LexIris-pt and LexBert-pt: Specialized Sentence Embeddings for Legal Similarity in Brazilian Portuguese

Willgnner Ferreira Santos[1], João Gabriel Grandotto Viana[2]
Antônio Pires de Castro Júnior[2], Fernando Ribeiro Trindade[2]
Nádia Félix Felipe da Silva[1]

[1]Institute of Informatics (INF), Federal University of Goiás (UFG), Brazil
[2]Directorate of Artificial Intelligence, Data Science and Statistics (DIACDE),
Court of Justice of the State of Goiás (TJGO), Brazil

willgnner_santos@discente.ufg.br, jggviana202@gmail.com
{apcastro, frtrindade}@tjgo.jus.br, nadia.felix@ufg.br

Abstract

This work presents and evaluates two specialized sentence embedding models for the Portuguese legal domain, LexIris-pt and LexBert-pt, obtained through supervised fine-tuning of BERT-based models using pairs of initial petitions. We propose a comparative evaluation protocol along three fronts: (i) zero-shot inference with pretrained embeddings, (ii) supervised fine-tuning on these pairs, and (iii) vector retrieval with incremental clustering over a corpus of 20,000 initial petitions. The results show that fine-tuning consistently increases correlations with reference scores and improves performance in vector retrieval; additionally, the vector retrieval stage indicates that the metric configured in the index (cosine similarity or inner product) can change the granularity of the partitioning under a fixed threshold, reinforcing the need for joint calibration among the encoder, metric, and threshold. After auditing by specialists from the partner institution, LexIris-pt and LexBert-pt were operationally adopted to support the screening and organization of repetitive claims and predatory litigation.

1 Introduction

The exponential growth of lawsuits in Brazil has posed challenges to procedural management in large courts (Armonas Colombo et al., 2017; Castelliano et al., 2024). In initial complaints, overlap between the description of the facts and the legal arguments is frequent, generating repetitive claims, that is, large volumes of different lawsuits concerning the same type of issue (De Martino et al., 2023; Sauvola et al., 2024) and, in extreme cases, predatory litigation[1] (Page and Soss, 2021). Manual triage of this volume overloads court staff and judges, affecting both timeliness and the consistency of case law (Zadgaonkar and Agrawal, 2021; Singh and Jha, 2024), which reinforces the need for automated solutions that identify legally relevant similarities between documents in large volumes of cases (Nithya et al., 2024). In courts of justice, the Recursive Electronic Search using Natural Language (BERNA) system (Portuguese: *Busca Eletrônica Recursiva usando Linguagem Natural*) already explores automatic similarity checking between facts, legal theses, and precedents (Campos, 2022; Ruiz, 2021; Grace, 2025), but challenges of granularity, robustness, and adaptation to the Portuguese legal domain remain (Zanuz and Rigo, 2022).

Pretrained language models such as Bidirectional Encoder Representations from Transformers (BERT) demonstrate strong performance on multiple Natural Language Processing (NLP) tasks (Devlin et al., 2019), with additional gains when adapted to specific domains, such as the legal (Chalkidis et al., 2020), biomedical (Lee et al., 2019), and scientific domains (Beltagy et al., 2019). In Portuguese, efforts such as BERTimbau (Souza et al., 2020) and resources from the LegalNLP ecosystem (Polo et al., 2021) indicate the potential for specialization (Viegas, 2022), but a large share of studies in legal NLP still focuses on narrow tasks, evaluated in a non-standardized manner, which makes it difficult to estimate the real benefits of domain adaptation and to establish comparisons. In this work, we use similarity between initial petitions as a case study, placing different models and strategies under a comparative experimental protocol.

The main contribution of this work is the training and release of two sentence similarity models specialized in Portuguese legal language, **LexIris-pt** and **LexBert-pt**, derived from models trained and adapted on a corpus of initial petitions. To

[1]Situations in which one of the parties uses the judicial system in an abusive or strategic manner, not to seek a fair solution to a dispute, but to obtain undue advantages, pressure the other party, or overload the Judiciary with repetitive lawsuits lacking solid legal grounds.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 540–550

support this contribution, we propose a common protocol that integrates three complementary axes: **(A1) zero-shot similarity with pretrained embeddings** (Atigh et al., 2025; Schuff et al., 2023), establishing a textual proximity baseline; **(A2) supervised Fine-Tuning (FN) of Portuguese legal embeddings** (Zheng et al., 2021; Mireshghallah et al., 2022), which yields the specialized models LexIris-pt and LexBert-pt; and **(A3) evaluation via vector retrieval and incremental clustering** (Guo et al., 2017), combining Hierarchical Navigable Small World (HNSW) indexing in Milvus (Ribeiro et al., 2025) with a similarity-threshold association rule to group similar pleadings in the insertion stream.

The study is supported by a **novel corpus of 20,000 initial petitions in Portuguese**, provided by a Brazilian public body, processed by a pipeline for cleaning, normalization, and deduplication. In accordance with legal and ethical requirements, the dataset may be made available on demand for academic purposes, subject to institutional authorization and the signing of a confidentiality and responsibility agreement, in compliance with the General Data Protection Law (LGPD) (Lorenzon, 2021). Furthermore, the corpus was anonymized in compliance with CNJ Resolution No. 615/2025 (Conselho Nacional de Justiça (Brasil), 2025).

2 Related Work

Research in legal NLP in Portuguese has advanced, but still faces challenges regarding resources and evaluation (Lima et al., 2022). In the Brazilian context, notable contributions include domain-specific models and data, such as LegalBERT-pt (Silveira et al., 2023) and Portuguese Semantic Textual Similarity (STS) datasets (da Silva Junior et al., 2024), as well as efforts in legal named entity recognition (Zanuz and Rigo, 2022; Duarte et al., 2022).

Several studies evaluate semantic representations and document organization strategies. Silva Junior et al. (2025) compare Term Frequency-Inverse Document Frequency and Best Match 25 to Transformers, showing that simple methods can be competitive, while clustering approaches explore BERT, Generative Pre-trained Transformer 2, and RoBERTa in judicial decisions (de Oliveira and Nascimento, 2021; Liu et al., 2019) and in topic discovery for legal collections (Vianna et al., 2024). In other legal systems, there are STS studies on Indian judgments (Jain et al., 2022; Mandal et al., 2021), proximity between German legal texts and reference laws (Darji et al., 2021), and judgment prediction in English with Transformers (Maqsood et al., 2024).

In the area of sentence embeddings, Sentence-BERT shows how siamese architectures enable cosine-similarity comparison in a zero-shot setting for semantic search and STS (Reimers and Gurevych, 2019). Self-supervised Simple Contrastive Learning of Sentence Embeddings methods refine these representations (Gao et al., 2022), and multilingual families such as E5 further improve performance on retrieval and STS (Wang et al., 2024). Additional gains come from continued pretraining by domain and task (Gururangan et al., 2020) and from legal benchmarks such as LEXGLUE (Chalkidis et al., 2022).

Finally, the organization of large collections of embeddings typically combines approximate nearest neighbors indexes (such as HNSW), nonlinear projections (t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection) and the density-based clustering algorithm Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) for geometric inspection and efficient vector retrieval (Malkov and Yashunin, 2018; Linderman and Steinerberger, 2019; McInnes et al., 2018, 2017).

In summary, despite recent resources and models, there is still a lack of studies that, in the Portuguese legal context, systematically compare pretrained models in zero-shot mode and their FN counterparts, including in vector retrieval tasks over large document collections. This work addresses this gap by training and releasing two specialized sentence embedding models for Portuguese legal language (LexIris-pt and LexBert-pt), evaluated under a unified experimental protocol that includes zero-shot scenarios, supervised FN, and vector retrieval on collections of initial petitions.

3 Corpora

For training the models proposed in this work, we use a corpus composed of 20,000 Portuguese legal documents, provided by a public body of the State of Goiás through the BERNA system. This collection contains first-instance initial petitions. Each record has four textual fields defined from the pairing provided by the system, namely **Paradigm**

Facts (description of the facts in the reference pleading), **Similar Facts** (description of the facts in the most similar pleading), **Paradigm Legal Basis** (legal reasoning in the reference pleading), and **Similar Legal Basis** (legal reasoning in the most similar pleading). In addition to the texts, the collection includes technical internal identifiers of the pairing and numerical proximity scores produced by the system.

Each document is decomposed into these four textual segments for analysis purposes, but, in the modeling stages, fact and law are subsequently reconsolidated into a single representation per petition, so that the corpus remains with 20,000 document-level instances.

The use of the data complies with the LGPD. The texts are processed for academic research purposes, and sensitive information, when present, is not included in the package to be shared. For any on-demand release, the identifiers will be replaced by one-way hash keys, with access granted only after institutional authorization from the public body[2].

3.1 Statistical analysis of the corpus

We performed descriptive analysis after preprocessing. The goal is to characterize the four textual fields and estimate the risk of truncation in BERT-type models. We compute the length in words and characters, number of sentences, and lexical diversity for each field. Lexical diversity is estimated by two measures, one of which is the Type–Token Ratio (TTR),

$$\mathrm{TTR} = \frac{V}{N},$$

where V represents the number of unique types and N the number of tokens; and Herdan's C measure,

$$C = \frac{\log V}{\log N},$$

which is more robust to length variation. A Table 1 summarizes the aggregated results of the corpus and shows a higher average length for the Law blocks compared to the Fact blocks; we observe a higher TTR for Facts and similar Herdan's C values across fields, suggesting a stabilized vocabulary after normalization.

To estimate the impact of the 512-token limit in BERT models, we tokenized each text with the WordPiece tokenizer from BERTimbau Base without truncation and counted its length in tokens. We then computed the fraction with length > 512 for each field, obtaining (61.50%) for Paradigm Facts, (61.94%) for Similar Facts, (79.18%) for Paradigm Legal Basis, and (79.55%) for Similar Legal Basis. Since BERT processes at most 512 tokens per input, in this work we process the texts with the 512-token limit (tokenizer truncation); we do not apply windowing nor extended-context models.

For geometric inspection, we generated sentence embeddings (BERTimbau Base, cosine metric) and produced t-SNE projections with simple random sampling by field (fixed seed). The projections reveal overlap between the Paradigm and Similar views of the same textual type (for example, Paradigm Facts versus Similar Facts and Paradigm Legal Basis versus Similar Legal Basis) and only partial separation between the Facts and Legal Basis blocks (Figure 1), which motivates supervised FN and the comparison of similarity functions.

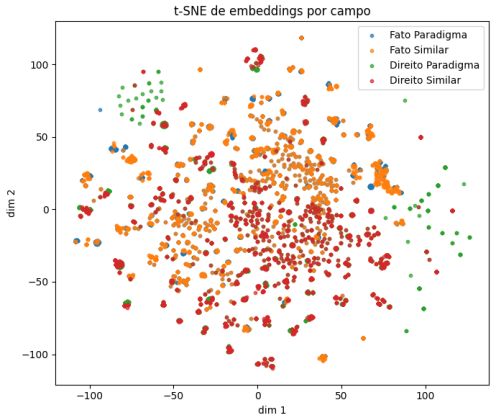

Figure 1: t-SNE projection of embeddings with cosine metric. Overlap is observed between Paradigm and Similar views and partial separation between Facts and Legal Basis.

In the experimental design, the 20,000 documents were partitioned into two complementary blocks. 10,000 constitute the material for supervised learning, used to construct pairs and in the $80/16/4$ (train/validation/test, seed 42) split employed for FN; the remaining 10,000 form a hold-out set dedicated to indexing and vector retrieval over large document volumes. The (A1) zero-shot similarity approach uses 500 pairs drawn exclusively from this hold-out block, without weight updates, whereas (A3) operates over the entire indexed hold-out set.

[2]Legal basis. Processing for studies by a research body, with anonymization whenever possible, pursuant to art. 7º, IV, art. 11, II, c, and art. 13 of Law No. 13.709/2018.

Text field	Length					Lexical diversity	
	Mean words	Standard deviation	Mean characters	Standard deviation characters	Mean sentences	TTR	Herdan C
Paradigm Facts	545.80	870.68	4340.68	6618.89	47.37	0.49	0.93
Similar Facts	543.40	856.55	4326.24	6556.96	47.07	0.49	0.93
Paradigm Legal Basis	1531.65	2295.55	12066.80	15928.19	162.14	0.40	0.91
Similar Legal Basis	1528.98	2291.16	12049.64	15898.56	161.70	0.40	0.91

Table 1: Statistical summary of the textual fields (after preprocessing and stopword removal).

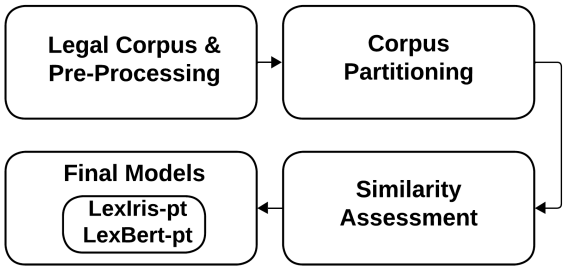

Figure 2: Overall architecture of the methodology, from corpus preprocessing to similarity evaluation and selection of the specialized LexIris-pt and LexBert-pt models in the partner institution environment.

Item	Configuration
Batch size	4 per device in training and validation
Epochs	4
Loss function	CosineSimilarityLoss
Optimizer	AdamW (trainer default)
Learning rate	1×10^{-5}
Warmup	10% of total steps
Numerical precision	FP16 enabled
Evaluation strategy	step-based; evaluation every 100 steps with SequentialEvaluator
Saving and logging	saving every 100 steps; at most 2 checkpoints; logging every 100 steps
Checkpointing	enabled, with state recovery
Hardware	GPU H100

Table 2: Hyperparameters and configurations used in supervised FN.

4 Methodology

This section describes the study methodology. Figure 2 summarizes the overall architecture of the proposed pipeline.

4.1 Preprocessing and data normalization

Preprocessing reduced noise and standardized the textual input, including the removal of stopwords, while preserving numbers and punctuation. We replaced line breaks with spaces, removed punctuation only at the beginning of the sequence, trimmed extra spaces, and limited repeated symbols to four occurrences. Next, we deduplicated Similar Facts and Similar Legal Basis by case, while preserving the reference–similar pairing. For the supervised experiments, we concatenated Facts and Legal Basis to compose two sequences, Paradigm (Paradigm Facts + Paradigm Legal Basis) and Similar (Similar Facts + Similar Legal Basis). These sequences form the pairs (sentence1, sentence2).

4.2 Zero-shot similarity with pretrained embeddings

We applied zero-shot inference to a subset of 500 pairs drawn from the hold-out of 10,000 documents. We used the same pretrained encoders evaluated during FN, without any weight updates. We kept the same corpus preprocessing and built two representations per pair: Paradigm, composed of Paradigm Facts and Paradigm Legal Basis, and Similar, composed of Similar Facts and Similar Legal Basis. Each representation was tokenized with the model's native scheme up to 512 tokens, with standard truncation and padding. We then computed proximity scores between the two sequences using cosine similarity and Inner Product (IP), and used these scores for ranking and descriptive analysis with respect to BERNA's reference scores. There was no threshold calibration or windowing, keeping the procedure strictly zero-shot and reproducible with a fixed seed.

4.3 Supervised FN of Portuguese legal embeddings

We evaluated BERTimbau Base, BERTimbau Large, STJ Iris (Melo et al., 2023; Fonseca et al., 2016; Real et al., 2020; May, 2021), Legalbert-pt, RoBERTaLexPT-base (Garcia et al., 2024), and JurisBERT (Viegas et al., 2023) with the same configurations and hyperparameters shown in Table 2. The sequences were limited to 512 tokens with trun-

cation and the native padding of each model, and the data split follows the corpora section. The evaluation considered cosine similarity, IP, Euclidean distance, and Manhattan distance.

4.4 Evaluation via vector retrieval and incremental clustering

After FN, we applied vector retrieval on the 10,000-document hold-out. This stage is distinct from zero-shot inference, which used 500 pairs from the same hold-out. For each petition, we generated two embeddings, one for the Facts block and another for the Legal Basis block, using the same encoder. We then concatenated the resulting vectors (dimension $2d$) and applied ℓ_2 normalization to the combined vector before indexing and search. Next, we indexed the vectors in Milvus with HNSW, using $M = 16$ and $efConstruction = 256$. To investigate partitioning behavior under different similarity settings in the index, we ran the same routine under two metric configurations in Milvus: cosine and IP in the implementation. Since the vectors are ℓ_2-normalized, cosine and IP are theoretically equivalent; nevertheless, we empirically report the observed partitioning under approximate search (HNSW) and incremental ingestion, as small variations in the returned neighborhood may affect association decisions near the threshold.

Searches used $ef = 64$ and returned the single nearest neighbor. Clustering was incremental and representative-based. For each new petition, we queried the index for the nearest neighbor and, when the returned score (the higher, the more similar for cosine/IP) was greater than or equal to 0.90, we assigned the petition to the cluster of the retrieved representative. Otherwise, we created a new cluster and inserted into the index only the representative vector of this new group. Thus, the index maintains one vector per cluster (the representative), rather than all vectors of the associated petitions. The results were audited by the business team of the granting agency through sample inspection and validation using score tables and undirected graphs, which allowed us to assess the internal consistency of the clusters and the suitability of the adopted threshold. We emphasize that the incremental procedure is sensitive to insertion order and to the threshold; therefore, all runs were performed with a fixed order and fixed parameters to ensure comparability across conditions.

4.5 Reproducibility

All code used in this work, including the preprocessing pipeline, training/evaluation scripts, and configurations, is available on GitHub[3].

5 Results and Discussion

In this section, we present the results and discuss implications for operational use in judicial case collections. Tables 3 and 4 summarize the comparison between two configurations, without FN (original model) and with FN. The first table reports the quality of the groupings induced by the embeddings through the Silhouette Score, while the second quantifies the agreement between similarity scores and the reference scores using Pearson and Spearman correlations, under different distance and similarity functions. To facilitate reading Table 4, **Cos-P** and **Cos-S** denote, respectively, the Pearson and Spearman correlations when similarity is computed with cosine; **Euc-P** and **Euc-S** refer to the correlations under Euclidean distance (mapped to similarity); and **Man-P** and **Man-S** to the correlations under Manhattan distance (with the same mapping). The comparison between Pearson and Spearman allows us to distinguish, respectively, linear agreement and monotonic agreement between the model scores and the reference scores.

Table 4 shows that, in the no-FN setting, there are substantial differences between models before any domain adaptation. STJ Iris presents the most consistent baseline, leading the six reported measures and reaching, for example, a Spearman correlation of 0.5214 under cosine. This result indicates that a relevant portion of the semantic signal captured by the reference is already recoverable with pretrained representations, but it also sets a practical ceiling for zero-shot use when the goal is to faithfully reproduce the ranking of the reference system. After FN, correlations increase systematically across all models, shifting performance to an approximate range of 0.62 to 0.73, which confirms the specialization effect for the textual pattern of initial petitions and for the proximity criterion represented in the supervised set. In this setting, STJ Iris maintains high and stable performance on rank-based metrics, for example, **Cos-S** of 0.7251, while peaks on specific measures appear in other models, for example, **Cos-P** of 0.7315 in JurisBERT. Taken together, the results indicate that FN improves both

[3] https://github.com/Willgnner-Santos/LexEmbed

Model	Status	Clustering (Silhouette Score)				
		Agg.	KMeans	Affinity	DBSCAN	HDBSCAN
STJ Iris	Original	0.4101	0.3892	0.1836	-0.1271	0.0876
	FN	0.3557	0.3310	0.1976	-0.0183	**0.1638**
BERTimbau Large	Original	0.4526	0.4344	**0.4426**	-0.0035	0.0504
	FN	0.3956	**0.3875**	0.2404	-0.0201	0.0611
BERTimbau Base	Original	0.4916	0.4584	0.4414	0.1228	0.0862
	FN	0.3902	0.3575	0.2501	-0.0031	0.0765
Legalbert-pt	Original	**0.5075**	0.4726	0.4315	**0.1448**	**0.1132**
	FN	0.3326	0.3055	**0.2590**	**0.0458**	0.0337
RoBERTaLexPT-base	Original	0.4986	**0.4806**	–	–	0.0637
	FN	**0.4172**	0.3757	0.1533	0.0101	0.0038
JurisBERT	Original	0.4731	0.4514	0.3566	-0.0481	0.0351
	FN	0.3727	0.3321	0.2522	-0.0755	0.0624

Table 3: Performance in clustering (Silhouette Score) comparing original models (without FN) and after FN. Symbol '–' indicates a measure not calculated.

Model	Original (without FN)						After FN					
	Cos-P	Cos-S	Euc-P	Euc-S	Man-P	Man-S	Cos-P	Cos-S	Euc-P	Euc-S	Man-P	Man-S
STJ Iris	**0.4030**	**0.5214**	**0.4766**	**0.5126**	**0.4767**	**0.5132**	0.7244	**0.7251**	0.7077	**0.7206**	0.7069	**0.7198**
BERTimbau Large	0.2577	0.3956	0.3639	0.4031	0.3659	0.4039	0.6928	0.6932	0.6919	0.6928	0.6914	0.6928
BERTimbau Base	0.2621	0.4529	0.3776	0.4495	0.3770	0.4494	0.7178	0.7182	**0.7185**	0.7196	**0.7176**	0.7187
Legalbert-pt	0.2681	0.4674	0.3981	0.4730	0.3993	0.4754	0.6822	0.6703	0.6699	0.6680	0.6692	0.6674
RoBERTaLexPT-base	0.2736	0.4989	0.3902	0.4977	0.4051	0.5013	0.6156	0.6481	0.6631	0.6566	0.6638	0.6570
JurisBERT	0.2855	0.4646	0.4130	0.4660	0.4144	0.4665	**0.7315**	0.7162	0.6720	0.6913	0.6722	0.6910

Table 4: Pearson and Spearman correlations between similarity scores and reference scores, comparing original models (without FN) and after FN.

linear agreement, Pearson, and monotonic agreement, Spearman, with expected variations across models depending on the induced geometry and the sensitivity of each similarity function.

Table 3 provides a different and necessary reading. Instead of evaluating pairs, it quantifies how much the vector space separates instances into cohesive and well-defined groups under classical clustering algorithms. In the no-FN configuration, centroid-based and hierarchical methods (Agglomerative Clustering (Agg.) and KMeans) tend to produce higher Silhouette values than density-based approaches, with Legalbert-pt standing out in Agg. (0.5075) and RoBERTaLexPT-base in KMeans (0.4806). After FN, a reduction in the Silhouette Score is observed for most models and algorithms, which is consistent with the nature of supervised training. By optimizing pairwise matching, the model can reorganize the vector space in favor of local neighborhoods that are important for ranking, without necessarily maximizing the global separation required by unsupervised partitions. Even so, there are locally positive effects in density-based methods, as in the HDBSCAN results for STJ Iris, which reaches 0.1638 after FN, suggesting greater regularity of densities in semantically close regions even when global separation does not increase.

The two readings, supervised fidelity on pairs and the unsupervised structure of the space, are deliberately treated as complementary, because high correlation does not, by itself, guarantee better overall organization of the collection under indexing and continuous ingestion. This distinction is further explored in the vector retrieval stage with incremental clustering on the hold-out set.

Figure 3 summarizes the combined effect of FN and the index configurations on partitioning granularity under the incremental rule with a 0.90 threshold. Although cosine and IP are theoretically equivalent when vectors are normalized ℓ_2, we observed differences in the partitioning under approximate

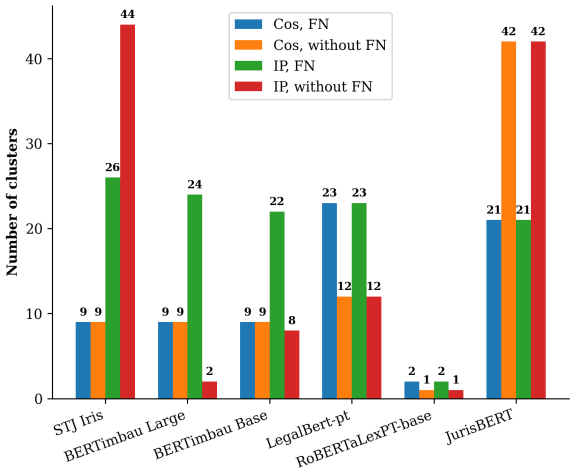

Figure 3: Partitioning of the hold-out set induced by the presence or absence of FN and by the metric configured in the vector index, cosine or IP, under approximate HNSW search and the association rule with a 0.90 threshold. Each bar represents the number of clusters resulting from each configuration. The labels "FN" and "no FN" indicate models with and without FN, respectively.

search (HNSW) and incremental ingestion in regions near the threshold, where small variations in the returned neighborhood can change association decisions. Therefore, we treat the index similarity configuration as an operational parameter to be calibrated jointly with the threshold and the encoder, supported by qualitative validation to avoid excessive merging of groups or artificial fragmentation of the collection. From an applied perspective, this reinforces the need for multi-metric validation and qualitative checks to prevent, on the one hand, excessive group merging and, on the other, artificial fragmentation of the collection.

The qualitative validation conducted by the business area of the partner institution was decisive to interpret the results beyond the aggregated metrics. Samples of clusters and retrieved links were audited, with checks of thematic coherence between Facts and Legal Basis, detection of duplications, and group stability under the incremental insertion of new documents. Based on this procedure, STJ Iris and BERTimbau Large were selected as bases to derive specialized variants, resulting in the models LexIris-pt and LexBert-pt. These variants, obtained through supervised FN, were incorporated into the institution's environment to support the screening and organization of repetitive claims and predatory litigation, constituting the main applied contribution of this work to the Portuguese legal

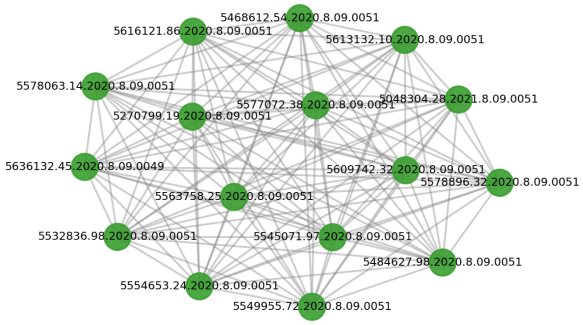

Figure 4: Slice of a cluster obtained by vector retrieval with the STJ Iris model. The nodes represent cases and the edges indicate similarity in the index.

NLP ecosystem.

Figure 4 illustrates a slice of a cluster generated by the vector retrieval approach to organize large document collections, using the STJ Iris model. The graph shows vertices representing cases and edges weighted by the similarity in the index. The image was selected as a visual example of the grouping produced by the incremental procedure.

In summary, the results support a coherent picture. Zero-shot inference establishes a reproducible baseline; FN consistently increases the fidelity of similarity scores; and vector retrieval with incremental clustering makes explicit the behavior at scale and the dependence on the metric and threshold, reinforcing the need for multi-metric evaluation and human validation in heterogeneous legal settings. Finally, the LexIris-pt[4] and LexBert-pt[5] variants were made available with documentation and weights ready for inference on Hugging Face.

6 Conclusion

This study presented and evaluated two specialized sentence-embedding models for the Portuguese legal domain, LexIris-pt and LexBert-pt, trained from corpora of initial petitions. The comparison across three complementary similarity-measurement tracks (zero-shot, supervised FN, and vector retrieval) indicates that FN consistently increases the faithfulness of similarity scores, reflected in higher correlations under multiple metrics, and improves performance in operational search and triage scenarios.

Expert validation by the partner agency confirmed the practical utility of the results and sup-

[4] https://huggingface.co/DIACDE/stjiris_tjgo_diacde_sts

[5] https://huggingface.co/DIACDE/bertimbaularge_tjgo_diacde_sts

ported the operational adoption of the models, particularly their domain-adapted versions. In practice, we observed gains in document triage and in organizing highly redundant sets, while preserving reproducibility and a computational cost compatible with continuous use. Nevertheless, improvements in pairwise matching do not necessarily imply maximization of the global separation of the vector space, reinforcing the importance of multi-metric evaluation and human validation in heterogeneous legal collections.

As next steps, we plan to incorporate sliding windows or larger-context models to mitigate truncation, perform continued pretraining with national legal collections, explore hard-negative mining and new contrastive losses, calibrate scores via temperature scaling or isotonic regression, and adopt active learning with experts. We also foresee explainability modules with anchored excerpts, multi-institutional evaluation, and monitoring of semantic drift in the production index.

Limitations

The corpus comes from a single jurisdiction, focuses on initial petitions, and covers a specific time interval, which limits generalization to other institutions and case classes and requires external validation. The texts were processed with a 512-token limit without windowing, which may lose information in long passages, and preprocessing and deduplication choices may have removed relevant variations. From an operational standpoint, we did not measure end-to-end cost and latency, nor did we isolate, in a controlled way, the causes of the differences between cosine and IP under approximate search; therefore, we treat these variations as an empirical configuration effect and reinforce the need for calibration with qualitative validation. Finally, we did not evaluate temporal drift, domain shift, or fairness, and there remains a risk of indirect re-identification in long texts, even after anonymization.

Ethical considerations

Data handling followed the LGPD guidelines, with data minimization, anonymization, and replacement of identifiers with hash keys when applicable, as well as role-based access control and audit logging. In addition, the corpus was anonymized in accordance with CNJ guidelines for data handling and protection in the context of the Judiciary. Access to the dataset is provided upon request and subject to formal institutional authorization, as described for the corpus.

The models should not be used as substitutes for judicial decision-making. Recommended use includes assisted screening, retrieval of similar cases, and support for procedural standardization, always under qualified human supervision. To mitigate risks of bias and spurious associations, we adopted pipeline documentation, expert review, and operational safeguards, including threshold calibration and human fallback.

Acknowledgements

The authors would like to acknowledge Brazilian Research Agencies FAPEG (Process: 202510267001606), CNPq (N 18/2024), Court of Justice of the State of Goiás (TJGO), and Center for Excellence in Artificial Intelligence (CEIA), Institute of Informatics, Federal University of Goiás (UFG).

References

Bruna Armonas Colombo, Pedro Buck, and Vinicius Miana Bezerra. 2017. Challenges when using jurimetrics in brazil—a survey of courts. *Future Internet*, 9(4):68.

Mina Ghadimi Atigh, Stephanie Nargang, Martin Keller-Ressel, and Pascal Mettes. 2025. Simzsl: Zero-shot learning beyond a pre-defined semantic embedding space. *International Journal of Computer Vision*, pages 1–17.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

Murillo Simiema Campos. 2022. O uso da inteligência artificial no poder judiciário: as contribuições do sistema berna para o tribunal de justiça de goiás.

Caio Castelliano, Tomas Aquino Guimaraes, and Adalmir de Oliveira Gomes. 2024. Fatores que aumentam o tempo do processo judicial no brasil. *Revista de Administração Pública*, 58:e2023–0175.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *Preprint*, arXiv:2010.02559.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Conselho Nacional de Justiça (Brasil). 2025. Resolution no. 615 of march 11, 2025. *Atos Normativos do Conselho Nacional de Justiça*. Accessed on: 9 dez. 2025.

Daniel da Silva Junior, Paulo Roberto dos Santos Corval, Daniel de Oliveira, and Aline Paes. 2024. Datasets for portuguese legal semantic textual similarity. *Journal of Information and Data Management*, 15(1):206–215.

Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2021. Exploring semantic similarity between german legal texts and referred laws. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pages 37–50. Springer.

Graziella De Martino, Gianvito Pio, and Michelangelo Ceci. 2023. Multi-view overlapping clustering for the identification of the subject matter of legal judgments. *Information Sciences*, 638:118956.

Raphael Souza de Oliveira and Erick Giovani Sperandio Nascimento. 2021. Clustering by similarity of brazilian legal documents using natural language processing approaches. In *Data clustering*. IntechOpen.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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.

Maria Duarte, Pedro A Santos, João Dias, and Jorge Baptista. 2022. Semantic norm recognition and its application to portuguese law. *arXiv preprint arXiv:2203.05425*.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings. *Preprint*, arXiv:2104.08821.

Eduardo Garcia, Nadia Silva, Felipe Siqueira, Juliana Gomes, Hidelberg O Albuquerque, Ellen Souza, Eliomar Lima, and André de Carvalho. 2024. Robertalexpt: A legal roberta model pretrained with deduplication for portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 374–383.

Angelina Grace. 2025. Leveraging natural language processing (nlp) for intelligent incident ticket classification.

Mingqiang Guo, Ying Huang, Qingfeng Guan, Zhong Xie, and Liang Wu. 2017. An efficient data organization and scheduling strategy for accelerating large vector data rendering. *Transactions in GIS*, 21(6):1217–1236.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

Sarika Jain, Deepak Jaglan, and Kapil Gupta. 2022. Investigating the similarity of court decisions. *ACI@ ISIC*, pages 316–326.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tiago Lima, Kellyton Brito, André CA Nascimento, George Valença, and Fábio Pedrosa. 2022. Using natural language processing to improve transparency by enhancing the understanding of legal decisions. In *EGOV-CeDEM-ePart-**.

George C Linderman and Stefan Steinerberger. 2019. Clustering with t-sne, provably. *SIAM journal on mathematics of data science*, 1(2):313–332.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Laila Neves Lorenzon. 2021. Análise comparada entre regulamentações de dados pessoais no brasil e na união europeia (lgpd e gdpr) e seus respectivos instrumentos de enforcement. *Revista do Programa de Direito da União Europeia*, 1:39–52.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Arpan Mandal, Kripabandhu Ghosh, Saptarshi Ghosh, and Sekhar Mandal. 2021. Unsupervised approaches for measuring textual similarity between legal court case reports. *Artificial Intelligence and Law*, 29(3):417–451.

Arooba Maqsood, Adnan Ul-Hasan, and Faisal Shafait. 2024. Transformer-based architecture for judgment prediction and explanation in legal proceedings. In *International Workshop on Document Analysis Systems*, pages 20–36. Springer.

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

Leland McInnes, John Healy, Steve Astels, and 1 others. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Rui Melo, Pedro A. Santos, and João Dias. 2023. A semantic search system for the supremo tribunal de justiça. In *Progress in Artificial Intelligence*, pages 142–154, Cham. Springer Nature Switzerland.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022. Memorization in nlp fine-tuning methods. *arXiv preprint arXiv:2205.12506*.

Maheshwaran Nithya, S Harini, S Kavyadharshini, and K Srinidhi. 2024. Ai-driven legal automation to enhance legal processes with natural language processing. In *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, pages 1246–1253. IEEE.

Joshua Page and Joe Soss. 2021. The predatory dimensions of criminal justice. *Science*, 374(6565):291–294.

Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia, and Renato Vicente. 2021. Legalnlp–natural language processing methods for the brazilian legal language. *arXiv preprint arXiv:2110.15709*.

Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Patrick Fernandes Rezende Ribeiro, Juliane de Lima Pires, Patrick Alves Bastos, Roberto Rigo, Henrique Assumpção dos Reis, Kamilly Voitkiv Hubner, Maria Fernanda Zandoná Casagrande, Bruno de Paula Marafiga, Dante Krol Simba, and Denise Fukumi Tsunoda. 2025. Bancos vetoriais e modelos de embedding: Avaliação comparativa de desempenho na recuperação semântica em língua portuguesa. *Research, Society and Development*, 14(10):e106141049768–e106141049768.

Rodrigo Amorim Ruiz. 2021. *Jurisprudence search based on facts similarity using NLP and ML techniques.* Ph.D. thesis, Universidade de São Paulo.

Jaakko Sauvola, Sasu Tarkoma, Mika Klemettinen, Jukka Riekki, and David Doermann. 2024. Future of software development with generative ai. *Automated Software Engineering*, 31(1):26.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 29(5):1199–1222.

Daniel da Silva Junior, Daniel de Oliveira, and Aline Paes. 2025. Evaluating text representations for unsupervised legal semantic textual similarity in brazilian portuguese. *Discover Data*, 3(1):23.

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 *Brazilian Conference on Intelligent Systems*, pages 268–282. Springer.

Saurabh Kumar Singh and Radhe Shyam Jha. 2024. Legal framework vs. practical realities: The effectiveness of fast track courts in achieving speedy justice. *Issue 3 Int'l JL Mgmt. & Human.*, 7:2117.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Daniela Vianna, Edleno Silva de Moura, and Altigran Soares da Silva. 2024. A topic discovery approach for unsupervised organization of legal document collections. *Artificial Intelligence and Law*, 32(4):1045–1074.

Carolina Castro Costa Viegas. 2022. *Dos limites do dever de informar e do dever de informar-se: análise a partir do impacto da assimetria da informação no direito contratual*. Ph.D. thesis, Universidade de São Paulo.

Charles FO Viegas, Bruno C Costa, and Renato P Ishii. 2023. Jurisbert: a new approach that converts a classification corpus into an sts one. In *International Conference on Computational Science and Its Applications*, pages 349–365. Springer.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Ashwini V Zadgaonkar and Avinash J Agrawal. 2021. An overview of information extraction techniques for legal document analysis and processing. *International Journal of Electrical & Computer Engineering (2088-8708)*, 11(6).

Luciano Zanuz and Sandro José Rigo. 2022. Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 219–229. Springer.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for

law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.