



Datasets for Portuguese Legal Semantic Textual Similarity

Daniel da Silva Junior   [Institute of Computing - UFF | danieljunior@id.uff.br]

Paulo Roberto dos Santos Corval  [Law School - UFF | paulocorval@id.uff.br]

Daniel de Oliveira  [Institute of Computing - UFF | danielcmo@ic.uff.br]

Aline Paes  [Institute of Computing - UFF | alinepaes@ic.uff.br]

 *Institute of Computing, Universidade Federal Fluminense, Av. Gal. Milton Tavares de Souza, s/n, São Domingos, Niterói, RJ, 24210-590, Brazil.*

Received: 21 July 2023 • Published: 5 April 2024

Abstract The Brazilian judiciary faces a significant workload, leading to prolonged durations for legal proceedings. In response, the Brazilian National Council of Justice introduced the Resolution 469/2022, which provides formal guidelines for document and process digitalization, thereby creating the opportunity to implement automatic techniques in the legal field. These techniques aim to assist with various tasks, especially managing the large volume of texts involved in law procedures. Notably, Artificial Intelligence (AI) techniques open room to process and extract valuable information from textual data, which could significantly expedite the process. However, one of the challenges lies in the scarcity of datasets specific to the legal domain required for various AI techniques. Obtaining such datasets is difficult as they require some expertise for labeling. To address this challenge, this article presents four datasets from the legal domain: two include unlabelled documents and metadata, while the other two are labeled using a heuristic approach designed for use in textual semantic similarity tasks. Additionally, the article presents a small ground truth dataset generated from domain expert annotations to evaluate the effectiveness of the proposed heuristic labeling process. The analysis of the ground truth labels highlights that conducting semantic analysis of domain-specific texts can be challenging, even for domain experts. Nonetheless, the comparison between the ground truth and heuristic labels demonstrates the utility and effectiveness of the heuristic labeling approach.

Keywords: Legal Dataset, Semantic Textual Similarity, Data Annotation

© Published under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0)

1 Introduction

According to the *Justice in Numbers* Report 2021 edition¹, the Brazilian Judiciary concluded 2020 with 75.3 million ongoing cases, of which 25.8 million were newly opened cases during that year. The high number of unsolved cases can be attributed to several factors, including an insufficient human workforce to cope with the demands and an extensive legal framework comprising over 34,000 laws². Additionally, Brazil's population, estimated at 213 million inhabitants in 2020³, ranks it as the sixth most populous country in the world, contributing to a large number of potential litigants. However, the *Justice in Numbers* Report highlights a positive trend in the productivity of the Brazilian Judiciary. This productivity increase results from the Judiciary's prioritization of reducing the backlog of ongoing cases. Nevertheless, even with this improvement, it could take over 50 years to clear the existing process inventory if the current pace continues.

Digitizing the inventory of legal processes⁴ represents one of the initiatives to alleviate the burden on the judicial system. This digital transformation also enables the utiliza-

tion of computational resources that facilitate the analysis of processes and, in certain instances, automate repetitive tasks involving processing a substantial volume of documents. The automation of tasks within the legal context has gained support from various legal entities⁵, with the adoption of Artificial Intelligence (AI) techniques, such as Legal Document Classification [Dal Pont *et al.*, 2020] and Semantic Textual Similarity [de Oliveira and Nascimento, 2022]. Machine Learning (ML) and Natural Language Processing (NLP) methods are predominantly employed to address these tasks.

The search for similar processes in the legal domain is conducted exhaustively since previous cases can serve as a foundation for new ones. The outcome of this search is advantageous for both the litigant, who can use similar cases as a reference for their petition, and for the judge, as it expedites the analysis of the current case. It is worth noticing that this type of search proves to be more effective when considering the textual components of the cases, particularly when evaluating the *semantic similarity* between them.

Automating tasks within the legal scenario is crucial to diminish the backlog of unresolved cases, making AI a valuable ally in this endeavor. However, delving into AI methods and devising new specialized techniques for the legal domain

¹<https://tinyurl.com/bdhbj244>

²<https://tinyurl.com/ytzrhc4t>

³<https://tinyurl.com/mr33fss7>

⁴<https://tinyurl.com/25ep43s8>

⁵<https://tinyurl.com/2v76r4d4>

requires the availability of datasets. Moreover, the automation of specific tasks demands specialized datasets to harness more sophisticated AI methods effectively. Additionally, numerous tasks in the legal domain, such as retrieving similar documents, necessitate *annotated datasets*. Nevertheless, the annotation process proves particularly challenging in the legal domain, as it requires experts who deeply understand the context and vocabulary used to describe the legal processes, which may not be trivial to find.

This article presents four datasets specific to the Portuguese legal domain, primarily focusing on semantic textual similarity to facilitate similar document retrieval. Two of these datasets, namely *TCU Votes*⁶ and *STJ Judgments*⁷, encompass texts and metadata extracted from the portals of both entities, but they do not include any annotations. The other two datasets, *TCU Votes for Textual Semantic Similarity* and *STJ Judgments for Textual Semantic Similarity*, were derived from the aforementioned datasets. These datasets were created using a heuristic method proposed in this article to annotate similar documents. Moreover, the article introduces a ground truth dataset for Semantic Textual Similarity, incorporating data from the STJ Semantic Textual Similarity dataset, which legal domain experts annotated. This ground truth dataset proved instrumental in evaluating the heuristic Semantic Textual Similarity dataset, revealing a moderate correlation between the expert and heuristic labels.

This article is an extension of the conference paper [Silva-Junior et al., 2022] published in the Proceedings of the 2022 Dataset Showcase Workshop. This extended version provides more details regarding the proposed datasets and a ground truth provided by legal domain experts. The article is organized into five sections besides this introduction. Section 2 discusses other datasets from the legal domain in Portuguese. Section 3 presents the datasets *TCU votes* and *STJ judgments*. Subsequently, Section 4 presents the datasets *TCU Votes for Textual Semantic Similarity* and *STJ Judgments for Textual Semantic Similarity*, as well as the heuristics used for their generation. Section 5 describes the annotation process for the ground truth dataset, data analysis, and comparison with the heuristic dataset. Finally, Section 7 concludes the article and discusses future work.

2 Related Work

To tackle the Semantic Textual Similarity (STS) task, Joshi et al. [2023] proposes the *U-CREAT* pipeline that enhances BM25 retrieval results by extracting structured events. To evaluate the proposed pipeline, Joshi et al. [2023] also proposes *Indian Legal Prior Case Retrieval corpus* (IL-PCR corpus), which contains a corpus of Indian legal documents in English. The literature lacks annotated datasets for the STS task with Portuguese legal data. However, some legal datasets already contain a corpus of textual data without annotation, and others have annotations for addressing other specific tasks. The Iudicium Textum Dataset [Willian Sousa and Fabro, 2019] comprises 41,353 documents of judgments from the Federal Supreme Court (*Superior Tribunal Federal*

in Portuguese - STF) published between 2010 and 2018. Additionally, de Oliveira and Júnior [2017] presents a dataset containing jurisprudence from the Supreme Court of the State of Sergipe, consisting of four collections: (i) judgments of the Justice Court (181,994 documents); (ii) monocratic decisions of the Justice Court (37,142 documents); (iii) judgments by Special Courts (37,161 records); and (iv) monocratic decisions by Special Courts (23,151 documents). The Iudicium Textum Dataset [Willian Sousa and Fabro, 2019] and the corpus provided by de Oliveira and Júnior [2017] are unlabeled datasets.

For the textual classification task, the VICTOR dataset [Luz de Araujo et al., 2020] stands out with over 692,000 documents from the Federal Supreme Court. A team of experts has manually annotated this dataset to facilitate document-type classification tasks and process topic assignments. The LeNER-BR dataset [de Araujo et al., 2018] consists of 70 documents from judicial courts and Brazilian laws. It serves the named entity recognition (NER) task and is annotated with both general-purpose entities and specific entities of legal knowledge, such as “Legislation” for laws and “Jurisprudence” for judicial decisions resulting from legal proceedings. Furthermore, the UlyssesNER-Br dataset [Albuquerque et al., 2022] has also been developed for the NER task, created within the scope of the Chamber of Deputies. This dataset includes general and specific legal entities, such as “Fundamental” and “Product of Law”. It is divided into two subsets: (i) the *PL-corpus*, containing 9,526 publicly available sentence bills, and (ii) the *ST-corpus*, which consists of private internal documents with 790 sentences of work requests.

3 TCU Votes and STJ Decisions Corpora

The first two datasets presented in this article, *TCU votes* and *STJ decisions*, were derived from judgments of the *Superior Tribunal de Justiça (STJ)* and votes from the *Tribunal de Contas da União (TCU)*. The STJ and the TCU are collegiate bodies, meaning that decisions are reached through evaluation and consensus among the responsible members. The *decisions* consist of texts of judgments from collegiate bodies, which cover only the main points of a discussion. On the other hand, a *vote*, in the context of collegiate bodies, refers to the exposition, evaluation, and opinion on the decision to be taken for a specific case, carried out by the responsible member known as the *rapporteur*⁸.

The uniqueness of these datasets lies in the inclusion of precedents of *jurisprudence* used by the legal bodies. Jurisprudences are interpretations or understandings adopted by these bodies, which serve as guiding principles for making decisions on specific subjects. These interpretations are formulated by analyzing previous decisions on the same subject, known as precedents, and they aim to standardize decisions and expedite the resolution of recurrent matters.

⁶TCU = Federal Court of Accounts in Brazil

⁷STJ = Superior Tribunal of Justice in Brazil

⁸https://www.congressonacional.leg.br/legislacao-e-publicacoes/glossario-legislativo/-/legislativo/termo/relator_quanto_ao_papel

The texts were obtained through a data scraping routine from the respective websites of the collegiate bodies involved. Following the data scraping process, any records with missing or duplicated values were removed. As depicted in Table 1, the resulting *STJ decisions* dataset contains a significantly higher number of records, represented by the row labeled *Decisions*, in comparison to the *TCU votes* dataset, indicated by the row labeled *Votes*. Additionally, Table 1 highlights the superiority of jurisprudence representation in the *STJ decisions* dataset. Furthermore, Table 1 details the categorization applied to the data from each body, arranged in a hierarchical descending order. Specifically, in the TCU data, a vote is associated with a Subtopic, further grouped under a Topic, and finally falls within an Area.

Table 1. Characteristics of the *STJ decisions* and *TCU votes* datasets.

TCU		STJ	
Votes	371	Decisions	7403
Jurisprudences	44	Jurisprudences	1458
Areas	4	Subjects	7
Themes	27	Natures	68
Sub-theme	38		

The datasets presented in this article are available in *CSV* format and can be accessed at the following URL: <https://osf.io/k2qpx/>. The *TCU votes* dataset contains the following attributes: AREA, THEME, SUB-THEME, STATEMENT, PROCESS, YEAR, TYPE_PROCESS, REPORTER, and VOTE. The STATEMENT attribute specifies the jurisprudence to which a VOTE, representing a precedent, is associated. On the other hand, the *STJ Judgments* dataset comprises the attributes SUBJECT, NATURE, THEME, PROCESS, REPORTER, BODY, JUDGMENT_DATE, PUBLICATION_DATE, and SUMMARY. In this case, the THEME attribute defines the jurisprudence to which a SUMMARY, serving as a precedent, is associated.

The charts illustrating the composition of the aforementioned datasets are presented following. Figure 1 presents a histogram of the *TCU votes* dataset, revealing that, on average, each case has between seven and eight previous votes serving as jurisprudence. Conversely, the histogram shown in Figure 2, related to the *STJ decisions* dataset, demonstrates that most case law has between five and six precedent decisions.

Figure 3 shows that the majority of precedents in the *TCU votes* dataset are primarily from the LICITAÇÃO (BIDDING) area, followed by the PESSOAL (PERSONAL) area. Additionally, the LICITAÇÃO and PESSOAL areas present the greatest dispersion of precedents across different themes. Regarding the *STJ decisions* dataset, Figure 4 demonstrates that the prevalent precedents are mainly associated with the Subjects: Administrative Law, Civil Law, and Criminal Law. The dispersion of precedents within these three Subjects is also more significant than in the others.

The tag cloud of the *TCU votes* dataset, shown in Figure 5, highlights words such as OBRA (WORK), SERVIÇO (SERVICE), CONTRATO (CONTRACT), and LICITAÇÃO (BIDDING) as the most frequently occurring in the precedents of the dataset. Meanwhile, Figure 6 showcases terms

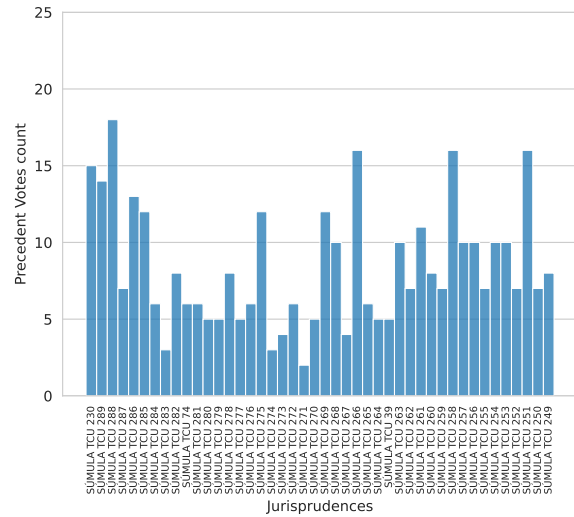


Figure 1. Histogram of Precedents X *TCU votes* Jurisprudences

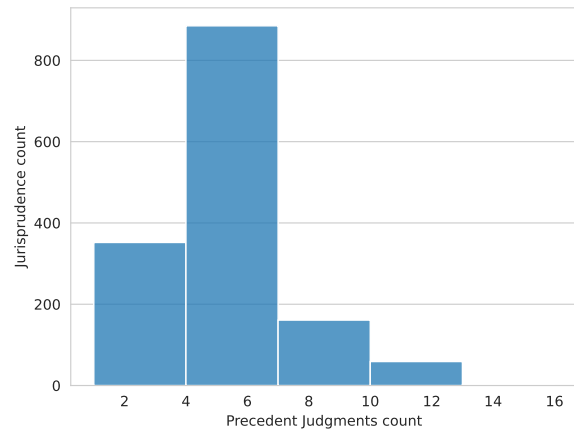


Figure 2. Histogram of Precedents X *STJ decisions* Jurisprudences.

like RECURSO ESPECIAL (SPECIAL APPEAL), HABEAS CORPUS, PROCESSUAL CIVIL (CIVIL PROCEDURE), and AGRAVO INTERNO (INTERNAL APPEAL) as the most frequent in the *STJ decisions* dataset.

The histogram in Figure 7 shows that most precedents in the *TCU votes* dataset contain up to 20,000 words. In this case, the words are defined by the spaces in the texts of the precedents. On the other hand, when considering the *STJ decisions*, Figure 8 reveals that most precedents in this dataset consist of up to 500 words.

4 Heuristic-Annotated Legal Semantic Textual Similarity Datasets

Given the significance of identifying similar legal cases and the lack of datasets to aid in training models for the Semantic Textual Similarity (STS) [Fonseca et al., 2016] task, the main contribution of this article lies in the creation of the datasets *TCU votes for Semantic Textual Similarity* and *STJ decisions for Semantic Textual Similarity*. These datasets were derived from the ones previously presented in Section 3 and specifically designed for the STS task [Fonseca et al., 2016]. Typically, an STS dataset consists of pairs of texts, each assigned a score reflecting their semantic similarity. A higher score indicates a stronger semantic resemblance between the texts.

Preparing datasets for the STS task can be laborious and

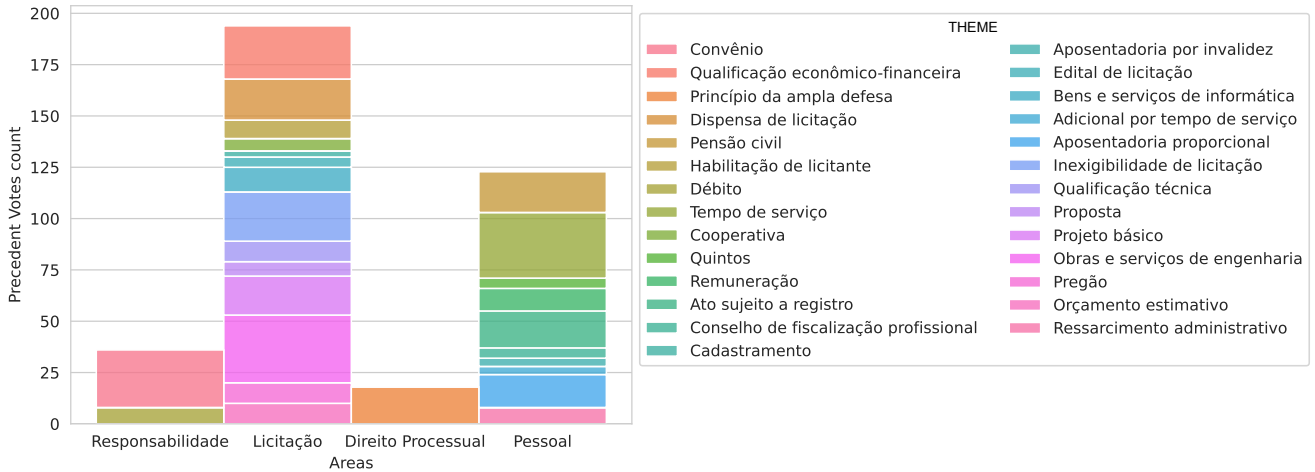


Figure 3. Histogram of VOTE X AREA X THEME of TCU votes

error-prone, often requiring human annotators who are experts in the data domain. This article presents a novel approach that automates the annotation process based on a heuristic derived from the metadata of the texts, thereby reducing the dependency on human annotators in STS tasks. Due to the distinct nature of both datasets, we propose two workflows to annotate the datasets from STJ and TCU. Specifically, to annotate the *STJ decisions for Semantic Textual Similarity* dataset, the workflow involves the following steps, considering the hierarchical order existing between the documents (Table 1):

1. Generate pairs between decisions of the *same Jurisprudence* and assign each pair a score with a base value of 4.5, along with a noise that follows a normal distribution across all generated pairs.
2. Generate pairs between decisions of the *same Nature* and assign each pair a score with a base value of 3 and a noise that follows a normal distribution across all generated pairs.
3. Generate pairs between decisions of *different Subjects* and assign each pair a score with a base value of 0.5 and a noise that follows a normal distribution across all generated pairs.
4. Generate the final set by joining the *same size* splits of subsets generated by steps 1,2 and 3.

The heuristic used to annotate the *STJ decisions for Semantic Textual Similarity* dataset assumes that decisions that served as precedents for the same Jurisprudence show a significant intrinsic similarity. Conversely, decisions that share the same Nature but are not precedents of the same Jurisprudence maintain a less pronounced similarity relationship. Finally, judgments dealing with different Subjects are notably dissimilar. However, this last set does not contain documents from different jurisprudence since, although they are not precedents of the same Jurisprudence, they may share the same Nature and thus retain some degree of similarity. The decision to choose three base values, (4.5, 3, 0.5), was made to simulate pairs of documents with high similarity, neutrality, or dissimilarity. Adding noise following a normal distribution aimed to simulate the uncertainty and variation in

annotations that might occur when a manual annotator performs the process.

The workflow to annotate the *TCU votes for Semantic Textual Similarity* dataset is similar to the previous one, except for the types of used metadata. In this case, the following steps are followed:

1. Generate pairs between votes of the *same Jurisprudence* and assign each pair a score with a base value of 4.5, along with noise that follows a normal distribution across all generated pairs.
2. Generate pairs between votes from the *same Area and Theme* and assign each pair a score with a base value of 3, along with the noise that follows a normal distribution across all generated pairs.
3. Generate pairs between votes from *different Areas* and assign each pair a score with a base value of 0.5 and noise that follows a normal distribution across all generated pairs.
4. Generate the final set by joining the *same size* splits of subsets generated by steps 1,2 and 3.

Compared to the previous procedure, a significant difference when labeling the *TCU votes for Semantic Textual Similarity* dataset lies in the second subset. This subset involves votes from the same Area and Theme. The data scraped from TCU include Themes with identical terminology but belonging to different Areas.

The TCU dataset comprises 4, 843 tuples, while the STJ dataset contains 51, 437. Following the automatic process of generating pairs and associated scores, as well as balancing between the subsets generated at each step, we further divide them into TRAINING, TEST, and VALIDATION sets, maintaining the proportion of pairs per similarity interval. As a result, the dataset for STS with TCU votes was divided into 3, 389 samples for training, 438 samples for validation, and 1, 016 samples for testing. Conversely, the dataset for STS with STJ decisions was divided into 36, 010 samples for training, 4, 613 samples for validation, and 10, 814 for testing.

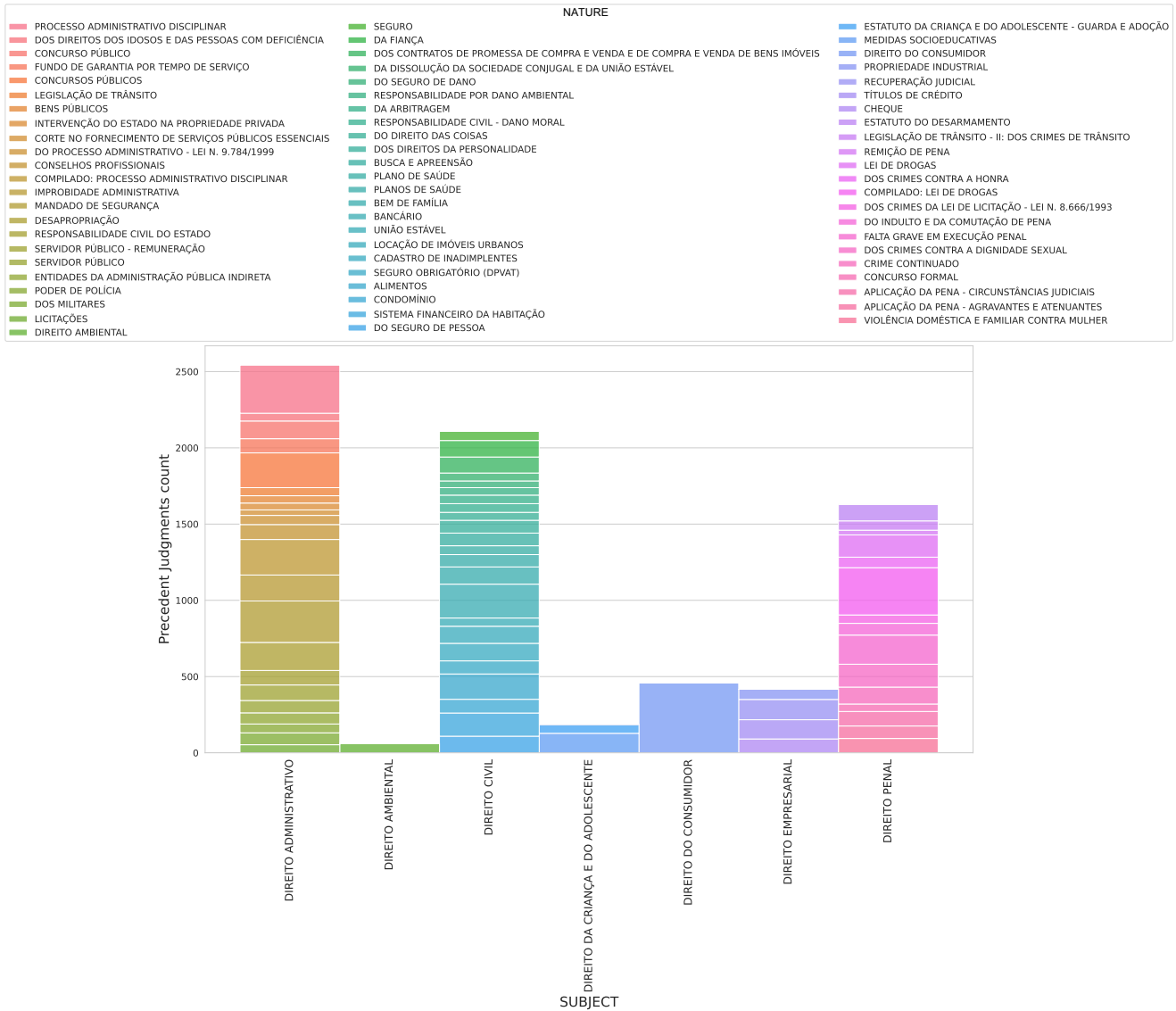


Figure 4. Histograms of DECISION X SUBJECT X NATURE of STJ decisions



Figure 5. Wordcloud TCU votes precedents.



Figure 6. Wordcloud STJ decisions precedents

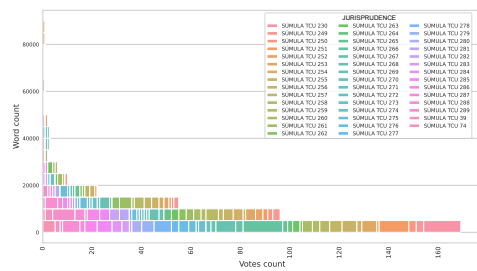


Figure 7. Histogram of words X TCU votes precedents

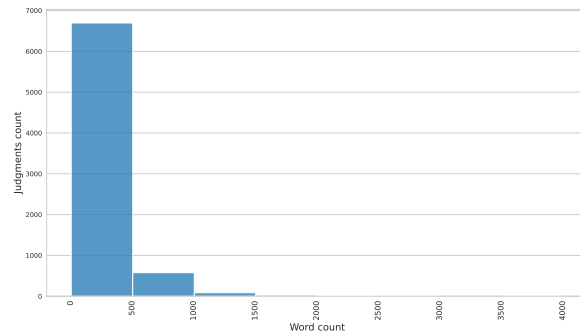


Figure 8. Histogram of words X STJ decisions precedents.

5 Building Datasets for the Legal STS Task Labeled by Experts

We collected labels from expert annotators and compared them to the labels generated by the proposed heuristic in this article. The annotation process was conducted using Google Forms⁹. Pairs of documents were presented to legal domain experts, and six questions were posed regarding these pairs:

1. How semantically similar are the two documents?
2. What is your level of confidence in the assigned similarity score?
3. Which part of the first document was most relevant for the attributed similarity?
4. Was the most relevant part of the first document in the header (initial part in capital letters) or in the body?
5. Which part of the second document was most relevant for the attributed similarity?
6. Was the most relevant part of the second document in the header (initial part in capital letters) or in the body?

The first question is of utmost importance as it allows us to evaluate the effectiveness of the heuristic labeling. The annotators are required to select one option from a Likert scale [Joshi *et al.*, 2015], which consists of five choices, with values ranging from 0 to 4:

- 0 - Not related
- 1 - Slightly related
- 2 - Moderately related but not similar
- 3 - Somewhat similar
- 4 - Highly similar

The annotators were provided with a guide, similar to the one proposed by Cer *et al.* [2017], which presented scenarios illustrating when each label is more likely to be applied. The second question aimed to gauge the uncertainty that even an expert annotator might find, thereby assisting in evaluating the heuristic labeling. The results of the third and fifth questions can benefit supervised Machine Learning methods by highlighting the relevant parts of the documents evaluated in the STS task. Conversely, the fourth and sixth questions aim to ascertain whether the document structure, which can be readily extracted, can be leveraged to enhance the heuristic labeling process.

As previously mentioned, obtaining access to expert annotators can be challenging. Therefore, we conducted the following experiment and obtained results using only the STJ decisions dataset. We chose to use STJ decisions instead of TCU votes because the TCU body encompasses a broader range of knowledge domains, which would require law experts with expertise in various subfields to evaluate the document pairs. For this annotation process, we enlisted the help of 27 students pursuing a Master’s in Law degree, who were invited to answer the questions in a classroom setting. We initially selected 140 document pairs from the STJ dataset’s test set to be annotated. As a result, we created fourteen *Google Forms*, each containing ten document pairs for annotation. Initially, we planned to use thirteen forms, with two experts annotating each form, and only one expert would annotate the remaining form. This approach would have provided us with a total of 270 labeled document pairs for the STS task. However, at the end of the form assignment, we collected 240 labeled document pairs, where two domain experts labeled 100 unique document pairs, and only one domain expert labeled 40 document pairs. Next, we investigated five research questions:

1. To what extent do the domain experts’ labels agree for the same pair of documents?
2. Where are the highlighted parts used to annotate the documents - in the body or in the header?
3. What is the distribution of the labels in the dataset annotated by the domain experts?
4. How closely do the domain expert and heuristic labels align?
5. What are the mean and standard deviation of the domain experts’ confidence in the assigned labels?

We examined the 100 unique document pairs that two domain experts labeled to address the first research question. Of these, only 32 pairs received identical labels from both experts. We then investigated the remaining 68 document pairs that were labeled differently by the two domain experts. We calculated the distance between the conflicting labels for each of these divergent pairs. We computed the mean, variance, and standard deviation based on these distances, which were 1.63, 0.58, and 0.76, respectively. To assess the correlation between the divergent labels, we used the Pearson and Spearman correlation coefficients and Krippendorff’s alpha [Krippendorff, 2004]. Although the latter would be more informative with more than two labels per document pair, we

⁹<https://www.google.com/forms/about/>

	Total	Mean distance
Partial divergences about non-similarity	28	1.11
Partial divergences on similarity	18	1.33
Total divergences	22	2.54

Table 2. Divergences in a stratified mode

still obtained meaningful results. The Pearson correlation was 0.63, and the Spearman correlation was 0.60, indicating a positive correlation between the divergent labels. This correlation is expected, given that all the labels are positive numbers and have a narrow range. Krippendorff’s alpha was -0.12 , which is appropriate since negative values imply less agreement than what would be expected by chance. This aligns with the fact that all the labels are divergent. In Figure 9, we present the distribution of distances between document pairs with divergent labels.

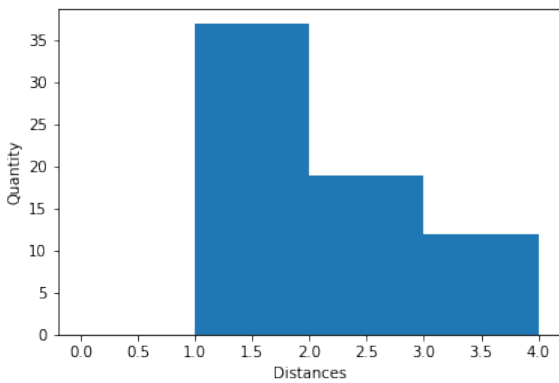


Figure 9. Histogram of the distance between labels, between the data labeled in divergence

We evaluated the divergent distances uniformly. However, in semantic similarity labeling using a Likert scale, some divergences can be considered as *partial divergences*, while others are *total divergences*. For instance, in a document pair where one domain expert assigns a label of 1 - *Slightly related*, and another domain expert assigns a label of 0 - *Not related*, this can be seen as a *partial divergence*, as neither expert perceives the documents as similar. On the other hand, if one domain expert gives a label of 0 - *Not related*, and another domain expert gives a label of 3 - *Somewhat similar*, it can be considered a *total divergence*. We thoroughly investigate and categorize cases where label divergences for document pairs fall within the range of $[0, 1, 2]$ or $[2, 3, 4]$ as *partial divergences*. In cases where one domain expert assigns a label in the range of $[0, 1]$, and another assigns a label in the range of $[3, 4]$, we consider this a *total divergence*.

Table 2 reveals that out of the 68 unique document pairs labeled divergently by two experts, 22 falls under the category of total divergence. This total divergence accounts for over 20% of the divergences observed among the 100 unique document pairs labeled by expert annotators. This finding highlights the challenging nature of the Semantic Textual Similarity task in the legal domain. As an example of a document pair with total divergent labels, consider the scenario where one domain expert assigns a similarity score of 0, while another expert assigns a score of 3:

Document 1

ADMINISTRATIVE AND CIVIL PROCEDURE. INTERNAL APPEAL IN THE APPEAL IN SPECIAL APPEAL. A QUO JUDGMENT THAT RESOLVED ALL THE CONTROVERSY POSTED IN THE FILES. SUFFICIENT RATIONALE. DENIAL OF JURISDICTIONAL PROVISION. NON OCCURRENCE. CONSUMER ACTION. REVERSAL OF THE BURDEN OF PROOF IN FAVOR OF PARQUET. POSSIBILITY.

1. In accordance with the jurisprudential guidance of this Superior Court, with the Court of origin having pronounced itself clearly and precisely on the issues raised in the case, based on sufficient grounds to support the decision, there is no talk of omission in the regional ruling since succinct reasoning does not mean the absence of grounds.

2. In the present case, the alleged offense to art. 1,022 of CPC/2015 did not occur insofar as the Court of origin resolved, with reason, the questions submitted to it, fully assessing the controversy placed in the case, and it cannot, furthermore, confuse a judgment unfavorable to the interest of the party with negative or lack of judicial provision.

3. Regarding the reversal of the burden of proof, the local Court aligned itself with the jurisprudence of this Sodality on the subject, whose understanding asserts that "in the consumer action initiated by the Public Prosecutor’s Office, there is no question of the plaintiff’s hypo sufficiency for the reversal of the burden of the proof, as the presence of Parquet as procedural substitute for the collective justifies it" (AgInt no AREsp 222.660/MS, Rel. Minister Gurgel de Faria, First Panel, DJe 12/19/2017).

4. Internal appeal that is dismissed.

Document 2

REGIMENTAL APPEAL. CIVIL PROCEDURE. THERE IS NO NEED TO TALK ABOUT A VIOLATION OF ARTICLE 535 OF THE CODE OF CIVIL PROCEDURE WHEN THE JUDGMENT RAISES, FOUNDALLY, THE ISSUES PERTINENT TO THE DISPUTE. UNDER THE TERMS OF SUMMARY 283 OF THE FEDERAL SUPREME COURT, WHEN THE DECISION APPELLED IS BASED ON MORE THAN ONE GROUND, THE APPEAL MUST COVER ALL OF THEM. SUMMARY 60 OF THIS COURT ORDERING VOID THE EXCHANGE OBLIGATION ASSUMED BY THE BORROWER’S ATTORNEY LINKED TO THE LENDER IN THIS IS THE EXCLUSIVE INTEREST. ADVISES SUMMARY 83 OF THIS COURT, WHICH IS NOT KNOWN OF AN APPEAL BASED ON DISAGREEMENT WHEN THE GUIDANCE OF THIS COURT WAS STATED IN THE SAME SENSE OF THE DECISION APPEALED TO.

APPLICATION OF THE FINE PROVIDED FOR ARTICLE 557, § 2, OF THE CIVIL PROCEDURE CODE. IMPROVED APPEAL.

To gain insights into the factors that may influence divergent annotation behavior in such scenarios, the document pair example and the two previous labels were presented to a third domain expert for examination. According to the domain expert, the label 3 - *Somewhat similar* is justified because both documents contain a procedural issue specific to the appeal before the STJ, which accounts for the perceived similarity. However, the expert also pointed out that the issues discussed on the merits of the appeal are entirely different, which justifies the label 0 - *Not related*. Those who assigned a label of 3 likely took into account this initial debate of a procedural nature. However, it should be noted that all appeals must examine this procedural aspect, making it less useful for distinguishing between them. On the other hand, label 0 is more interesting because it considers the appeals' fundamental issues, which are the primary focus of differentiation.

In addressing research question 2, we discovered that the positions of text portions are irrelevant in determining whether they are more likely to be highlighted in the document. This is evident from the ground truth dataset, where the text highlighted positions are uniformly distributed between the header and body of the documents. Furthermore, in 83% of cases, the highlighted text position in document 1 differs from that in document 2, indicating no consistent pattern in the positioning of highlighted text between document pairs.

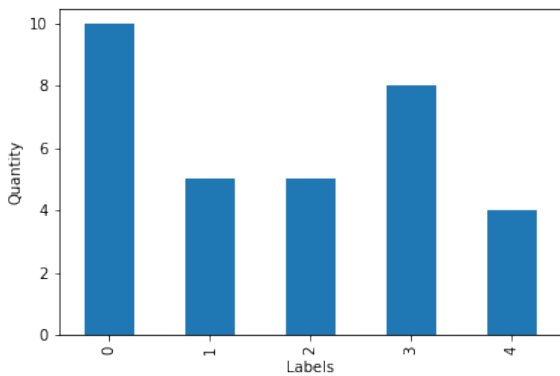


Figure 10. Domain expert labels distribution

With the assistance of Figure 10, we can address the third research question. As depicted in Figure 10, the labels assigned by the domain experts are predominantly 'Not related' between the documents, followed by 'Somewhat similar'. Given that the selected document pairs for annotation were drawn from the heuristic dataset, the heuristic can at least discern some similarity and dissimilarity between a pair of texts.

The fourth research question is paramount, as its answer directly measures the heuristic's effectiveness in labeling a dataset for STS with the STJ data. Only the 32 document pairs that were equally labeled by both domain experts were utilized for this analysis. To assess the heuristic's

performance, we calculated Pearson and Spearman correlations and Krippendorff's alpha using the labels from both the heuristic and ground truth datasets. The results revealed a Pearson correlation of 0.45, a Spearman correlation of 0.43, and a Krippendorff's alpha of 0.40. These metrics indicate a moderate positive correlation between the labels generated by the heuristic and the ground truth [Evans, 1996; Altman, 1990].

To address the fifth research question, we analyzed the responses regarding the self-confidence of the domain experts in their annotations. The confidence levels were collected using a Likert scale ranging from 0 to 4. The mean confidence level was 3.28, with a variance of 0.93 and a standard deviation of 0.96.

6 Discussion and Limitations

This article presents datasets for Legal STS tasks in Brazilian Portuguese through a proposed heuristic and by expert's legal domain annotation. The contribution of this paper follows the efforts employed in other languages, like English [Joshi et al., 2023] [Rabelo et al., 2022], to diminish the lack of legal annotated resources in Brazilian Portuguese. The results show that the proposed heuristic to annotate the dataset is helpful as the labels generated correlate reasonably to the expert annotations, and generating a gold standard set is too costly. As the proposed heuristic relies on latent text categories, one can generate variations of the methodology to generate datasets for other domains. However, without latent categories, the heuristic cannot be helpful. Although the comparative analyses between labels generated by the heuristic and expert annotation indicate the value of the heuristic, the expert-annotated data is still small, which limits more in-depth analysis. The availability of the proposed datasets enables the evaluation of several techniques for tasks like Legal Semantic Textual Similarity or Legal Information Retrieval [Sansone and Sperli, 2022]. Using the datasets, one can calculate the similarity between embeddings generated by several techniques as an unsupervised approach to retrieving similar cases or propose new architectures to classify document pairs as similar or non-similar.

7 Conclusions

This article contributes with resources to the legal domain by proposing two unlabelled datasets, a heuristic process for generating labeled datasets for the Semantic Textual Similarity task, two heuristic-labeled datasets for use in the Semantic Textual Similarity task, and a ground truth dataset derived from a subset of one of the heuristic-labeled datasets. The first two datasets were constructed from data collected from the Federal Court of Auditors (TCU) websites and the Superior Court of Justice (STJ). The data collection process yielded the *TCU votes* and *STJ decisions* datasets, both related to case law precedents. In addition to the textual content of the precedents, these datasets also include metadata related to categorizing the documents within the context of their respective bodies.

The main contribution of this article is the proposal of a heuristic for automatically annotating datasets for the Semantic Textual Similarity task using legal domain data. Leveraging the proposed heuristic process, the article also provides access to two heuristic-labeled datasets: *TCU Votes for Semantic Textual Similarity* and *STJ decisions for Semantic Textual Similarity*. These datasets were constructed from the collected precedents and the application of the heuristic. In addition to offering legal domain datasets and developing the heuristic, this article also includes an exploratory analysis of these datasets, further contributing to the understanding and utilization of the data.

The effectiveness of the proposed heuristic annotation was evaluated using a ground truth dataset generated through a data annotation process involving legal domain experts in the form of a question-answer experiment. This experiment revealed that the domain-specific annotation of semantic textual similarity can lead to relevant divergences in labeling between domain experts, underscoring the challenges of automating such a process. Furthermore, when comparing the heuristic labels with the ground truth labels, it was observed that the heuristic process can be used with confidence in generating labels.

Future work includes evaluating the heuristic's performance in contrast to similarity calculated by embedding techniques, as embeddings can be employed to assess unsupervised methods for Legal Information Retrieval. Additionally, the datasets provide an opportunity to adapt Language Models to the legal domain with Portuguese data [Paul et al., 2023].

Acknowledgements

The authors express their sincere gratitude to the master students of the Law School at Fluminense Federal University for their invaluable assistance in conducting the experiments.

Funding

This research was funded by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Financing Code 001*. The research was also partially supported by CNPq and FAPERJ.

Authors' Contributions

DSJ, DdO, AP and PRSC contributed to the conception of this study. DSJ performed the experiments. DSJ is the main contributor and writer of this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available at the following link: <https://osf.io/mct8s/>. The source code used in the study is also accessible at: <https://github.com/danieljunior/jidm>.

References

- Albuquerque, H., Costa, R., Silvestre, G., Souza, E. P., Félix, N., Vitória, D., and Carvalho, A. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. DOI: 10.1007/978-3-030-98305-5_1.
- Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.
- Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/S17-2001.
- Dal Pont, T. R., Sabo, I. C., Hübner, J. F., and Rover, A. J. (2020). Impact of text specificity and size on word embeddings performance: An empirical evaluation in brazilian legal domain. In *Brazilian Conference on Intelligent Systems*, pages 521–535. Springer.
- de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: A dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer.
- de Oliveira, R. A. N. and Júnior, M. C. (2017). Assessing the impact of stemming algorithms applied to judicial jurisprudence - an experimental analysis. In *Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 99–105. INSTICC, SciTePress. DOI: 10.5220/0006317100990105.
- de Oliveira, R. S. and Nascimento, E. G. S. (2022). Brazilian court documents clustered by similarity together using natural language processing approaches with transformers.
- Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.
- Fonseca, E., Santos, L., Criscuolo, M., and Aluisio, S. (2016). Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.
- Joshi, A., Kale, S., Chandel, S., and Pal, D. K. (2015). Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.
- Joshi, A., Sharma, A., Tanikella, S. K., and Modi, A. (2023). U-CREAT: Unsupervised case retrieval using events extrAc-Tion. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13899–13915, Toronto, Canada. Association for Computational Linguistics. DOI: 10.18653/v1/2023.acl-long.777.
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.
- Luz de Araujo, P. H., de Campos, T. E., Ataide Braz, F., and Correia da Silva, N. (2020). VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*,

- pages 1449–1458, Marseille, France. European Language Resources Association.
- Paul, S., Mandal, A., Goyal, P., and Ghosh, S. (2023). Pre-trained language models for the legal domain: A case study on indian law. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23*, page 187–196, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3594536.3595165.
- Rabelo, J., Goebel, R., Kim, M.-Y., Kano, Y., Yoshioka, M., and Satoh, K. (2022). Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.
- Sansone, C. and Sperlí, G. (2022). Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967. DOI: <https://doi.org/10.1016/j.is.2021.101967>.
- Silva-Junior, D., de Oliveira, D., and Paes, A. (2022). Criação de conjuntos de dados textuais jurídicos em português a partir de processo de extração e heurística. In *Anais do IV Dataset Showcase Workshop*, pages 91–100, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/dsw.2022.226253.
- Willian Sousa, A. and Fabro, M. (2019). Iudicium textum dataset uma base de textos jurídicos para nlp. In *Dataset Show Case Proceedings of 34th Brazilian Symposium on Databases*. SBC.