PLN e Segurança Jurídica
Identificação de divergências jurisprudenciais com Processamento de Linguagem Natural

Marcella Queiroz de Castro[1]**; Ana Régia Mendonça**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Brasília (IFB)
Brasília, DF – Brasil
marcella.castro@estudante.ifb.edu.br

Abstract. *This paper proposes the use of Natural Language Processing (NLP) and Machine Learning techniques to identify judicial rulings that are divergent to the majority understanding of the subject in Brazilian courts, aiming to enhance judicial security. The methodology used includes data preprocessing of a summary of the rulings, the use of the Word2Vec neural network technique for word embedding, and the analysis of 3.165 court decisions via k-means clustering to identify semantic similarities and divergences. Specific examples of jurisprudential divergences are presented, demonstrating how technology can assist in the uniformity of judicial decisions.*

Resumo. *Este artigo propõe a utilização de técnicas de Processamento de Língua Natural (PLN) e Aprendizado de Máquina para identificar divergências jurisprudenciais nos tribunais brasileiros, visando aumentar a Segurança Jurídica. A metodologia inclui a extração de 3.165 acórdãos do Tribunal de Justiça de Minas Gerais, o pré-processamento dos dados, uso da técnica de Word2Vec para definição de embeddings para cada palavra e análise de acórdãos via clustering para identificar semelhanças textuais e divergências nas decisões judiciais. Exemplos específicos de divergências jurisprudenciais são apresentados, demonstrando como a tecnologia pode auxiliar na uniformização das decisões.*

1. Introdução

Atualmente, no caso brasileiro, espera-se de uma IA (Inteligência Artificial) no judiciário o auxílio aos servidores para superar o enorme acervo de processos que aguardam julgamento, almejando celeridade na tramitação processual. Conforme o relatório Justiça em Números do Conselho Nacional de Justiça (CNJ), em 2023, os 92 tribunais do país possuíam pendentes de julgamento 81,4 milhões de casos [Brasil 2023], um cenário de contingente de processos alarmante, em que se faz essencial o desenvolvimento de ferramentas que não somente assegurem celeridade do processo, mas que garantam a estabilidade, previsibilidade e coerência das decisões proferidas pelos tribunais.

Com o intuito de incrementar a Segurança Jurídica das Cortes brasileiras, propõe-se a utilização de técnicas de PLN e Aprendizado de Máquina para apontar divergências jurisprudenciais. Para tanto, o trabalho realizou a extração de acórdãos do Tribunal de Justiça de Minas Gerais (TJMG) com python[1], o pré-processamento e passagem desses pelo processo de definição de *Embbeds* via a técnica de rede neural chamada *Word2Vec* [Mikolov et al. 2013].

[1]O código de raspagem e análise dos dados pode ser encontrado neste link: https://colab.research.google.com/drive/1GLp9jUlqMLKDf8P6OPVo-QkdFcp9KCp1?usp=sharing.

Uma vez que os textos passaram pela etapa de extração de características, tornou-se possível a construção de ferramenta de análise e classificação de acórdãos em *clusters*, capazes de identificar semelhanças entre as decisões por método de Similaridade Textual Semântica (STS) e agrupá-las de acordo com o grau de similaridade fático e decisório que possuem, apontando possíveis divergências jurisprudenciais no conjunto de decisões em estudo.

2. Conceitos Jurídicos

Um acórdão é todo pronunciamento judicial proferido por um órgão colegiado [Didier 2019] e está, usualmente, contido em um documento de cinco partes, das quais a mais relevante para este estudo é a ementa. A ementa é um resumo da decisão, mostrando quais argumentos foram acolhidos e a decisão tomada pelos magistrados. Estudos analisados [Wilton 2022, Gomes 2021] e a empiria da prática judicial por aplicadores do Direito demonstram que a ementa é suficiente para identificar os termos principais do acórdão, razão pela qual somente essa parte do todo foi selecionada para a análise por PLN. Ademais, são os acórdãos que possuem a tarefa legal de trazer estabilidade ao sistema de precedentes judiciais[2], demonstrando-se como adequado o estudo de acórdãos para a averiguação da Segurança Jurídica de uma Corte.

3. Análise das Ementas

3.1. Pré-processamento e extração de características

Assim como em trabalhos correlatos [Ciurlino 2021, Polo et al. 2021], as decisões judiciais utilizadas para análise foram obtidas pela extração automatizada a partir de *scripts*, almejando viabilizar a execução em larga escala e para outros temas além do escolhido nesta pesquisa. Para selecionar os dados foi feita pesquisa no site do Tribunal de Justiça de Minas Gerais [3] pelas palavras chave "direito do consumidor" e "apelação" nas ementas dos acórdãos. O período de busca escolhido foi de 1º de janeiro de 2021 até 31 de dezembro de 2023, resultando em 3.163 acórdãos extraídos. Uma vez selecionado o *corpus*, é preciso executar a limpeza do texto escolhido. No caso da busca por divergências jurisprudenciais, o contexto jurídico e as práticas dentro de textos desse tipo foram determinantes para definição das etapas de pré-processamento textual.

Para além do pré-processamento e da normalização, os textos passaram por outras etapas prévias que merecem destaque. Palavras comuns do discurso jurídico foram removidas, como "excelência" e "douto", e, além dos n-gramas existentes dentro do modelo de linguagem adotado, foram adicionados outros específicos do *corpus*, como "responsabilidade_fornecedor". Inspirado no trabalho correlato de Martins [2018], o qual incluiu formas de pré-processamento como a remoção de textos entre parênteses a partir de análises sobre a forma de redação de documentos jurídicos, este trabalho, de uma análise do *corpus* escolhido, verificou que a prática de escrita dos desembargadores do TJMG inclui colocar entre parênteses ou aspas referências a outros acórdãos citados, a artigos de lei ou a súmulas de outros tribunais, informações estas que também foram retiradas por serem consideradas irrelevantes para a análise computacional.

Quanto à vetorização das palavras, estudos que argumentam pela utilização conjunta de um modelo Word2Vec [Mikolov et al. 2013] com *corpus* de treinamento jurídico

[2]Art. 926: "Os tribunais devem uniformizar sua jurisprudência e mantê-la estável, íntegra e coerente" [Brasil 2015]

[3]O site utilizado foi: https:www5.tjmg.jus.brjurisprudenciaformEspelhoAcordao.do

para alcançar o melhor desempenho na análise de similaridade de textos [Xia et al. 2019] e também que destacam como é essencial o treinamento do modelo em uma base jurídica para realização de tarefas específicas de cada tipo de domínio de conhecimento [Gomes 2021] foram norteadores para decidir o tipo de modelo, arquitetura e *corpus* de treinamento.

Após pesquisa sobre os modelos disponíveis para utilização, optou-se pelo modelo LegalNLP [Polo et al. 2021]. Os pesquisadores desse projeto apresentaram um modelo já pré-treinado com textos jurídicos, mais especificamente, com decisões judiciais de Cortes Estaduais, de forma que a ferramenta está alinhada com os textos que serão analisados por similaridade textual neste artigo. Dentre as versões do modelo Word2Vec do projeto, escolheu-se a arquitetura Skip-gram [Mikolov 2013], com *size*=200, *window*=15 e *epochs*=20.

3.2. Análise de Similaridade Textual Semântica

Em estudos da área foi difundido o cálculo de similaridade por 'soft cossine' em que fora feita modificação no cálculo da similaridade por cosseno para passar a considerar uma matriz de similaridade das características, a qual acrescenta um peso que aproxima ou afasta os vetores a depender da similaridade de cada palavra do documento [Sidorov 2014].

Essa forma de cálculo e abordagem para o problema da similaridade textual semântica se demonstra útil em cenários onde palavras diferentes podem ter significados semelhantes, como nos textos jurídicos. Por exemplo, "provimento" e "desprovimento" são palavras que ocupam espaços semelhantes dentro da vetorização do modelo pois, ainda que sejam opostas, seguindo a lógica do modelo Word2Vec, costumam vir acompanhadas de outras palavras idênticas. Para combater essa particularidade do contexto jurídico, o cálculo de similaridade por 'soft cossine' se demonstra adequado, pois permite que características relacionadas contribuam para o cálculo de similaridade dos vetores, gerando uma medida de similaridade textual robusta e focada na semântica do texto.

3.3. K-Means

Para lidar com vetores de múltiplas dimensões e focar somente nas mais significativas foi aplicada a técnica de t-SNE (t-Distributed Stochastic Neighbor Embedding) [Maaten 2008], visando permitir uma melhor representação bidimensional do objeto de estudo. A redução de dimensionalidade é etapa considerada relevante em estudos conexos [Wilton 2022, Xia et al. 2019, Martins 2018] e foi seguida neste artigo para garantir uma análise atualizada ao estado da arte da área.

Após a redução de dimensionalidade, o algoritmo do K-Means [Pajankar 2022] necessita que seja definido o número *k* ideal de *clusters*. Para tal, foi utilizado o coeficiente de *Silhouette* [Rousseeuw 1987], que é uma métrica interna que avalia a qualidade dos *clusters* formados pelo algoritmo K-Means, medindo tanto a coesão quanto a separação dos *clusters*. Os parâmetros para a construção da curva do coeficiente de *Silhouette* para os dados deste estudo seguiram as indicações de estudos símiles [Magalhães 2023, Lima 2022], com *n_init=10*, *random_state=0* e *max_iter=300*. O melhor coeficiente de *Silhouette* obtido foi 0.48 para um número de 15 grupamentos e o resultado encontra-se na Figura 1.

4. Divergências Jurisprudenciais

Com base na Figura 1, a investigação foi voltada para os *clusters* com pontos mais dispersos, como os grupos 3 e 11, na busca por divergências jurisprudenciais. Analisando-se manualmente os grupos citados, foram encontrados pontos de retoque na jurisprudência do Tribunal para que esta se torne mais coesa e uniforme.

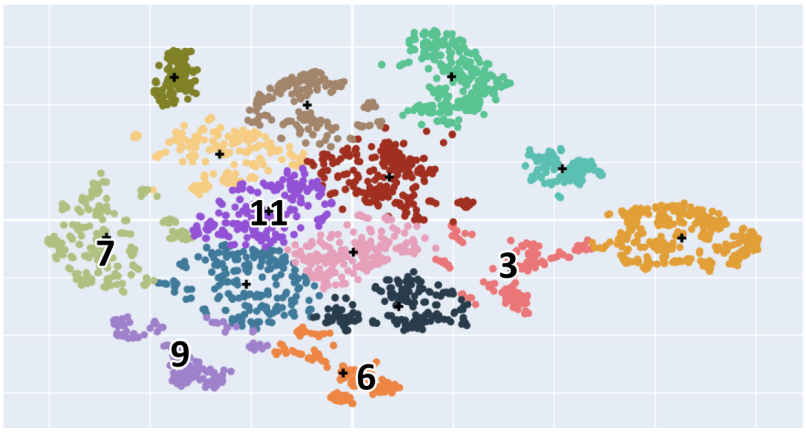

Figura 1. K-Means com 15 *clusters* realizado sobre as ementas selecionadas e reduzidas por técnica de t-SNE

No cluster 3, comparam-se as ementas proferidas nos processos 5000575.70.2020.8.13.0393 (ementa nº 719 da extração) e 5001067.08.2022.8.13.0453 (ementa nº 496 da extração). Em ambos os casos, trata-se de um contrato de crédito firmado sem o conhecimento da pessoa e que gerou descontos indevidos em seus benefícios previdenciários. Para o primeiro acórdão o desconto indevido só é capaz de gerar dano moral quando comprometer parcela significativa dos proventos. No outro acórdão, no entanto, foi proferido entendimento de que o desconto indevido gera dano moral por decorrência do próprio fato, independente do montante comprometido. De igual maneira, no cluster 11, encontram-se acórdãos que suscitam questionamentos sobre a uniformidade dos julgamentos. Em um caso, é determinada a repetição em dobro de valores que foram cobrados indevidamente a um consumidor, considerando somente a comprovação de que os valores não eram corretos de serem cobrados (Apelação Cível 5016736.60.2019.8.13.0145, ementa nº 2317 da extração). Noutro acórdão, do mesmo cluster, exige-se do consumidor que comprove a má-fé da empresa que realizou a cobrança indevida para ser ressarcido em dobro (Apelação Cível 6787163.48.2009.8.13.0024, ementa nº 3055 da extração).

Sem o objetivo de exaurir as divergências jurisprudenciais possíveis de serem encontradas nessa amostra de ementas, mostra-se interessante para a análise que, de uma verificação manual feita por pessoa formada na área jurídica, sem automatização de qualquer tipo, foi possível encontrar com relativa facilidade precedentes em que situações semelhantes se deparam com decisões de teor significativamente distinto. A incursão no conjunto de dados, embora manual, foi orientada pelas etapas prévias de PLN e Aprendizagem de Máquina, demonstrando a capacidade da tecnologia de auxiliar na identificação de uma quebra do princípio da Segurança Jurídica.

5. Conclusões

Inspirada na problemática de falta de Segurança Jurídica, o presente estudo buscou incorporar a ideia de qualidade, coesão e uniformidade - e não somente eficiência - para entender melhor os precedentes resolutivos de demandas reais, ajuizadas por cidadãos brasileiros que se socorreram ao judiciário para terem uma dor sanada. Com o apoio das ferramentas tecnológicas, foi possível observar coesão de assunto de acórdãos incluídos dentro de um mesmo cluster e, a partir de uma análise manual com enfoque jurídico, foram encontrados acórdãos em que as decisões podem ser apontadas como divergências jurisprudenciais dignas de retoques por parte da Corte Estadual. As decisões divergentes foram encontradas com maior

facilidade do que seria possível, mesmo para um especialista, caso tivesse que olhar o contingente completo de processos. O auxílio da análise construída está na possibilidade de tornar um problema relevante, mas que facilmente se perde entre os números, possível de ser encontrado e corrigido.

Referências

Brasil (2015). *Código de Processo Civil*. Senado Federal. Lei No 13.105, de 16 de março de 2015. Edição atualizada.

Brasil, C. N. d. J. (2023). Justiça em Números - 2023. Disponível em `https://t.ly/RVMXs`. Acessado em 05/08/2024.

Ciurlino, V. H. (2021). BertBR : a pretrained language model for law texts.

Didier, F. (2019). *Curso de Direito Processual Civil, Vol. 1: Introdução ao Direito Processual Civil, Parte Geral e Processo de Conhecimento*. JusPODIVM, 21 edition.

Gomes, T. A. (2021). Avaliação de técnicas de similaridade textual na uniformização de jurisprudência. Disponível `https://repositorio.unb.br/handle/10482/40798`. Acessado em 08/06/2024.

Lima, João Pedro e Costa, J. A. (2022). Comparing clustering techniques on brazilian legal document datasets. In *Hybrid Artificial Intelligent Systems*, pages 98–110, Cham. Springer International Publishing.

Maaten, Laurens van der e Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, Vol. 9:p. 2579 – 2605.

Magalhães, Dimmy; Pozo, A. e. M. S. (2023). Técnicas de aprendizado de máquinas aplicadas à classificação de decisões judiciais. *Revista de Estudos Empíricos em Direito*.

Martins, A. D. M. (2018). Agrupamento automático de documentos jurídicos com uso de inteligência artificial. Disponível `https://repositorio.idp.edu.br//handle/123456789/2635`. Acessado em 08/06/2024.

Mikolov, Tomas; Chen, K. C. G. e. D. J. (2013). Efficient estimation of word representations in vector space.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Pajankar, Ashwin e Joshi, A. (2022). *Hands-on Machine Learning with Python: Implement Neural Network Solutions with Scikit-learn and PyTorch*. Apress, Berkeley, CA.

Polo, F. M., Mendonça, G. C. F., Parreira, K. C. J., Gianvechio, L., Cordeiro, P., Ferreira, J. B., de Lima, L. M. P., do Amaral Maia, A. C., and Vicente, R. (2021). Legalnlp-natural language processing methods for the brazilian legal language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 763–774. SBC.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Sidorov, Grigori e Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, pages 324 – 344.

Wilton, Pabloe Vigneaux e Rover, A. J. (2022). Clustering of brazilian legal judgments about failures in air transport service: an evaluation of different approaches. *Artificial Intelligence and Law*, 30:21–57. Accepted: 8 April 2021 / Published online: 17 April 2021.

Xia, C., He, T., Li, W., Qin, Z., and Zou, Z. (2019). Similarity analysis of law documents based on word2vec. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 354–357.