Criação de Conjuntos de Dados Textuais Jurídicos em Português a partir de Processo de Extração e Heurística*

Daniel Silva Junior[1], **Daniel de Oliveira**[1], **Aline Paes**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Caixa Postal 24210-310 – Niteroi, RJ, Brasil

danieljunior@id.uff.br, {danielcmo, alinepaes}@ic.uff.br

Abstract. *The Brazilian judiciary has a large workload, resulting in a long time to finish legal proceedings. Several digitization initiatives have emerged, opening up the possibility of using computational resources to help with everyday tasks in the legal field. The legal domain deals mainly with textual data, and Artificial Intelligence (AI) has techniques that can aid in daily tasks by speeding up the process. However, datasets from the legal domain required by some AI techniques are scarce and difficult to obtain as they need labels from experts. This article presents four datasets from the legal domain, two with corpus of documents and metadata but unlabeled, and another two labeled with a heuristic aiming at its use in textual semantic similarity tasks.*

Resumo. *O judiciário brasileiro possui uma grande carga de trabalho, o que acaba acarretando um longo tempo para conclusão dos processos judiciais. Diversas iniciativas de digitalização têm surgido, abrindo a possibilidade do uso de recursos computacionais no auxílio das tarefas cotidianas do domínio jurídico. O domínio jurídico lida, em sua maioria, com dados textuais e a Inteligência Artificial tem técnicas que podem ajudar a apoiar as tarefas cotidianas, dando maior celeridade ao processo. No entanto, conjuntos de dados do domínio jurídico necessários para algumas técnicas atuais de Inteligência Artificial são escassos e de difícil obtenção, uma vez que requerem anotações por parte de especialistas. Este artigo apresenta quatro conjuntos de dados do domínio jurídico, dois com corpus de documentos e alguns metadados mas sem rótulo, e outros dois anotados com uma heurística visando seu uso na tarefa de similaridade semântica textual.*

1. Introdução

Segundo a edição de 2021 do Relatório *Justiça em Números*[1], o Poder Judiciário Brasileiro encerrou o ano de 2020 com 75,3 milhões de processos em andamento, dos quais 25,8 milhões foram novos casos abertos no ano de referência. Entre as causas para uma quantidade tão grande de casos não resolvidos estão uma força de trabalho humana que é insuficiente para atender às demandas, e uma Legislação extensa, que conta com mais

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. A pesquisa foi também apoiada parcialmente por CNPq e FAPERJ.

[1]https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/

de 34.000 leis[2]. Além disso, o Brasil é o sexto país mais populoso do mundo, com uma população estimada de 213 milhões de habitantes em 2020[3], que grosseiramente reflete o número de possíveis litigantes. Por outro lado, o Relatório *Justiça em Números* indica um aumento na produtividade do Poder Judiciário Brasileiro, induzido pela prioridade do judiciário em reduzir a quantidade de processos em andamento, pois se o sistema continuar nesse ritmo pode-se levar mais de 50 anos para zerar o estoque de processos.

A digitalização do estoque de processos[4] é uma das iniciativas para agilizar o desafogo do sistema judiciário. Essa digitalização também torna possível o uso de recursos computacionais que facilitem a análise de processos, e, em alguns casos, automatizem tarefas repetitivas que envolvem o processamento de um grande volume de documentos. A automatização de tarefas no contexto jurídico tem sido apoiada por técnicas de Inteligência Artificial adotadas por diversos órgãos legais[5]. Exemplos de uso de técnicas de Inteligência Artificial no contexto jurídico incluem a tarefa de Classificação de documentos [Dal Pont et al. 2020] da sub-área de Aprendizado de Máquina, e a tarefa de Similaridade Semântica Textual [de Oliveira and Nascimento 2022] da sub-área de Processamento de Linguagem Natural (PLN).

Este segundo exemplo tem como foco a busca por processos semelhantes. Essa tarefa é um exemplo de tarefa realizada exaustivamente no domínio jurídico e passível de automatização. Tal busca se torna importante, pois processos anteriores podem servir de base para um novo processo. O resultado dessa busca é útil tanto para o litigante, que pode considerar processos semelhantes como base para sua petição, quanto para o julgador conseguir agilizar a análise do processo. Salienta-se também que esse tipo de busca se mostra mais eficaz quando se consideram os componentes textuais do caso, sobretudo quando se considera a *similaridade semântica* entre os processos.

Nota-se que automatização de tarefas no cenário jurídico é importante para diminuição do estoque de processos não resolvidos, e, dessa forma, as técnicas de Inteligência Artificial podem ser grandes aliadas nesse processo. Contudo, a experimentação de métodos de Inteligência Artificial que apresentam resultados relevantes em outros domínios, e até mesmo a proposta de novos métodos específicos para o domínio jurídico se encontra bastante relacionada à disponibilidade de conjunto de dados do domínio jurídico. Além disso, a automatização de tarefas específicas necessitam de conjuntos de dados especializados para tornar possível o uso de métodos de Inteligência Artificial mais sofisticados. Ainda, diversas tarefas do domínio jurídico, incluindo a recuperação de documentos similares, requer *conjuntos de dados anotados*. Entretanto, a tarefa de anotação é particularmente desafiadora para o domínio jurídico, pois requer especialistas no assunto que entendam o contexto e o vocabulário utilizado para descrever os processos.

Este artigo apresenta quatro conjuntos de dados do domínio jurídico em português. Os conjuntos de dados *Votos de TCU* e *Acórdãos do STJ* contém os textos e metadados

[2]https://educacao.uol.com.br/disciplinas/cidadania/legislacao-mais-de-34-mil-leis-ordenam-a-vida-dos-brasileiros.htm

[3]https://paises.ibge.gov.br/\#/mapa/ranking/brasil?indicador=77849&tema=5&ano=2020

[4]https://www.conjur.com.br/2021-set-22/justica-receber-apenas-processos-eletronicos-partir-marco-2022

[5]https://www.cnj.jus.br/justica-4-0-inteligencia-artificial-esta-presente-na-maioria-dos-tribunais-brasileiros/

relativos aos referidos documentos extraídos dos portais de ambos os órgãos, mas sem anotação. Também são apresentados os conjuntos de dados *Votos de TCU para Similaridade Semântica Textual* e *Acórdãos do STJ para Similaridade Semântica Textual*, gerados a partir dos dois conjuntos iniciais, mas, com a utilização de uma heurística proposta neste artigo para anotar os documentos que são similares uns aos outros. O artigo está organizado em quatro seções além desta introdução. A Seção 2 cita outros conjuntos de dados do domínio jurídico em português. A Seção 3 apresenta os conjuntos de dados *Votos de TCU* e *Acórdãos do STJ*. Posteriormente, a Seção 4 apresenta os conjuntos de dados *Votos de TCU para Similaridade Semântica Textual* e *Acórdãos do STJ para Similaridade Semântica Textual*, bem como a heurística utilizada para suas gerações. E, finalmente, a Seção 5 apresenta as considerações finais.

2. Trabalhos Relacionados

A literatura não apresenta conjuntos de dados para a tarefa de Similaridade Semântica Textual com dados jurídicos em português como é proposto neste artigo, no entanto, têm-se alguns conjuntos de dados jurídicos, sejam contendo apenas *corpus* de dados textuais, como também dados com anotações para abordar outras tarefas específicas. O *Iudicium Textum Dataset* [Willian Sousa and Fabro 2019] é um *corpus* de documentos jurídicos contendo 41.353 documentos relativos a acórdãos do Supremo Tribunal Federal (STF) publicados entre os anos de 2010 a 2018. De Oliveira [de Oliveira and Júnior 2017] também apresenta um *corpus* de documentos jurídicos contendo jurisprudências do Supremo Tribunal do Estado de Sergipe, formado por quatro coleções: a) acórdãos do Tribunal de Justiça(181.994 documentos); b) decisões monocráticas do Tribunal de Justiça (37.142 documentos); c) acórdãos de Juizados Especiais (37.161 autos); e d) monocrático decisões de Juizados Especiais (23.151 documentos). Tanto o *Iudicium Textum Dataset* [Willian Sousa and Fabro 2019] quanto os *corpus* disponibilizados por De Oliveira [de Oliveira and Júnior 2017] são conjuntos de dados não rotulados.

Para a tarefa de classificação textual, o *VICTOR* [Luz de Araujo et al. 2020] é um conjunto de dados com mais de 692 mil documentos do Supremo Tribunal Federal anotados manualmente por uma equipe de especialistas para as tarefas de classificação do tipo de documento e atribuição de tema de processo. Para a tarefa de reconhecimento de entidades nomeadas (NER, do inglês *Named Entity Recognition*), o *LeNER-BR* [de Araujo et al. 2018] é um conjunto de dados com 70 documentos de tribunais judiciais e leis brasileiras. Os dados são anotados com entidades de uso genérico e com entidades específicas do conhecimento jurídico, a saber, "Legislação" para leis, e "Jurisprudência" para decisões judiciais resultantes de processos judiciais. Também construído para a tarefa de NER, o *UlyssesNER-Br* [Albuquerque et al. 2022] é um conjunto de dados criado no âmbito da Câmara de Deputados que também contém entidades de domínio genérico acrescido das entidades específicas do domínio jurídico em questão, como "Fundamento" e "Produto de Lei". O o *UlyssesNER-Br* é dividido em dois subconjuntos: *PL-corpus*, com projetos de lei que são documentos públicos, com 9.526 sentenças e *ST-corpus*, que são solicitações de trabalho, documentos internos com 790 sentenças.

3. *Corpus* de Votos do TCU e Acórdãos do STJ

Os dois primeiros conjuntos de dados apresentados neste artigo, *Votos do TCU* e *Acórdãos do STJ*, foram produzidos a partir de textos provenientes de acórdãos do *Superior Tribunal*

de Justiça (STJ) e votos do *Tribunal de Contas da União (TCU)*. O STJ e o TCU são órgãos colegiados, ou seja, órgãos onde a decisão é proferida após avaliação e consenso dos membros responsáveis. Os *acórdãos* são textos de julgamentos de órgãos colegiados, que abrangem somente os pontos principais de uma discussão. Por outro lado, um *voto*, no contexto dos órgãos colegiados, é a exposição, avaliação e opinião sobre a decisão a ser tomada para um caso em questão realizada pelo membro responsável, denominado relator[6].

A particularidade desses dados consiste em serem precedentes de *jurisprudências* usadas pelos órgãos. Jurisprudências são entendimentos adotados por órgãos jurídicos que orientam qual deve ser a decisão para um determinado assunto. Esses entendimentos são formulados a partir da análise de decisões anteriores sobre o mesmo assunto, ou seja, precedentes, e visam uniformizar as decisões e dar celeridade aos processos de assuntos recorrentes. Os conjuntos de dados apresentados aqui podem servir para o *Fine-Tuning* [Howard and Ruder 2018] de Modelos de Linguagem [Chen and Goodman 1999] que são usados como base para tarefas de Processamento de Linguagem Natural como classificação textual.

Os textos foram obtidos a partir de uma rotina de raspagem de dados dos sítios dos respectivos órgãos. Após a execução da rotina de raspagem de dados, um pré-processamento para eliminação de registros com informações nulas ou registros duplicados foi realizado. Como visto na Tabela 1, após o pré-processamento, o conjunto de dados provenientes de acórdãos do STJ tem uma quantidade de registros, representada pela linha *Acórdaos*, bem superior ao conjunto de dados de votos do TCU, onde o total de registros é representado pela linha *Votos*, além da superioridade de jurisprudências representadas. A Tabela 1 também traz informações sobre categorizações utilizadas para os dados de cada órgão e estão exibidas em ordem de superioridade hierárquica, ou seja, nos dados do TCU, um voto tem um Subtema que pertence a um Tema que por sua vez pertence a uma Área.

Tabela 1. Características dos dados do *STJ* e *TCU* usados nos experimentos.

TCU		**STJ**	
Votos	371	Acórdãos	7403
Jurisprudências	44	Jurisprudências	1458
Áreas	4	Matérias	7
Temas	27	Naturezas	68
Subtemas	38		

Os conjunto de dados apresentados neste artigo são disponibilizados em formato *CSV* na URL `https://osf.io/k2qpx/`. O conjunto de dados *Votos do TCU* possui os atributos: `AREA`, `TEMA`, `SUBTEMA`, `ENUNCIADO`, `PROCESSO`, `ANO`, `TIPO_PROCESSO`, `RELATOR` e `VOTO`. O atributo `ENUNCIADO` define a jurisprudência a qual um VOTO, que é um precedente, está associado. Já o conjunto de dados *Acórdãos do STJ* possui os atributos `MATERIA`, `NATUREZA`, `TEMA`, `PROCESSO`,

[6] `https://www.congressonacional.leg.br/legislacao-e-publicacoes/glossario-legislativo/-/legislativo/termo/relator_quanto_ao_papel`

RELATOR, ORGAO, DATA_JULGAMENTO, DATA_PUBLICACAO e EMENTA. Neste caso, o atributo TEMA define a jurisprudência a qual uma EMENTA, que é um precedente, se encontra associada.

A seguir são apresentados os gráficos que exploram a composição dos conjuntos de dados supracitados. A Figura 1 apresenta um histograma do conjunto de dados *Votos do TCU*, indicando que neste conjunto de dados as jurisprudências têm, em média, entre sete e oito votos precedentes. Por outro lado, o histograma apresentado na Figura 2 relacionada aos *Acórdãos do STJ* mostra que a maioria das jurisprudências têm entre cinco e seis acórdãos precedentes.

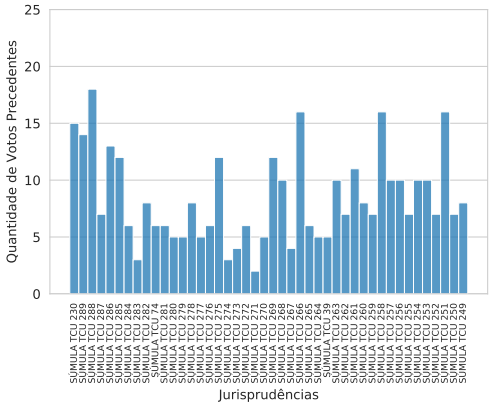

Figura 1. Histograma Precedentes x Jurisprudências dos *Votos do TCU*.

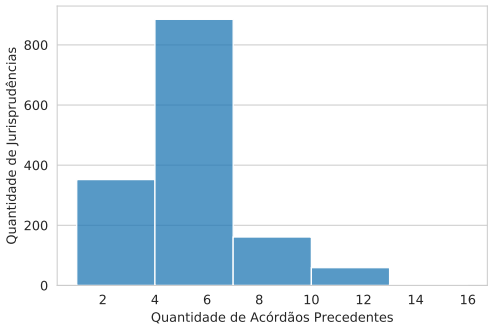

Figura 2. Histograma Precedentes x Jurisprudências dos *Acórdãos do STJ*.

A Figura 3 mostra que os precedentes do conjunto de dados *Votos do TCU* são, em sua maioria, da área de Licitação, seguida da área Pessoal. Além disso, as áreas Licitação e Pessoal também são as que apresentam a maior dispersão dos precedentes entre diferentes Temas. Já com relação aos *Acórdãos do STJ*, a Figura 4 mostra que os precedentes são, em sua maioria , das Matérias: Direito Administrativo, Direito Civil e Direito Penal. A dispersão por Natureza dos precedentes nessas três Matérias também são maiores que nas demais.

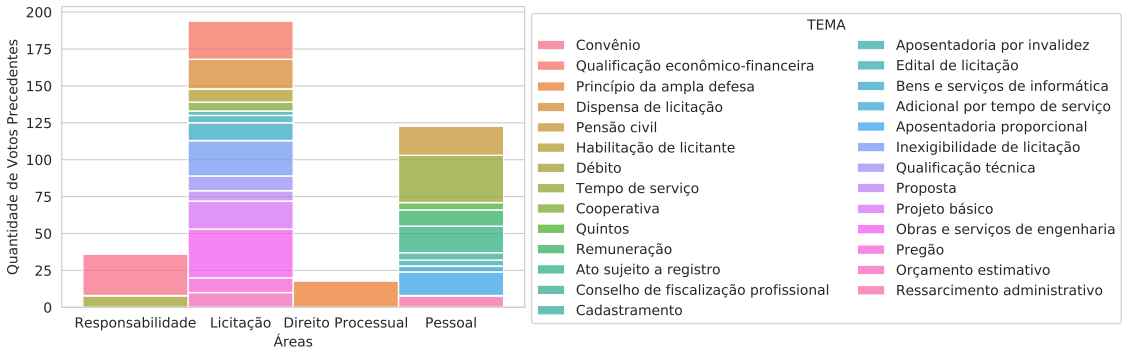

Figura 3. Histograma VOTO x AREA x TEMA dos *Votos do TCU*

A nuvem de palavras dos *Votos do TCU* na Figura 5 destaca palavras como OBRA, SERVIÇO, CONTRATO e LICITAÇÃO como as mais frequentes

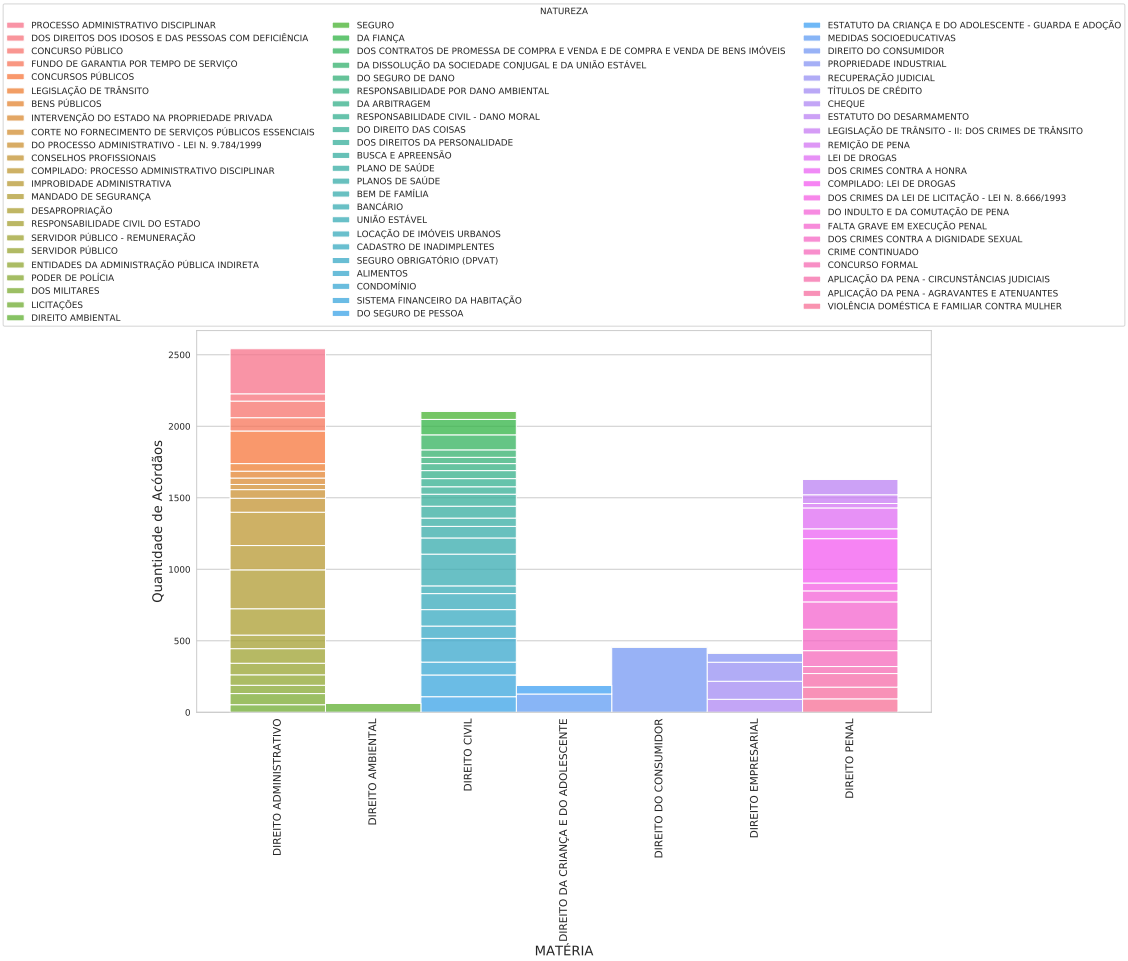

Figura 4. Histograma ACÓRDÃO x MATERiA x NATUREZA dos *Acórdãos do STJ*

nos precedentes do conjunto de dados. Enquanto isso, a Figura 6 destaca termos como `RECURSO ESPECIAL`, `AGRAVO REGIMENTAL`, `HABEAS CORPUS`, `PROCESSUAL CIVIL` e `AGRAVO INTERNO` como os mais frequentes no conjunto de dados *Acórdãos do STJ*.

Figura 5. Nuvem de palavras dos precedentes do *Votos do TCU*.

Figura 6. Nuvem de palavras dos precedentes do *Acórdãos do STJ*

O histograma apresentado na Figura 7 indica que a maioria dos precedentes do conjunto de dados *Votos do TCU* possui até 20.000 palavras. Nesse caso, as palavras são definidas a partir da quebra por espaços nos textos dos precedentes. Ao considerar os *Acórdãos do STJ*, a Figura 8 mostra que a maioria dos precedentes desse conjunto de dados possui até 500 palavras.

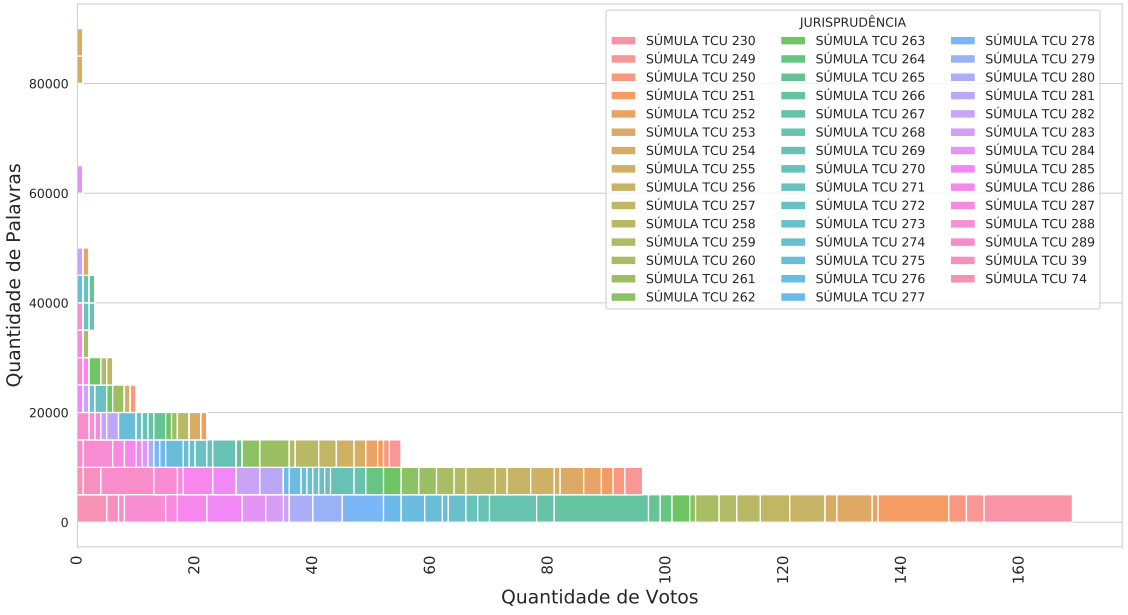

Figura 7. Histograma Palavras x Precedentes dos *Votos do TCU*

4. Geração de Conjuntos de Dados para Similaridade Semântica Textual a partir de Heurística

Dada a importância da recuperação de processos similares no contexto jurídico, e a ausência de conjuntos de dados que auxiliem no processo de treinamento de modelos para a tarefa de Similaridade Semântica Textual, este artigo também propõe os conjuntos

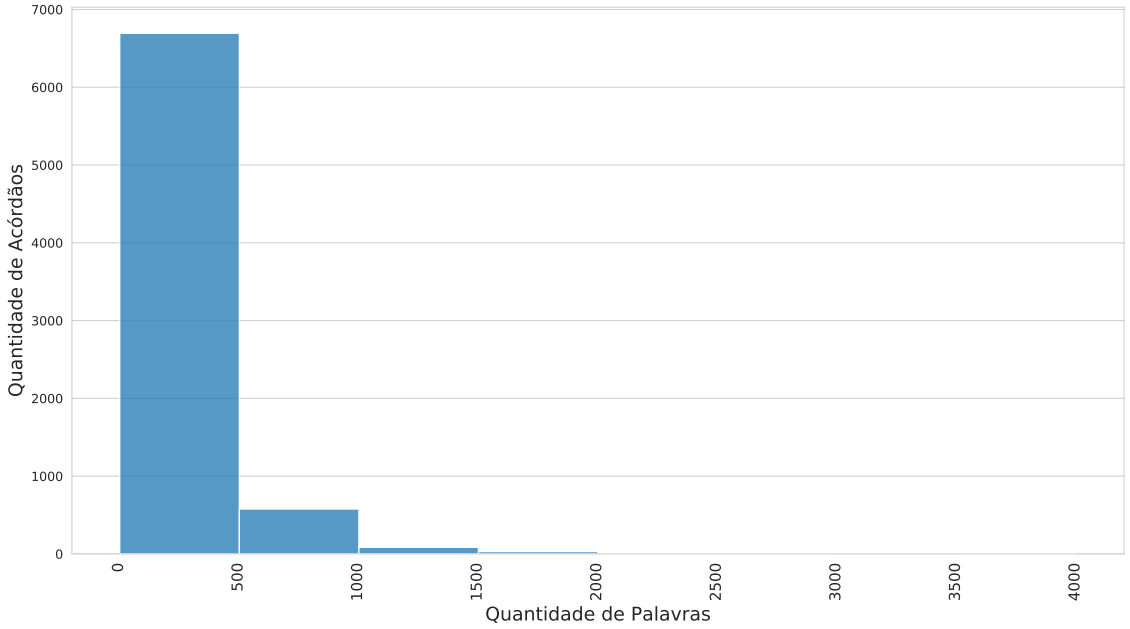

Figura 8. Histograma Palavras x Precedentes dos *Acórdãos do STJ*.

de dados *Votos de TCU para Similaridade Semântica Textual* e *Acórdãos do STJ para Similaridade Semântica Textual*.

Os dois conjuntos de dados apresentados nesta seção foram construídos a partir dos conjuntos de dados apresentados da Seção 3 deste artigo, e sintetizados para tarefa de Similaridade Semântica Textual (SST) [Fonseca et al. 2016]. Um conjunto de dados para a tarefa de SST é composto por um par de textos e uma pontuação associada à similaridade semântica entre os dois textos. Quanto maior a pontuação, maior a similaridade semântica entre os textos. Conjuntos de dados para a tarefa de SST são de elevado custo de elaboração, pois demandam a mobilização de usuários humanos para a anotação dos dados e em muitos casos precisam ser especialistas do domínio dos dados.

Para aliviar a necessidade de anotadores humanos na criação do conjunto de dados de SST a partir dos dados do STJ e TCU, utilizou-se um processo automático baseado em uma heurística proveniente dos metadados dos textos. Para construir o conjunto SST a partir dos dados do STJ, as seguintes etapas devem ser seguidas:

1. Gerar pares entre acórdãos de uma mesma Jurisprudência e atribuir a cada par uma pontuação com um valor base de 4.5, acrescido de um ruído que obedece uma distribuição normal entre todos os pares gerados.
2. Gerar pares entre acórdãos de uma mesma Natureza e atribuir a cada par uma pontuação com um valor base de 3, acrescido de um ruído que obedece uma distribuição normal entre todos os pares gerados.
3. Gerar pares entre acórdãos de diferentes Matérias e atribuir a cada par uma pontuação com um valor base de 0.5, acrescido de um ruído que obedece uma distribuição normal entre todos os pares gerados.
4. Gerar o conjunto final com uma combinação balanceada dos subconjuntos gerados anteriormente.

A heurística por trás do processo de geração do conjunto de dados para SST com

os acórdãos do STJ é baseada no conceito que acórdãos que serviram como precedentes para uma mesma Jurisprudência guardam uma grande similaridade intrínseca. Por outro lado, acórdãos que são de uma mesma Natureza, mas não são precedentes de uma mesma Jurisprudência guardam uma relação de similaridade menos intensa. Por fim, acórdãos que versam sobre diferentes Matérias são bem diferentes. Para o último caso, optou-se por não utilizar dados de diferentes jurisprudências, pois apesar de não serem precedentes de uma mesma Jurisprudência, documentos podem ter uma mesma Natureza e desse modo guardar algum grau de similaridade. A escolha de três valores de base, $(4.5, 3, 0.5)$, para a geração do graus de similaridade teve como intuito simular pares de documentos com similaridade alta, neutra ou dissimilaridade. A inclusão de ruídos seguindo uma distribuição normal, visou simular a incerteza e diferença entre anotações que fossem realizadas por um anotadores manuais.

Para construir um conjunto de SST partir dos dados de votos do TCU, foi empregado um processo similar ao utilizado na construção dos dados de SST do STJ, exceto por diferenças nos metadados utilizados. Assim, para construir o conjunto SST a partir dos dados do TCU, as seguintes etapas devem ser seguidas:

1. Gerar pares entre votos de uma mesma Jurisprudência e atribuir a cada par uma pontuação com um valor base de 4.5, acrescido de um ruído que obedece uma distribuição normal entre todos os pares gerados.
2. Gerar pares entre votos de uma mesma Área e Tema e atribuir a cada par uma pontuação com um valor base de 3, acrescido de um ruído que obedece uma distribuição normal entre todos os pares gerados.
3. Gerar pares entre votos de diferentes Áreas e atribuir a cada par uma pontuação com um valor base de 0.5, acrescido de um ruído que obedece uma distribuição normal entre todos os pares gerados.
4. Gerar o conjunto final com uma combinação balanceada dos subconjuntos gerados anteriormente.

Uma diferença marcante na construção dos dados para SST do TCU é o segundo subconjunto utilizar votos de uma mesma Área e Tema, isso porque, para os dados raspados do TCU, existem Tema com uma mesma nomenclatura, mas pertencentes a Àreas diferentes. O conjunto de dados TCU tem um total de 4.843 tuplas, enquanto o conjunto de dados do STJ tem um total de 51.437 tuplas. Após o processo automático de geração de pares e pontuação associados, e do balanceamento entre os subconjuntos gerados em cada passo, ainda foram gerados estratos no conjunto de dados, dividindo-o em `TREINAMENTO`, `TESTE` e `VALIDAÇÃO`, guardando a proporção dos pares por intervalo de similaridade. Desse modo, o conjunto de dados para SST com Votos do TCU ficou dividido em 3.389 para treinamento, 438 para validação e 1.016 para teste. Já o conjunto de dados para SST com os Acórdãos do STJ ficou dividido em 36.010 para treinamento, 4.613 para validação e 10.814 para teste.

5. Considerações Finais

Este artigo apresenta quatro conjuntos de dados do domínio jurídico, a partir de dados coletados dos portais do Tribunal de Contas da União e Superior Tribunal de Justiça. Os conjuntos de dados *Votos do TCU* e *Acórdãos do STJ* são relativos a precedentes de

jurisprudências, e além do conteúdo textual dos precedentes também possui metadados relacionados a categorizações que os documentos têm no contexto dos respectivos órgãos.

Também são disponibilizados os conjuntos de dados *Votos de TCU para Similaridade Semântica Textual* e *Acórdãos do STJ para Similaridade Semântica Textual* que foram construídos a partir dos precedentes coletados e a aplicação de uma heurística. Além da disponibilização dos conjuntos de dados, este artigo também tem como contribuição análises exploratórias desses conjuntos, bem como a criação de uma heurística para a anotação automática dos conjuntos de dados para a tarefa de Similaridade Semântica Textual. Os conjuntos de dados podem ser obtidos em `https://osf.io/k2qpx/`.

Referências

Albuquerque, H., Costa, R., Silvestre, G., Souza, E. P., Félix, N., Vitório, D., and Carvalho, A. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

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.

de Oliveira, R. S. and Nascimento, E. G. S. (2022). Brazilian court documents clustered by similarity together using natural language processing approaches with transformers.

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.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Luz de Araujo, P. H., de Campos, T. E., Ataides 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.

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.