Combining Semantic Embeddings and Knowledge Graphs for Identifying Decision Patterns in Brazilian Judicial Decisions

Gustavo Soares Silva[1], **Omar Andres Carmona Cortes**[1,2],
Fábio Manoel França Lobato[1,3], **Antonio Fernando Lavareda Jacob Junior**[1],
[1]State University of Maranhão (UEMA), [2]Federal Institute of Maranhão (IFMA),
[3]Federal University of Western Pará (UFOPA),
Correspondence: antoniojunior@professor.uema.br

Abstract

Approaches based solely on textual representations have limitations in capturing structural relations between legal entities, particularly in documents with high lexical similarity. This paper presents ongoing work on a dynamic clustering system for judicial decisions that integrates hybrid representations, combining semantic embeddings from legal-domain Portuguese models with knowledge graphs automatically constructed from documents. The architecture supports incremental clustering and generates cluster justifications using Large Language Models grounded on knowledge graph relations. Preliminary evaluation combines the quantitative metrics Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

1 Introduction

Judicial decision-making systems increasingly rely on mechanisms that support consistency across large case volumes, especially in contexts of repetitive litigation(Castro and Mendonça, 2024; Oliveira and Nascimento, 2025). In Brazil, the Judiciary accumulates millions of pending cases (CNJ, 2025), motivating the use of computational methods to assist courts in managing precedents and supporting decision uniformity (Polo et al., 2021). The Brazilian Code of Civil Procedure reinforces this demand by establishing binding precedents (Mentzingen et al., 2024).

Existing approaches frequently rely on textual similarity measures or semantic representations (Silva et al., 2021; Costa et al., 2023). Recent studies have explored hybrid representations that integrate semantic embeddings with Knowledge Graphs (KGs) (Tang et al., 2024; Aguiar et al., 2022), yet questions remain about their performance in dynamic environments and their capacity to support legally meaningful explanations. Methods requiring complete corpus reprocessing face scalability limitations, and the use of Large Language Models (LLMs) to generate justifications raises legal validity concerns (Oliveira and Nascimento, 2025).

This paper presents ongoing work on a dynamic clustering system combining semantic embeddings and automatically constructed knowledge graphs. The system supports incremental clustering and structured justifications, addressing three Research Questions regarding (RQ1) hybrid representations, (RQ2) incremental clustering, and (RQ3) legal validity of generated justifications.

2 Related Work

Text mining in the legal domain presents challenges due to specialized vocabulary, complex syntactic structures, and extensive use of normative references (Polo et al., 2021). For Brazilian Portuguese, BERTimbau (Souza et al., 2020) established itself as a reference model, pre-trained on 2.68 billion tokens. Subsequent domain specialization efforts produced models adapted to legal texts, such as LegalBERT-pt (Silveira et al., 2023) and BumbaBERT (do Carmo, 2024).

In the Brazilian legal context, Silva et al. (2021) compared combinations of textual representation techniques (TF-IDF, Word2Vec) with clustering algorithms (K-Means, Agglomerative, Spectral) on 1,515 initial petitions, finding that TF-IDF with PCA and K-Means produced more coherent clusters. Aguiar et al. (2022) expanded the analysis to 16,000 petitions from a state court, integrating HDBSCAN and BERTimbau with a legal KG. Oliveira and Nascimento (2025) used LLMs to interpret clusters generated from 210,000 labor court documents.

The analysis of related work reveals three limitations. First (L1), all analyzed works process static document sets, without addressing continuous incorporation of new documents. Second (L2),

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 181–185

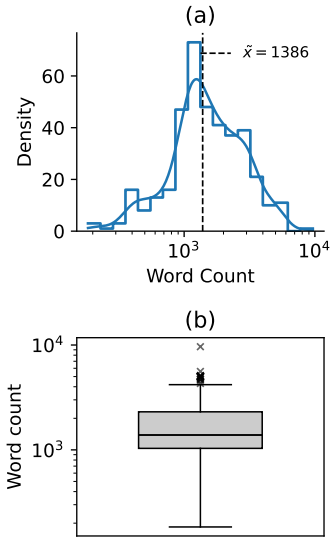

Figure 1: Word count distribution in the corpus: (a) frequency histogram and (b) box plot.

representations capture linguistic patterns but do not explicitly model structural relations such as citations and normative references. Third (L3), when LLMs are used for interpretation, they generate textual descriptions but do not validate the legal consistency of clusters.

This work addresses these gaps by proposing a system that incorporates incremental clustering (L1), hybrid representation combining semantic embeddings and KGs (L2), and LLM-based justification grounded on graph relations (L3).

3 Methodology

This section describes the research methodology following the Data Science Trajectories (DST) framework (Martínez-Plumed et al., 2019), including dataset description, system architecture, and evaluation procedures.

3.1 Business Understanding

The study was conducted in collaboration with the Tribunal de Justiça do Maranhão (TJMA), where decision-pattern identification remains manual and time-consuming. Payroll loan disputes were selected as a pilot domain due to their high volume.

3.2 Data Acquisition and Data Understanding

The corpus comprises 388 judicial decisions from the TJMA (2016–2023) related to payroll loan disputes. Figure 1 shows the word count distribution.

As shown in Figure 1, document length varies substantially, exceeding the token limit of BERT-based models and requiring chunking strategies. Figure 2 compares structural similarity using Jaccard distance over tokens, with semantic similarity using cosine similarity between LegalBERT-pt embeddings. The concentration of document pairs in the upper-right quadrant (structural similarity: 0.86; semantic similarity: 0.94) indicates high lexical and semantic overlap, which may limit the discriminative capacity of clustering approaches based exclusively on textual representations.

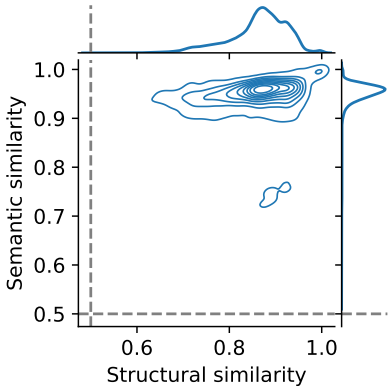

Figure 2: Relationship between structural and semantic similarity across document pairs.

To characterize the corpus's thematic structure, topic modeling was performed using BERTopic with LegalBERT-pt embeddings, UMAP, and HDBSCAN. Figure 3 shows the two-dimensional projection of documents by topic.

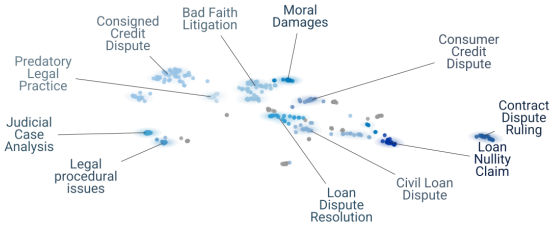

Figure 3: Two-dimensional projection of documents by topics identified via BERTopic.

As shown in Figure 3, the results indicate thematic variation within the corpus, including multiple topics and outliers.

3.3 Data Preparation

Text preprocessing was limited to merging paragraphs fragmented by spurious line breaks and removing excess whitespace; no lowercasing, accent removal, or header stripping was applied, since contextual language models benefit from preserving the original text. Documents were then split into

non-overlapping chunks with a maximum length of 1,500 characters, empirically determined, using recursive character splitting to respect the token limits of extraction models while avoiding duplicate entities in the KG. Chunks and their embeddings are stored in a ChromaDB vector database, which is selected for its local execution capabilities and compliance with data protection requirements. Named entity anonymization is applied after KG extraction to preserve contextual integrity during the extraction stage.

3.4 Modeling

Four embedding models are compared: LegalBERT-pt (Silveira et al., 2023); BERTimbau (Souza et al., 2020); BumbaLM-Embedding, a Qwen3-Embedding-4B variant fine-tuned for the Brazilian legal domain; and a CBOW-based Word2Vec model as a non-contextual baseline. For documents exceeding the token limit of transformer-based models, chunk-level embeddings are averaged to produce a single document vector.

KG construction uses Jurema-7B, a legal-domain LLM for Brazilian Portuguese, hosted on HuggingFace and executed locally, to perform schema-free extraction of entities and relations from each chunk. The resulting graph captures parties, legal provisions, and juridical concepts along with their semantic relations. Graph embeddings are obtained via TransE (Bordes et al., 2013), which models relations as translations in vector space, chosen for its computational efficiency and dimensional compatibility with the textual embeddings.

3.5 Evaluation

Evaluation uses Silhouette, Davies-Bouldin, and Calinski-Harabasz metrics to assess cluster cohesion and separation. The contribution of KG embeddings is assessed by contrasting text-only and hybrid representations across all three metrics. Qualitative evaluation involves legal experts from the TJMA assessing cluster coherence, justification validity, and practical utility through a structured questionnaire.

3.6 Deployment

The system will be delivered as a Docker-containerized microservice that exposes an API that accepts judicial documents and returns cluster assignments with structured justifications, ensuring local execution to ensure data protection compliance.

4 Framework Architecture

The system is organized into five stages. First (a), judicial decisions are segmented into chunks, entities and relations are extracted using an LLM, entity linking resolves coreferences, and the results are incorporated into the KG.

Second (b), semantic embeddings from the domain-specialized model and structural embeddings from TransE are L2-normalized and concatenated. Let $e_t \in \mathbb{R}^n$ denote the textual embedding and $e_g \in \mathbb{R}^m$ the graph embedding. The hybrid representation is defined as

$$e_h = \left[\frac{e_t}{\|e_t\|_2}; \frac{e_g}{\|e_g\|_2} \right] \in \mathbb{R}^{n+m}. \quad (1)$$

Third (c), embeddings are reduced via UMAP with empirically selected target dimensions $d \in \{2, 10, 50\}$, and initial clustering is performed over the reduced vectors using K-Means, HDBSCAN, Agglomerative, and Spectral methods, producing reference clusters $C = \{C_0, \ldots, C_K\}$.

Fourth (d), incremental clustering assigns each new document j to the most similar existing cluster or creates a new one. Let $s_j^{(k)}$ denote the cosine similarity between $e_h^{(j)}$ and the centroid of cluster C_k. Given a threshold θ_s, the document is assigned to C_{k^*} if $\max_k s_j^{(k)} \geq \theta_s$; otherwise, a new cluster C_{K+1} is created. Centroids are updated incrementally without reprocessing the full corpus. The threshold θ_s is initially defined as a user-configurable parameter, allowing legal professionals to control the granularity of the cluster according to operational needs. Data-dependent estimation strategies are planned for future work.

Finally (e), an LLM generates cluster-level justifications grounded in the subgraph of entities and relations shared by the cluster members.

5 Preliminary Results

This section presents preliminary results from experiments with text-only representations, which serve as a baseline for subsequent evaluation of hybrid representations. To obtain a single ranking that balances all three metrics, each configuration is ranked independently by Silhouette Score, Davis-Bouldin Index, and Calinski-Harabasz Index, and the mean of the three positional ranks is computed.

Embedding	Reducer	Algorithm	k	S ↑	DB ↓	CH ↑	Mean Rank ↓
BumbaLM	UMAP(d=2)	HDBSCAN	5	0.752	0.289	8,232.49	37.7
BERTimbau	UMAP(d=2)	HDBSCAN	5	0.745	0.309	3,540.24	94.7
BumbaLM	UMAP(d=10)	HDBSCAN	5	0.723	0.373	3,743.54	123.0
CBOW	UMAP(d=10)	HDBSCAN	5	0.719	0.378	3,212.93	146.3
BumbaLM	UMAP(d=2)	HDBSCAN	10	0.663	0.382	3,984.78	148.3

Table 1: Top-5 clustering configurations by average internal positional rank across Silhouette Score (S), Davies-Bouldin Index (DB), and Calinski-Harabasz Index (CH)

Table 1 presents the top-5 configurations by this aggregated criterion.

The configuration that achieved the highest overall ranking according to the aggregated metric combines BumbaLM embeddings with UMAP ($d = 2$) for dimensionality reduction and HDBSCAN ($k = 5$) for clustering, yielding S = 0.752, DB = 0.289, and CH = 8232.49. BumbaLM appears in three of the five top-ranked configurations, which may indicate that the legal-domain fine-tuning of this embedding model contributes to more discriminative representations in this context.

The presence of CBOW among the top configurations indicates that simpler embedding approaches can still produce competitive results when combined with suitable dimensionality reduction and clustering techniques. Additionally, all top-performing configurations rely on the combination of UMAP and HDBSCAN.

These results indicate that evaluation metrics do not always favor the same configurations, reinforcing the need for qualitative expert assessment and KG-based representations.

6 Preliminary Conclusions

This paper presented ongoing work on a dynamic clustering system that integrates semantic embeddings and KGs, addressing limitations related to static document processing (L1), text-only representations without structural relations (L2), and cluster interpretation without legal validation (L3).

Preliminary results with text-only representations indicate that legal-domain fine-tuned models, particularly BumbaLM, produce more discriminative embeddings for clustering judicial decisions, and that density-based clustering (HDBSCAN) with UMAP consistently outperforms partition-based alternatives. However, disagreements among internal validation metrics highlight that quantitative evaluation alone is insufficient to assess legally meaningful cluster quality, motivating both the integration of KG-based structural features and qualitative expert validation.

Future stages involve implementing and evaluating hybrid representations, incremental clustering, and extending to additional legal domains.

Limitations

This work has limitations inherent to its current stage. Although the proposed pipeline is designed to be domain-agnostic and applicable across legal jurisdictions, empirical validation is currently limited to 388 documents from a single court addressing payroll loan disputes. The high semantic overlap in this corpus makes it a challenging testbed where KG-based differentiation is most needed, but evaluation across additional domains is required to confirm the approach's generalizability. An expanded corpus has been requested from the court.

Also, a few components remain pending: the empirical comparison between text-only and hybrid representations, the assessment of incremental clustering sensitivity to θ_s, and qualitative validation by legal experts.

The schema-free KG extraction relies on Jurema-7B without independent validation of entity and relation quality, and the entity linking step has not been formally evaluated. Similarly, the LLM-produced justifications have not yet been assessed for legal adequacy; while grounding on KG relations is designed to mitigate hallucination, its effectiveness remains an open question.

Acknowledgments

This study was supported by the National Council for Scientific and Technological Development (CNPq) - DT-303031/2023-9; by the Maranhão Foundation for Research and Scientific and Technological Development; the Financing Agency for Studies and Projects (FINEP) – (ProAmazonia - 2373/24 - CTCCA-II); and by Technical Cooperation Agreement Nº. 02/2021 (case Nº. 38328/2020-TJ/MA).

References

André Aguiar, Raquel Silveira, Vasco Furtado, Vládia Pinheiro, and João A. Monteiro Neto. 2022. *Using Topic Modeling in Classification of Brazilian Lawsuits*, page 233–242. Springer International Publishing.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2 of *NIPS'13*, page 2787–2795. Curran Associates Inc.

Marcella Queiroz de Castro and Ana Régia Mendonça. 2024. PLN e segurança jurídica identificação de divergências jurisprudenciais com processamento de linguagem natural. In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 457–462, Belém do Pará, Brazil. Association for Computational Linguistics.

CNJ. 2025. Justiça em números.

José Alfredo F. Costa, Nielsen Castelo D. Dantas, and Esdras Daniel S. A. Silva. 2023. Evaluating text classification in the legal domain using bert embeddings. In *Intelligent Data Engineering and Automated Learning – IDEAL 2023*, pages 51–63, Cham. Springer Nature Switzerland.

Fabrício Almeida do Carmo. 2024. Representações *Embeddings* Orientadas à Linguagem Jurídica Brasileira. Mestrado em engenharia da compuatação e sistemas, Universidade Estadual do Maranhão, São Luís - MA.

Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cesar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana, and Peter Flach. 2019. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3048–3061.

Hugo Mentzingen, Nuno António, Fernando Bacao, and Marcio Cunha. 2024. Textual similarity for legal precedents discovery: Assessing the performance of machine learning techniques in an administrative court. *International Journal of Information Management Data Insights*, 4:100247–100247.

Raphael Souza de Oliveira and Erick Giovani Sperandio Nascimento. 2025. Analysing similarities between legal court documents using natural language processing approaches based on transformers. *PLOS ONE*, 20(4):e0320244.

Felipe Polo, Gabriel Mendonça, Kauê Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Ferreira, Leticia Lima, Antônio Maia, and Renato Vicente. 2021. Legalnlp - natural language processing methods for the brazilian legal language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 763–774, Porto Alegre, RS, Brasil. SBC.

Ingrid L. A. da Silva, Rafael Ferreira Mello, Péricles B. C. Miranda, André C. A. Nascimento, Isabel W. S. Maldonado, and José L. M. Coelho Filho. 2021. Assessment of text clustering approaches for legal documents. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021)*, ENIAC 2021, page 37–48. Sociedade Brasileira de Computação.

Raquel Silveira, Caio Ponte, Vitor Almeida, Vládia Pinheiro, and Vasco Furtado. 2023. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Intelligent Systems*, pages 268–282, Cham. Springer Nature Switzerland.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024. Casegnn: Graph neural networks for legal case retrieval with text-attributed graphs. In *Advances in Information Retrieval*, pages 80–95, Cham. Springer Nature Switzerland.