461 ± 0.118	0.260 ± 0.113	0.759 ± 0.082
	Chg. + Just.	$\mathbf{0.777 \pm 0.050}^*$	$\mathbf{0.479 \pm 0.088}^*$	$\mathbf{0.501 \pm 0.100}^*$	$\mathbf{0.286 \pm 0.113}^*$	$\mathbf{0.775 \pm 0.078}^*$
	Chg.	0.760 ± 0.056	0.424 ± 0.085	0.438 ± 0.093	0.206 ± 0.120	0.728 ± 0.093

* Statistically significant difference relative to the *Raw* baseline (one-tailed Wilcoxon test, $p < 0.05$).

Table V: Retrieval results using the proposed changes and justifications of the amendments.

Model	MPV P@1	MPV R-P	MPV MAP	PEC P@1	PEC R-P	PEC MAP	PLP P@1	PLP R-P	PLP MAP	PLP NDCG@all
BM25L	0.760	0.516	0.543	**0.739**	0.491	0.523	0.792	0.505	0.531	0.792
BM25L (w/ pre-proces.)	**0.782**	**0.560**	**0.596**	0.735	**0.511**	**0.545**	**0.793**	**0.529**	**0.559**	**0.809**
F2LLM-v2-14B	0.721	0.577	0.624	0.826	0.640	0.684	**0.824**	0.566	0.598	0.826
harrier-oss-v1-27b	0.788	0.615	0.663	0.856	0.653	0.699	0.811	0.569	0.604	**0.830**
ICT-TIME-and-Querit	0.777	0.576	0.624	0.799	0.568	0.617	0.818	0.535	0.561	0.812
jina-v5-text-small (tm)	0.760	0.523	0.551	0.856	0.591	0.642	0.809	0.529	0.556	0.798
KaLM-Gemma3-12B	**0.810**	0.599	0.648	0.795	0.508	0.547	0.788	0.461	0.477	0.773
nemotron-8B	0.777	0.526	0.570	0.788	0.579	0.617	0.809	0.528	0.557	0.803
Octen-Embedding-8B	0.765	**0.625**	**0.670**	**0.867**	**0.659**	**0.713**	0.818	**0.571**	**0.606**	0.827
Qwen3-Embedding-8B	0.765	0.601	0.644	0.852	0.645	0.695	0.816	0.554	0.588	0.819
gemini-embedding-2 (def)	0.721	0.491	0.525	0.731	0.402	0.437	0.741	0.413	0.415	0.746
gemini-embedding-2 (ss)	**0.782**	**0.627**	**0.667**	**0.852**	**0.622**	**0.671**	**0.837**	**0.588**	**0.622**	**0.838**
text-embedding-3-large	0.771	0.583	0.621	0.799	0.576	0.626	0.803	0.520	0.549	0.806
BERTimbau-FT (mp)	0.654	0.454	0.477	0.693	0.443	0.475	0.710	0.386	0.406	0.739
BERTimbau-FT (trunc)	0.676	0.432	0.455	0.712	**0.458**	**0.498**	0.703	0.347	0.358	0.717
Serafim Base (mp)	0.687	0.453	0.478	0.693	0.403	0.442	0.755	0.423	0.439	0.747
Serafim Base (trunc)	0.721	0.437	0.476	0.655	0.301	0.342	0.670	0.311	0.311	0.707
Serafim IR (mp)	**0.726**	**0.461**	**0.483**	**0.739**	0.438	0.479	**0.767**	**0.437**	**0.456**	**0.756**
Serafim IR (trunc)	0.693	0.424	0.449	0.693	0.338	0.377	0.694	0.332	0.326	0.717

domain-specialized encoders and matching or exceeding proprietary APIs such as *text-embedding-3-large* and default *gemini-embedding-2*, consistent with prior findings in the Brazilian legislative domain [Pressato et al. 2024; Vitório et al. 2025]. In clustering, BM25L performed strongly on MPV (ARI: 0.224) and PLP (ARI: 0.363, V-Measure: 0.813), surpassing several compact dense models. However, its lower performance on PEC (ARI: 0.270), which has a more balanced thematic distribution, suggests limitations of sparse representations in inducing semantic partitions under such scenarios.

At the bottom of the performance hierarchy, domain-specialized encoders (*Serafim* and *BERTimbau-FT*) performed substantially worse than all other families across both tasks. Within this group, *Serafim-IR (mp)* consistently outperformed *BERTimbau-FT* in retrieval (e.g., PLP MAP: 0.456 vs. 0.406); sharing the *BERTimbau-Large* backbone, this suggests that functional task alignment is more useful than domain-specific legislative fine-tuning. In clustering, however, neither model dominated, and their gap to modern architectures widened substantially. Although mean pooling (*mp*) mitigated the effects of short context windows for longer documents, such heuristics cannot overcome the architectural constraints of compact encoders [Vitório et al. 2025]. More broadly, while performance scales non-linearly with model size, parameter count is entangled with generational advances in training and native context windows (32k vs. legacy 128–512 tokens). Thus, these findings do not show that scale inherently surpasses domain expertise; rather, they indicate that language and domain adaptation on legacy backbones cannot compensate for limited representational capacity and context when competing with modern billion-parameter architectures or strong lexical baselines such as BM25L.

Finally, thematic granularity in PLP 68/2024 (Table VI) poses a trade-off: aggregating into broader General themes improves local precision ($P@1$) but severely degrades global structure, halving ARI and V-Measure and reducing MAP to 0.376. At the fine-grained Original level, graded NDCG@all further favors dense models, reaching 0.838 with *gemini-embedding-2* (ss) and 0.830 with *harrier-oss-v1-27b* versus 0.809 with BM25L. Therefore, preserving fine-grained themes while leveraging dense semantic representations appears to be the most robust strategy for legislative text consolidation.

Table VI: Retrieval and Clustering Average Performance across Thematic Levels (PLP 68/2024) (Mean ± Std. Dev.)

Level	Retrieval P@1	Retrieval R-Precision	Retrieval MAP	Clustering ARI	Clustering V-Meas.
Original Theme	0.777 ± 0.050	0.479 ± 0.088	0.501 ± 0.100	0.286 ± 0.113	0.775 ± 0.078
Intermediate Theme	$\mathbf{0.868 \pm 0.040}^*$	0.419 ± 0.082	0.435 ± 0.095	0.279 ± 0.140	0.663 ± 0.097
General Theme	$\mathbf{0.903 \pm 0.028}^*$	0.360 ± 0.051	0.376 ± 0.058	0.141 ± 0.062	0.391 ± 0.078

* Statistically significant difference relative to the Original Theme (Wilcoxon test, $p < 0.05$).

Table VII: Clustering results using the proposed changes and justifications of the amendments.

Model	MPV		PEC		PLP	
	ARI	**V-M**	**ARI**	**V-M**	**ARI**	**V-M**
BM25L	0.186	0.689	0.265	0.605	0.349	0.807
BM25L (w/ pre-proces.)	**0.224**	**0.712**	**0.270**	**0.609**	**0.363**	**0.813**
F2LLM-v2-14B	0.208	0.690	**0.529**	**0.787**	0.376	0.828
harrier-oss-v1-27b	**0.278**	0.715	0.452	0.756	0.380	0.827
ICT-TIME-and-Querit	0.255	**0.719**	0.371	0.717	0.357	0.821
jina-v5-text-small (tm)	0.122	0.653	0.344	0.700	0.357	0.814
jina-v5-text-small (cls)	0.096	0.600	0.091	0.475	0.262	0.781
KaLM-Gemma3-12B	0.254	0.709	0.281	0.654	0.293	0.795
nemotron-8B	0.202	0.704	0.361	0.693	0.354	0.818
Octen-Embedding-8B	0.241	0.711	0.416	0.749	**0.396**	**0.833**
Qwen3-Embedding-8B	0.171	0.687	0.467	0.781	0.356	0.820
gemini-embedding-2 (def)	0.174	0.653	0.182	0.569	0.289	0.788
gemini-embedding-2 (cls)	0.144	0.662	0.346	0.676	0.323	0.798
gemini-embedding-2 (ss)	0.180	**0.711**	**0.499**	**0.783**	**0.395**	**0.837**
text-embedding-3-large	**0.221**	0.707	0.349	0.708	0.347	0.819
BERTimbau-FT (mp)	0.113	0.609	0.257	0.626	**0.225**	0.739
BERTimbau-FT (trunc)	0.098	0.603	**0.295**	**0.631**	0.168	0.709
Serafim Base (mp)	**0.142**	0.628	0.168	0.538	0.171	0.739
Serafim Base (trunc)	0.114	**0.629**	0.119	0.516	0.022	0.562
Serafim IR (mp)	0.128	0.637	0.213	0.619	0.184	**0.753**
Serafim IR (trunc)	0.119	0.622	0.159	0.542	0.029	0.569

5. CONCLUSION AND FUTURE WORK

This article presents a comparative evaluation of 14 text representation models for the retrieval and clustering of Brazilian legislative amendments across three legislative proposals. Large-scale open-source multilingual models provide the strongest overall performance, with *Octen-Embedding-8B* showing the most consistent results across corpora and tasks, while *gemini-embedding-2* in sentence-similarity mode achieves competitive or superior retrieval performance. Preprocessed BM25L also remains a strong baseline, outperforming the specialized encoders. Beyond model choice, the results show that removing structural metadata improves performance, whereas removing justifications substantially degrades global structure. Thematic aggregation increases local precision ($P@1$) but reduces MAP, ARI, and V-measure, indicating that preserving fine-grained thematic distinctions is preferable for legislative text consolidation.

This study is limited by the absence of fine-tuning of contemporary large-scale embedding models on legislative data, which could further improve performance in domain-specific settings. Future work will focus on expanding the corpus to include additional proposals and investigating the adaptation of contemporary embedding architectures to the legislative context, particularly instruction-tuned and long-context models. Furthermore, vector space compression and organization techniques, such as Matryoshka Representation Learning (MRL) and UMAP, will be explored to reduce computational overhead and assess representation stability across different embedding dimensionalities.

ACKNOWLEDGMENT

The author would like to thank Alberto Zouvi for providing the themes for MPV 612/2013, and Prof. Nádia Silva and José dos Santos for sharing the article about *HIRS* [dos Santos et al. 2024].

REFERENCES

Agnoloni, T., Marchetti, C., Battistoni, R., and Briotti, G. Clustering Similar Amendments at the Italian Senate. In *Proceedings of the Workshop ParlaCLARIN III at the 13th LREC*. Marseille, France, pp. 39–46, 2022.

Andrade, P., Souza, E., Silva, N., and Carvalho, A. Semantic Clustering in the Context of Legislative Amendments. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*. Fortaleza, Brazil, pp. 569–579, 2025.

Chung, I., Kerboua, I., Kardos, M., Solomatin, R., and Enevoldsen, K. Maintaining MTEB: Towards Long Term Usability and Reproducibility of Embedding Benchmarks. In *Proceedings of the Workshop on Championing Open-source Development in Machine Learning at the 42th ICML*. Vancouver, Canada, 2025.

DeepMind, G. Gemini Embedding 2: A Native Multimodal Embedding Model from Gemini. https://arxiv.org/abs/2605.27295, 2026.

dos Santos, J. A., Souza, E., Bastos Filho, C. J. A., Albuquerque, H. O., Vitório, D., de Lucena, D. C. G., Silva, N., and de Carvalho, A. HIRS: A Hybrid Information Retrieval System for Legislative Documents. In *Proceedings of the 23rd EPIA Conference on Artificial Intelligence, Part I*. Viana do Castelo, Portugal, pp. 320–331, 2024.

Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemiński, D., Winata, G., Sturua, S., Utpala, S., Ciancone, M., Schaeffer, M., Misra, D., Dhakal, S., Rystrøm, J., Solomatin, R., Çağatan, O., Kundu, A., Bernstorff, M., Xiao, S., Sukhlecha, A., Pahwa, B., Poświata, R., GV, K. K., Ashraf, S., Auras, D., Plüster, B., Harries, J., Magne, L., Mohr, I., Zhu, D., Gisserot-Boukhlef, H., Aarsen, T., Kostkan, J., Wojtasik, K., Lee, T., Suppa, M., Zhang, C., Rocca, R., Hamdy, M., Michail, A., Yang, J., Faysse, M., Vatolin, A., Thakur, N., Dey, M., Vasani, D., Chitale, P., Tedeschi, S., Tai, N., Snegirev, A., Hendriksen, M., Günther, M., Xia, M., Shi, W., Lu, X. H., Clive, J., K, G., Anna, M., Wehrli, S., Tikhonova, M., Panchal, H., Abramov, A., Ostendorff, M., Liu, Z., Clematide, S., Miranda, L. J. V., Fenogenova, A., Song, G., Bin Safi, R., Li, W.-D., Borghini, A., Cassano, F., Hansen, L., Hooker, S., Xiao, C., Adlakha, V., Weller, O., Reddy, S., and Muennighoff, N. MMTEB: Massive Multilingual Text Embedding Benchmark. In *International Conference on Learning Representations*. pp. 101715–101771, 2025.

Gomes, L., Branco, A., Silva, J. a., Rodrigues, J. a., and Santos, R. Open Sentence Embeddings for Portuguese with the Serafim PT* Encoders Family. In *Proceedings of the 23rd EPIA Conference on Artificial Intelligence, Part III*. Viana do Castelo, Portugal, pp. 267–279, 2024.

Górski, Towards Legal Change Analysis: Clustering of Polish Civil Code Amendments. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts, co-located with the 17th International Conference on Artificial Intelligence and Law (ICAIL)*. Montreal, Canada, pp. 1–5, 2019.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* vol. 12, pp. 2825–2830, 2011.

Presidência da República. Medida Provisória nº 927, de 22 de março de 2020. `https://www.congressonacional.leg.br/materias/medidas-provisorias/-/mpv/141145`, 2020.

Pressato, D., de Andrade, P. L. C., Junior, F. R., Siqueira, F. A., Souza, E. P. R., da Silva, N. F. F., de Souza Dias, M., and de Leon Ferreira de Carvalho, A. C. P. Natural Language Processing Application in Legislative Activity: a Case Study of Similar Amendments in the Brazilian Senate. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. Santiago de Compostela, Galicia/Spain, pp. 614–619, 2024.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 3982–3992, 2019.

Rosenberg, A. and Hirschberg, J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic, pp. 410–420, 2007.

Sajeva, A., Iannucci, S., Marchetti, C., Merialdo, P., and Torlone, R. Clustering Amendments with Semantic Embeddings. In *Proceedings of the 32nd Symposium of Advanced Database Systems*. Villasimius, Italy, pp. 312–320, 2024.

Senado Federal. Relatório de Gestão 2025. `https://www12.senado.leg.br/transparencia/gestgov/egov/rg2025_final_22042026-1.pdf`, 2026.

Smith, R. An Overview of the Tesseract OCR Engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*. Curitiba, Brazil, pp. 629–633, 2007.

Souza, F., Nogueira, R., and Lotufo, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems, Part I*. Rio Grande, Brazil, pp. 403–417, 2020.

Vitório, D., Souza, E., Dos Santos, J. A., De Carvalho, A. C. P. d. L. F., Oliveira, A. L. I., and F. da Silva, N. F. BM25 x Vila Sésamo: avaliando modelos Sentence-BERT para Recuperação de Informação no cenário legislativo brasileiro. *Linguamática* 17 (1): 17–33, 2025.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin* 1 (6): 80–83, 1945.