

Inteligência Artificial e direito: o impacto transformador na organização de grandes acervos de textos jurídicos.¹

Marina Soares Marinho
Universidade Federal de Minas Gerais (UFMG e Jusbrasil)

 <https://orcid.org/0000-0002-4963-1570>
marinasoaresmarinho@gmail.com

Daniela Vianna (Jusbrasil)

 <https://orcid.org/0000-0003-2943-5211>
dvianna@gmail.com

Tiago Ferreira
Universidade de Brasília (Jusbrasil)

 <https://orcid.org/0009-0002-9208-6202>
hello@tiagoferreira.me

Altigran Soares da Silva
Universidade Federal do Amazonas (UFAM)

 <https://orcid.org/0000-0002-8992-495X>
altigran@jusbrasil.com.br

RESUMO

Os recentes avanços na área de inteligência artificial e de processamento de linguagem natural têm impulsionado diversas mudanças no campo jurídico. No Brasil, o Direito segue um movimento constante de modernização, principalmente direcionado à transparência e ao acesso à informação. O grande volume de documentos jurídicos abre espaço para o desenvolvimento e uso de

¹ Esse trabalho é parcialmente apoiado pelo Jusbrasil; pela FAPEAM no âmbito do POS-Programa GRAD 2022 e Projeto NeuralBond (Universal 2023 Proc. 01.02.016301.04300/2023-04); pela Coordenação do Financiamento de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES) código social 001; e pelo CNPq no âmbito do Projeto IAIA (406417/2022-9) e bolsa de produtividade para Altigran da Silva (307248/2019-4).

ferramentas inteligentes que busquem organizar e facilitar a gestão desse acervo. Neste trabalho, mostramos como o uso de modelos de linguagem em conjunto com técnicas de modelagem de tópicos são capazes de organizar e extrair conhecimento desses grandes acervos jurídicos, revelando temas que muitas vezes são implícitos e desconhecidos, o que traz benefícios para diversas aplicações, como a pesquisa por documentos similares e a recomendação de textos jurídicos.

Palavras-chave: processamento de linguagem natural; modelos de linguagem; modelagem de tópicos; direito; organização de acervos jurídicos.

Artificial Intelligence and law: the transformative impact on the organization of large collections of legal texts.

ABSTRACT

Recent advances in the area of artificial intelligence and natural language processing have driven several changes in the legal field. In Brazil, Law follows a constant movement of modernization, mainly aimed at transparency and access to information. The large volume of legal documents opens up space for the development and use of intelligent tools that seek to organize and facilitate the management of this collection. In this work, we show how the use of language models in conjunction with topic modeling techniques are capable

Keywords: natural language processing; language models; topic modeling; right; organization of legal collections.

Submissão em: 21/08/2023 | **Aprovação em:** 14/04/2024

1. INTRODUÇÃO

Inteligência Artificial (IA) é a área de estudo dedicada a criar sistemas que sejam capazes de replicar e potencializar tarefas realizadas tipicamente por humanos. Um dos seus principais motores de desenvolvimento é o aprendizado de máquina (*machine learning*), uma sub-área da IA que permite que as máquinas aprendam padrões e tomem decisões a partir de dados. Embora as aplicações que utilizam inteligência artificial com base no aprendizado de máquina não sejam novas, é inegável a sua expansão nos últimos anos. Parte desse movimento pode ser atribuída ao aumento da quantidade de dados disponíveis, mas não apenas. Houve aumento expressivo na capacidade de processamento desses dados.

No campo da linguagem escrita, que nos interessa para tratar do Direito, os avanços mais notáveis ocorreram na área do processamento de linguagem natural (PLN - *natural language processing*, NLP). PLN é uma aplicação da IA que permite que a máquina compreenda a linguagem humana, permitindo a extração e a geração de textos. Em sua versão mais moderna, se vale de representações vetoriais (*embeddings*) para mapear as palavras a partir de um modelo de linguagem, como é o caso dos modelos Word2Vec (Mikolov et al., 2013), Doc2Vec (Le; Mikolov, 2014), Bert (Devlin et al., 2019b) e GPT (Brown et al., 2020). Estes, por sua vez, utilizam estatística e probabilidade para prever palavras de um texto a partir do seu contexto. Eles se baseiam em redes neurais artificiais, que visam a simular o processo de pensamento humano e a forma como processamos a linguagem. Esses modelos “adquirem conhecimento” por meio do treinamento em grandes corpora de texto, acumulando informações sobre como as palavras são utilizadas em diferentes tipos de documentos. Atualmente, os grandes modelos de linguagem são capazes de realizar diversas tarefas de processamento de linguagem natural como por exemplo a tradução automática, a sumarização e a geração de textos.

Justamente pela característica predominante da linguagem escrita, os últimos avanços integrando modelos de linguagem e processamento de linguagem natural são promissores para o meio jurídico. O enorme volume de documentos com os quais os acadêmicos e os profissionais da área lidam abre espaço para que ferramentas que buscam organizar e facilitar a gestão desse acervo anunciem verdadeiras revoluções. Não seria a primeira vez. O mercado jurídico e até o Poder Judiciário seguem um movimento constante de modernização e investimento no acesso facilitado à informação. O desenvolvimento tecnológico permitiu, por exemplo, que o conteúdo dos autos processuais migrasse da Secretaria para as sedes dos escritórios de advocacia e até mesmo a residência dos jurisdicionados. A busca por jurisprudência, antes limitada ao acesso a dezenas de decisões, saiu das páginas de periódicos para a consulta rápida a uma base de milhões de documentos, filtrada a partir de palavras-chave. O processo de reconhecimento de entidades permitiu que as citações e intimações sejam mapeadas e informadas a partir de alertas por emails. Para mais, a tecnologia também permitiu que mesmo durante o isolamento provocado por uma pandemia global, por aproximadamente dois anos, a Justiça brasileira não apenas seguisse funcionando, como ampliasse a sua capacidade de julgamento.

Neste artigo vamos descrever um dos recentes casos de uso que desenvolvemos utilizando IA aplicada ao contexto do Direito: a utilização de modelos de linguagem e técnicas de modelagem de tópicos para a organização de grandes acervos de textos jurídicos.

Na primeira seção deste trabalho, Inteligência Artificial - Visão Geral e Estado da Arte do Processamento de Linguagem Natural, realizamos uma revisão abrangente das tecnologias e métodos mais recentes em IA, com uma ênfase especial naquelas que são relevantes para o campo do Direito e a modelagem de tópicos. Destacamos as principais pesquisas e avanços tecnológicos recentes nessa área e também identificamos as limitações atuais dessas tecnologias. Essa análise do estado da arte proporciona um contexto fundamental para se familiarizar com os projetos executados no campo jurídico,

apresentados na seção seguinte, e para compreender o estudo de caso apresentado ao final, demonstrando assim a contribuição significativa e relevante que ele traz para a área do Direito.

Já na seção seguinte, *A Utilização da IA no Direito*, analisamos com mais detalhes os impactos recentes da IA para o Direito, a partir da discussão sobre as aplicações práticas da IA. Abordamos principalmente o mencionado processo de modernização da gestão da informação no campo jurídico e a estreita relação entre o desenvolvimento tecnológico e a viabilização do acesso à Justiça, um direito fundamental, constitucionalmente garantido.

Finalmente, na terceira seção, *Estudo de Caso: Organização de Documentos Jurídicos*, apresentamos um estudo de caso no qual investigamos o emprego de técnicas de topificação visando à organização não supervisionada de um extenso volume de textos jurídicos. Inicialmente, descrevemos o desafio que essa abordagem visa superar, detalhando a implementação das técnicas utilizadas e os dados empregados. Em seguida, analisamos os resultados obtidos no estudo de caso, considerando suas inferências e comparando-os com a atualidade do estado da arte. A partir deste trabalho, conseguimos identificar conteúdo jurídico relacionado por temas mais gerais e mais específicos, o que permite recuperar documentos com maior precisão. Concluímos o estudo de caso ao explorar suas implicações potenciais, futuras aplicações e destacando a necessidade de pesquisas adicionais neste campo específico.

2. INTELIGÊNCIA ARTIFICIAL - VISÃO GERAL E ESTADO DA ARTE DO PROCESSAMENTO DE LINGUAGEM NATURAL

Quando Alan Turing propôs o famoso teste que levou seu nome, o objetivo foi criar uma forma de atestar a capacidade de pensamento de uma máquina. O critério de aprovação era simples: se a máquina conseguisse emular um diálogo a ponto de ser indistinguível de um ser humano, essa máquina teria passado no teste (Russel, 2010). Tal

capacidade conversacional é hoje um campo inteiro dentro da Inteligência Artificial, chamado de *Processamento de Linguagem Natural* (PLN) (Jojri, 2021).

Esse campo surgiu da combinação de esforços entre a Linguística e a Ciência da Computação para permitir a criação de programas de computador que pudessem se comunicar de forma fluida com seres humanos (Jojri, 2021). Atualmente, desdobramentos diretos e indiretos do PLN são ubíquos, podendo ser encontrados em sistemas de busca, aplicativos de navegação, serviços de e-mail, editores de texto, e os inúmeros assistentes virtuais disponíveis em smartphones ou aparelhos eletrônicos em nossas casas. O PLN é composto por várias áreas, tais como *Reconhecimento de Fala*, *Classificação de Texto*, *Tradução Automática*, *Autocorreção*, *Question Answering*², *Sumarização* e *Modelagem de Tópicos* – assunto deste artigo (Dieng, 2020; Jojri, 2021). Apesar de serem empregadas com finalidades distintas, todas essas áreas têm como objetivo projetar e construir sistemas que sejam capazes de entender a linguagem natural, falada ou escrita, ou de gerar conteúdo também em linguagem natural.

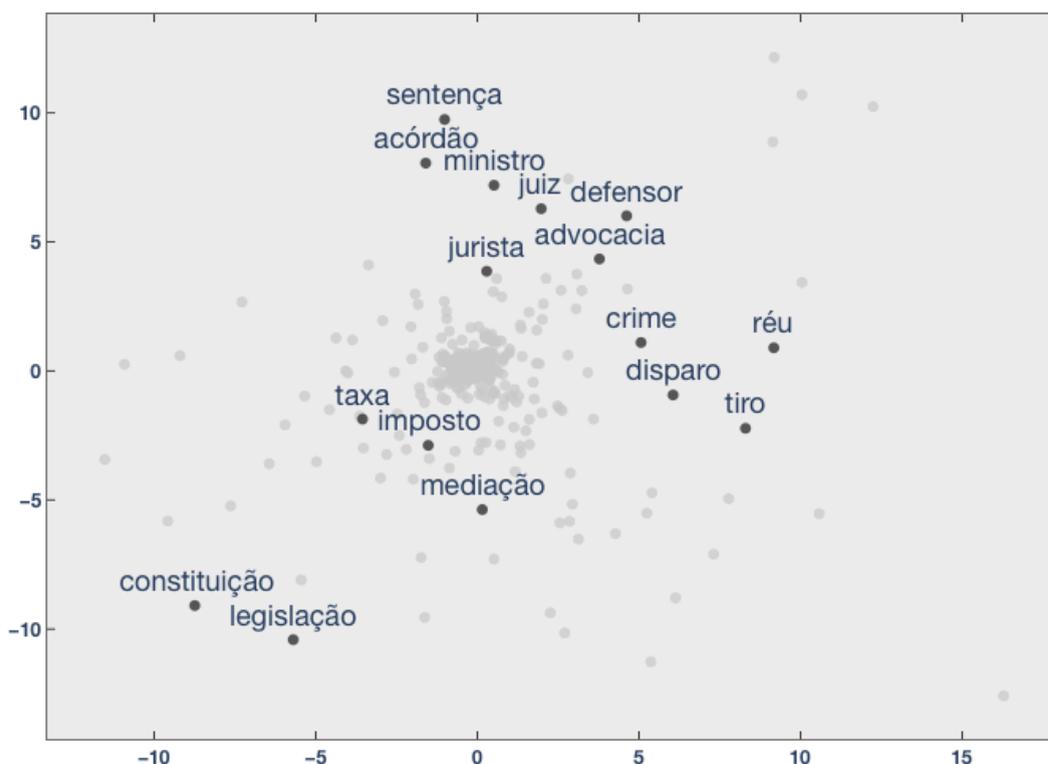
Em suas primeiras décadas, o PLN foi abordado de forma mais morfológica, ou seja, com regras rígidas e uma forte conexão com o campo da linguística. Isso requeria um profundo conhecimento da língua e definições inequívocas de regras sintáticas (Jojri, 2021). Mais tarde, com a evolução do poder computacional, essa abordagem foi gradualmente complementada com técnicas estatísticas, permitindo o desenvolvimento dos chamados "*word embeddings*", que são representações vetoriais baseadas na probabilidade da ocorrência das palavras de um vocabulário a partir do contexto onde elas ocorrem (Liu, Q, 2020).

Em sua essência, *embeddings* podem ser visualizados como palavras distribuídas num plano cartesiano, onde a proximidade espacial entre palavras diferentes também significa uma proximidade semântica entre elas (Lund, 1996). Os *embeddings* utilizados atualmente podem possuir centenas e até milhares de dimensões, mas para uma

² Sistemas de perguntas e respostas.

visualização simplificada, podemos utilizar uma técnica de *redução de dimensionalidade* para representá-los em duas ou três dimensões (Figura 1).

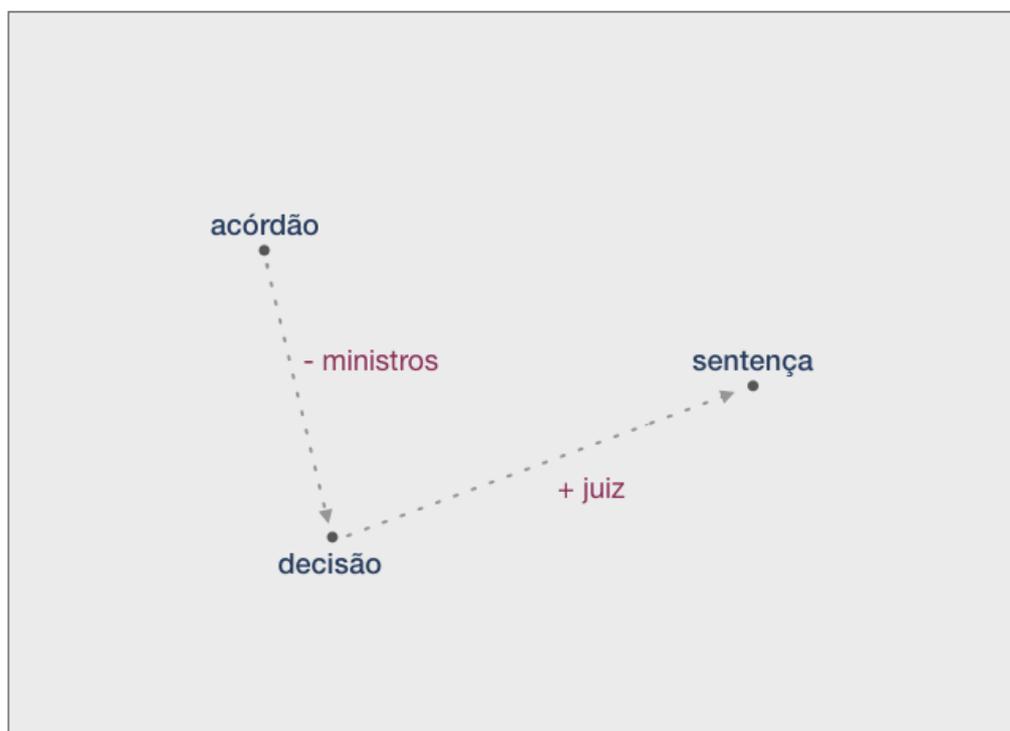
Figura 1 - Representação gráfica de *word embeddings* em duas dimensões



Fonte: elaborado pelos autores

Uma vez sendo possível representar textos de forma estritamente matemática, o PLN passou por um ponto de inflexão e avançou substancialmente o estado da arte na área (Hofmann, 2020). Com suas representações vetoriais, passamos a analisar palavras e textos aplicando a Álgebra Linear, e então simples analogias como "*Rei - Homem + Mulher = Rainha*" e "*Acórdão - Ministros + Juiz = Sentença*" passam a ser demonstráveis com um rigor matemático (Figura 2). E além de elementos em operações aritméticas, textos agora também podem ser usados como parâmetros em funções.

Figura 2 - Representação de operações aritméticas baseadas em *word embeddings*



Fonte: elaborado pelos autores

Até 2018, as técnicas de geração de *word embeddings* mais comuns, dentre elas Word2Vec e Doc2Vec, eram as chamadas "estáticas" e não tinham a capacidade de lidar com polissemia, homonímia e contextos diferentes (Wiedmann, 2019). Por exemplo, com essas técnicas, a palavra "pena" teria a mesma representação vetorial nos trechos "a pena foi fixada em 2 anos" e "o réu teve pena da vítima". Foi então com o surgimento dos chamados *word embeddings contextualizados*, como o Bert (Devlin, 2019), que a relação entre as palavras e seu contexto passou a ser levada em consideração na geração de cada *embedding*, permitindo um aumento na performance de várias tarefas do PLN (Liu, Q, 2020). Assim, a área passou por mais um ponto de inflexão (Ethayarajh, 2020), e essa nova arquitetura abriu caminho para aplicações mais robustas que, apesar de requerer

mais recursos computacionais (ARORA, 2020), avançou mais uma vez o estado da arte (Wiedmann, 2019).

Especialmente nos últimos meses, o PLN tem ganhado destaque na mídia devido aos avanços nos *Grandes Modelos de Linguagem* e o lançamento do *ChatGPT* (Barthlomew, 2023), um *chatbot* tão sofisticado que não só passaria com sucesso no Teste de Turing (Biever, 2023), como já vem sendo utilizado profissionalmente em diversas áreas, incluindo a prática do Direito (Schwaecz, 2023). Levantando questões sobre privacidade, direitos autorais, geração de conteúdo malicioso e impactos no mercado de trabalho, a evolução da performance no PLN é tão fascinante quanto é alarmante e nos desafia a enfrentar as implicações tecnológicas, éticas e econômicas da adoção da IA em nossa sociedade.

No Direito, as aplicações de IA já permeiam boa parte das atividades relacionadas à prática da profissão. Desde tarefas de revisão de textos, como contratos e peças processuais, até pesquisa jurídica e análise preditiva, softwares e plataformas se utilizam das diversas técnicas de aprendizado de máquina e PLN para automatizar e acelerar o trabalho de advogados, magistrados e operadores do Direito. Na seção seguinte analisamos com mais detalhes algumas dessas aplicações.

3. A UTILIZAÇÃO DA IA NO DIREITO

De acordo com o Painel de Projetos com Inteligência Artificial no Poder Judiciário, o relatório mais detalhado, publicado pelo Conselho Nacional de Justiça (CNJ), estavam em curso no país, em 2020, 41 projetos envolvendo a IA aplicada ao Direito, distribuídos entre 32 Tribunais (Brasil, 2020). Destes, a maioria pertencia a Tribunais Estaduais (19) e foi desenvolvida inteiramente pela equipe técnica de suas instituições (24), mas não apenas: há relatos de projetos executados por inteiramente ou em parte por empresas privadas³ e em parceria com Universidades. A maioria dos projetos foi realizado em

³ Entre eles, destaca-se o LEIA, realizado pela empresa privada Softplan, o chatbot do TRE-DF, desenvolvido em parceria com o Whatsapp e o Vitor, do STF, construído em parceria com a Universidade de Brasília.

Python (36) e aplica métodos de redes neurais e redes neurais convolucionais (*Convolutional Neural Networks, CNN*) (16), além de árvores de decisão e floresta aleatória (12) (Brasil, 2020).

Entre as ferramentas, destacavam-se desde já os chatbots e assistentes virtuais. Por exemplo, o Bot Eleitoral (BEL) do Tribunal Regional Eleitoral do Espírito Santo (TRE/ES), um dos projetos descritos no Painel, é um assistente virtual que utiliza processamento de linguagem natural para responder a um menu de perguntas usuais feitas ao Disque Eleições, relativas ao número de título eleitoral, local de votação, situação eleitoral, justificativa etc.⁴ De forma semelhante funciona a Celina, um assistente virtual construído pelo TRE de Rondônia (TER/RN), disponível no site oficial do Tribunal ou no Telegram para responder questões sobre local de votação e dúvidas sobre os serviços prestados no Tribunal.⁵

Além das ferramentas criadas para interagir diretamente com a usuária, respondendo a perguntas, a mais numerosa das iniciativas desenvolvidas pelos Tribunais brasileiro para lidar com grandes volumes de documentos é o classificador de documentos de um processo. É o caso do Hércules, construído pelo Tribunal de Justiça do Estado do Alagoas (TJ/AL), a ferramenta propõe uma classificação automática de petições intermediárias na 15ª vara de execuções fiscais de Maceió e a posterior distribuição por filas de acordo com a classificação obtida. Além disso, há a validação do preenchimento de Certidões de Dívida Ativa, considerando os requisitos legais. No mesmo sentido é o LEIA Petições Intermediárias, criada por uma empresa do mercado privado, mas implementada pelo TJAL e pelo Tribunal de Justiça do Estado do Amazonas (TJ/AM) e que também promove a classificação automática de petições intermediárias, a partir da análise, identificação e sugestão do tipo e da categoria da

⁴ O chatbot Bel foi premiado na categoria Inovação Tecnológica no Prêmio de Inovação Judiciário Exponencial, em 2021 (Brasil, 2021a).

⁵ O nome Celina homenageia à natalense Celina Guimarães, primeira mulher a exercer o direito de voto no Brasil.

petição.⁶ Ambos os projetos de classificação se propõem a facilitar a gestão da documentação jurídica e garantir celeridade na tramitação processual.

Também já utilizavam ferramentas dedicadas a classificar documentos o Tribunal de Justiça do Estado do Amapá (TJ/AM), a partir do Tucujuris Inteligência Artificial (TIA); o Tribunal de Justiça do Estado do Goiás (TJ/GO), com o Berna; o Tribunal de Justiça do Estado do Pernambuco (TJ/PE), que adota a Elis; o Tribunal de Justiça do Estado do Paraná (TJ/PR), a partir do Robo Larry; o Tribunal de Justiça do Estado de Santa Catarina (TJ/SC), com o Classificador de Petições de Execução Fiscal; o Tribunal de Justiça do Estado do Tocantins, com o MinerJus; o Tribunal Federal da 1ª Região (TRF1), que implementou o Análise Legal Inteligente (ALEI); o Tribunal Federal da 4ª Região (TRF4), que agrupa apelações por similaridade de sentenças; o Tribunal Federal da 5ª Região (TRF5), o qual agrupa petições iniciais para realização de perícia médica; o Tribunal Superior do Trabalho (TST), que criou o Bem-te-vi para triar documentos processuais; o Tribunal do Trabalho da 7ª Região (TRT7) e da 20ª Região (TRT20), que agrupa documentos para facilitar a distribuição de trabalho entre funcionários públicos com o GEMINI.⁷

Entre as principais aplicações utilizadas em Tribunais Superiores, a IA Victor, recentemente remodelada como VictorIA, é um sistema que apoia a atividade de análise de admissibilidade recursal por meio da sugestão de um ou mais temas de repercussão geral, posteriormente sujeita à validação pelos servidores e pelos Ministros (Salomão,

⁶ O LEIA também possui um módulo de precedentes, adotado pelo Tribunal de Justiça do Mato Grosso do Sul (TJ/MS). Este módulo auxilia Poder Judiciário na identificação de possíveis candidatos à vinculação à um dos temas de precedentes em tramitação nos Tribunais Superiores. Com base em matrizes de entendimento criadas pela empresa e validadas pelo Núcleo de Gerenciamento de Precedentes do TJ/MS (NUGEP/TJMS), a aplicação utilizou técnicas de processamento de linguagem natural e automação de consultas em base indexada baseado em regras definidas por seres humanos para realizar a "leitura e interpretação" da petição inicial do processo e verificar se está relacionada à um dos temas selecionados. (Brasil, 2019)

⁷ O Toth, criado pelo Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT) também é um classificador, criado para recomendar a classe e os assuntos do processo, durante a etapa do fluxo do PJe, identificada como petição inicial. O Tribunal é pioneiro em aplicações utilizando IA. A época do lançamento do Painel, estavam em produção 4 projetos: além do Toth, o Amon (para implementar um sistema de reconhecimento facial a partir de fotografias), o Artiu (dedicado ao encaminhamento correto de mandados para distribuição e cumprimento nos endereços dos jurisdicionados), o Hórus (para inserir automaticamente no Projeto Judicial Eletrônico (PJe) os processos digitalizados e Natureza Conciliação (para atendimento aos Centros Judiciários de Solução de Conflitos e Cidadania – CEJUSCs a partir da classificação de procedimentos que gerassem uma ou mais sessões de conciliação) (Brasil, 2021b).

2023, p. 27-31). A análise de admissibilidade é o procedimento de verificar se todos os requisitos para propor um recurso no STF foram atendidos; quando um recurso é muito similar a outro anteriormente proposto, aquele deve aguardar o julgamento deste. Com os avanços descritos em NLP, a VictorIA passou a agrupar processos por similaridade de temas, de forma que consegue identificar novas controvérsias (Brasil, 2023).

De acordo com o Segundo e o Terceiro Relatórios “Tecnologias Aplicadas à Gestão de Conflitos no Poder Judiciário com ênfase no uso da inteligência artificial”, coordenado pelo Ministro Luis Felipe Salomão, do Superior Tribunal de Justiça (STJ), em 2021 o mapeamento de iniciativas de IA aplicadas ao Direito no Poder Judiciário identificou 64 ferramentas do tipo, 23 a mais que no ano anterior. Já em 2022, foram identificados 85 novos projetos de IA.⁸ A conclusão dos Relatórios é de que a maioria das ferramentas tem a sua construção motivada pelo aumento de produtividade, a busca pela inovação, a melhoria da qualidade dos serviços judiciários e a redução de custos (Salomão, 2023). Objetivos relacionados com o princípio constitucional da eficiência aplicado ao Poder Público e com a garantia do acesso à justiça e do devido processo legal, expressamente destacados no art. 5º da Constituição da República Federativa do Brasil de 1988 (CFRB/1988).

A carta constitucional assegura, como direito e garantia fundamental dos brasileiros e estrangeiros residentes no país, a vida, a liberdade, a igualdade, a segurança e a propriedade (art. 5º). Faz isso especificando direitos e garantias individuais e coletivas, entre elas determinando que ninguém “será privado da liberdade ou de seus bens sem o devido processo legal” (art. 5º, inc. LIV); garantindo o “direito de petição aos Poderes Públicos em defesa de direitos ou contra ilegalidade ou abuso de poder” e protegendo a coisa julgada (art. 5º, inc. XXIV) e assegurando que “a lei não excluirá da apreciação do Poder Judiciário lesão ou ameaça a direito” (art. 5º, inc. XXV).

⁸ 12 projetos identificados em 2021 foram cancelados ou suspensos [REF].

Os mencionados incisos do art. 5º viabilizam a vida, a liberdade, a igualdade, a segurança e a propriedade porque promovem previsibilidade, confiabilidade e razoável rapidez na solução de conflitos. Trata-se de requisitos basilares do Estado Democrático de Direito (Waldron, 2008), os quais estruturam as relações sociais por meio do ordenamento jurídico. O julgamento parcial, as condições desiguais de acesso à Justiça, a demora na apreciação de pedidos litigiosos, muitas vezes implicando em privação de liberdade e danos materiais, não se conformam, de nenhuma maneira, ao desenho constitucional do Estado brasileiro. Democratizar o acesso à informação é garantir igualdade, dignidade e cidadania.

O investimento patente em classificadores e agrupadores de documentos por similaridade reflete a demanda de o Sistema Jurídico brasileiro prover condições de materialidade ao art. 5º da CRFB/1988. A partir da formação de grupos de conteúdo e documentos é mais fácil selecionar o mais adequado à necessidade informacional do jurisdicionado. Além disso, facilita a descoberta de novas oportunidades de atuação, o que promove equalização das condições de trabalho em um mercado tão desigual quanto o jurídico. Na seção seguinte, investigamos o emprego de técnicas de modelagem de tópicos e modelos de linguagem visando a organização e sumarização de textos jurídicos.

4. ESTUDO DE CASO: ORGANIZAÇÃO DE DOCUMENTOS JURÍDICOS

Investir em ferramentas inteligentes para apoiar especialistas jurídicos e acelerar decisões legais é crucial para muitos países devido à complexidade do sistema judiciário em todo o mundo. Nesse estudo de caso, focamos na tarefa desafiadora de organizar grandes e complexas coleções de documentos legais, sendo este um passo importante para a construção de soluções inteligentes eficientes para diversos problemas no domínio jurídico. A organização de documentos jurídicos serve de base para a realização de diversas outras tarefas, como a busca/pesquisa de casos jurídicos similares (Shao et al., 2021; Sansone; Sperlí, 2022; Novaes et al., 2023), recomendação de artigos e livros

jurídicos (Yang et al., 2021; Huang et al., 2021) e, classificação de texto jurídico (Chen et al., 2022; Chalkidis et al., 2019; Sulis et al., 2022).

Textos jurídicos são na maioria das vezes longos, escritos num jargão muito próprio e quase sempre de difícil compreensão para o público em geral. Além disso, um único texto pode tratar de uma variedade de temas diferentes. A tarefa de organizar textos jurídicos envolve o desenvolvimento de técnicas computacionais capazes de extrair conhecimento desses documentos fracamente estruturados e complexos, revelando temas, que muitas vezes são implícitos e desconhecidos, principalmente quando consideramos a natureza dinâmica do universo jurídico.

Nos últimos anos, com os avanços na área de PLN, mais especificamente em modelos de tópicos neurais (Dieng Ruiz; Blei, 2020; Bianchi; Terragni; Hovy, 2021; Srivastava; Sutton, 2017) e representações contextuais para texto (Devlin et al., 2019), as técnicas de descoberta e modelagem de tópicos foram aprimoradas e hoje podem ser aplicadas de forma eficaz a toda e qualquer coleção de documentos, incluindo documentos jurídicos com toda peculiaridade que esses apresentam em sua linguagem.

Nessa seção, mostramos a viabilidade do uso de técnicas modernas de descoberta de tópicos em conjunto com representações contextuais de texto para permitir a organização de uma coleção de documentos jurídicos composta por aproximadamente 14 mil acórdãos do STJ (Supremo Tribunal de Justiça) relacionados ao assunto "Contribuições Previdenciárias". Acórdãos são decisões proferidas por um colegiado em tribunais de segunda instância, no caso desse estudo, o STJ. Na subseção seguinte, detalhamos esse estudo de caso, seguido de uma avaliação quantitativa e qualitativa da solução desenvolvida. Uma versão detalhada deste estudo foi apresentada em (Vianna et al., 2023).

5. DESCOBERTA DE TÓPICOS PARA O DOMÍNIO JURÍDICO

No contexto das técnicas de modelagem e descoberta de tópicos, um tópico consiste em um agrupamento de palavras-chave que juntas possuem algum significado expressivo, permitindo a interpretação e organização de coleções de dados. Um algoritmo de descoberta de tópicos precisa analisar todos os documentos da coleção em busca de padrões de palavras e expressões que sejam semanticamente semelhantes e que melhor representem a temática de subconjuntos de documentos. Todo o processo é não supervisionado; ou seja, não existe nenhum conhecimento prévio dos tópicos que compõem a coleção de dados.

O estudo começa com a criação da coleção de dados a ser organizada. Para isso, expressões regulares contendo as palavras-chave 'contribuições previdenciárias' e 'contribuição previdenciária' são utilizadas para extrair, a partir de uma base de dados maior (Vianna et al., 2023), acórdãos do STJ que mencionam alguma variação desses termos. Em seguida, diversas técnicas de pré-processamento são aplicadas ao texto original com o intuito de limpar e normalizar a coleção. Entre as técnicas empregadas está a conversão de todo o texto para minúscula e a remoção de URLs, *stopwords* e pontuação. Essas duas etapas, extração e pré-processamento, resultaram em uma coleção contendo aproximadamente 14 mil acórdãos do STJ sobre o assunto "Contribuições Previdenciárias". É importante ressaltar que essa coleção contém apenas acórdão publicados até 2022.

A partir da criação e pré-processamento da coleção de dados, um método de descoberta de tópicos, denominada BERTopic (Grootendorst, 2022) é aplicada a todos os documentos da coleção resultando na descoberta de aproximadamente 300 tópicos dentro do assunto principal 'Contribuições Previdenciárias'.

BERTopic é um método de modelagem de tópicos baseada na descoberta de *clusters* (agrupamentos) densos de palavras a partir de representações vetoriais de

documentos, denominados *embeddings*. Nesta técnica, documentos são representados no espaço vetorial de maneira que a similaridade semântica entre eles pode ser calculada. BERTopic parte do princípio de que documentos que abordam os mesmos tópicos são semanticamente similares. Uma vez que os documentos são projetados no espaço vetorial, uma técnica tradicional de aprendizado de máquina, denominada de *redução de dimensionalidade*, é aplicada aos vetores que representam os documentos, com intuito de reduzir a complexidade de tais vetores e permitir a aplicação de outras técnicas de aprendizado de máquina não supervisionado, como é o caso dos algoritmos de agrupamento. Depois de reduzidos os vetores, um algoritmo de agrupamento, chamado Hdbscan (Rahman et al., 2016), é aplicado as representações vetoriais dos documentos. O objetivo é agrupar documentos relacionados em um mesmo *cluster* ou grupo. As palavras-chave que representam um tópico são extraídas a partir desses documentos usando uma técnica chamada de *c-TF-IDF* (Grootendorst, 2022), responsável por calcular a relevância de um termo a partir dos documentos de um tópico. É relevante mencionar, que o algoritmo HDBSCAN é uma técnica de agrupamento conhecida como *soft-clustering*, ou seja, um documento pode 'pertencer' a diferentes clusters (tópicos) com probabilidades diferentes. Neste estudo, vamos focar no cluster (tópico) *dominante* de cada documento. Cluster dominante é o cluster com maior probabilidade com relação a um determinado documento.

Como mencionado no parágrafo anterior, antes de serem agrupados, os documentos da coleção precisam ser convertidos em uma representação vetorial. Dada que a coleção adotada é formada por documentos longos, contendo em média 3041 palavras, decidimos por adotar o modelo Doc2Vec (Le; Mikolov 2014). Diferentemente dos modelos atuais baseados em *Transformers* (Vaswani et al., 201) tais como Bert (Devlin et al., 2019b) e GPT (Brown et al., 2020), que sofrem com limitações com relação ao tamanho do dado de entrada, Doc2Vec é capaz de gerar uma representação vetorial para todo o documento sem a necessidade de particionar esses documentos em blocos menores de texto.

6. EXPERIMENTOS

Nesta seção apresentamos os resultados obtidos, descrevemos a metodologia de avaliação aplicada e uma breve discussão sobre os impactos de tais resultados para o universo jurídico brasileiro.

Metodologia de avaliação. Os resultados obtidos durante este estudo foram avaliados por um time de especialistas do direito. Os especialistas julgaram a qualidade de um tópico descoberto pelo método de acordo com: (1) a qualidade das palavras-chave geradas para representarem o tópico e; (2) a qualidade do relacionamento entre os documentos agrupados em um mesmo tópico e a relação entre os documentos e o conjunto de palavras-chave gerado. Com relação a qualidade das palavras-chave, os especialistas atribuíram a cada tópico analisado uma nota variando de 1 a 5, onde 5 representa um tópico de excelente qualidade. Ou seja, o conjunto de palavras-chave é composto por termos que além de trazerem uma ideia clara do tópico em questão, permitem também uma fácil distinção entre os demais tópicos gerados. Por exemplo, os termos '*rural, comercialização, empregador rural, comercialização produção, rural pessoa, inconstitucionalidade, pessoa física, reprivatização, declaração inconstitucionalidade, funrural*' obtiveram a nota máxima (nota 5), uma vez que permitiram que os especialistas prontamente identificassem o assunto principal desse tópico '*Contribuição exigida do empregador rural pessoa física*'. Já tópicos com notas mais baixas, são aqueles cujas palavras-chave não trazem clareza com relação aos assuntos presentes nos documentos desse tópico. Nesse caso, as palavras-chave são muito gerais e sem significado claro. Por exemplo, os termos '*cofins, ipi, receitas, pis, financeiras, monetária, creditamento, correção monetária, capital próprio, correção*' foram considerados como sendo de baixa qualidade (nota 1), uma vez que os especialistas do direito não foram capazes de identificar o assunto principal discutido nesse tópico. Para os especialistas, os termos foram considerados bem gerais, englobando uma variedade de assuntos.

Para avaliar a qualidade dos documentos agrupados em um tópico, os especialistas consideraram o mínimo de 10 documentos e o máximo de 20 documentos por tópico.⁹ Para cada um dos documentos avaliados, os especialistas avaliaram se os documentos são relacionados entre si e se abordam a temática sugerida pelas palavras-chave que representam o tópico. Sendo assim, documentos foram marcados como relevantes ou não para os tópicos a que pertencem. Para garantir que os tópicos avaliados sejam representativos de toda coleção, uma amostra dos tópicos foi selecionada considerando um intervalo de confiança de 95% e uma margem de erro de 10%. No total, os especialistas avaliaram 163 tópicos dos 303 tópicos encontrados pelo algoritmo de modelagem de tópicos.

7. RESULTADOS

O primeiro passo para validar os resultados obtidos, foi a análise dos especialistas das palavras-chave encontradas pelo método de modelagem de tópicos para a amostra contendo 163 tópicos. Seguindo a metodologia proposta, os especialistas deram notas entre 1 e 5 para cada um dos tópicos avaliados, onde 5 representa tópicos de excelente qualidade. O Quadro 1 apresenta o número total e percentagem de tópicos para cada uma das cinco notas possíveis. Podemos notar que 58% dos tópicos foram considerados como sendo de alta qualidade, ou seja, as palavras-chave representam de forma clara a temática do tópico e ainda ajudam a diferenciar o tópico analisado dos demais tópicos encontrados pelo modelo. Apenas 3% dos tópicos receberam notas 1 e 2, indicando que são raros os casos onde os tópicos são irrelevantes, pelo menos no que se refere a qualidade das palavras-chave.

⁹

As pessoas especialistas que ajudaram na definição dos tópicos de pesquisa e na avaliação dos documentos agrupados integravam, à época, a equipe de Legal Consultants do Jusbrasil (Isabela Franco e Letícia Leite) e de Legal Engineering, mais especificamente do time de pesquisa jurídica (Luiz Azevedo; Maria Flávia Freitas Ribeiro; Catarina Wodzik; Ana Laura Monteiro e Victória Hollanda).

Quadro 1 - Avaliação das palavras-chave da amostra de 163 tópicos dos 303 tópicos encontrados pelo algoritmo BERTopic para a coleção de dados sobre 'Contribuições Previdenciárias'

Nota 1	Nota 2	Nota 3	Nota 4	Nota 5
2 (1%)	4 (2%)	16 (10%)	46 (28%)	95 (58%)

Fonte: autoria própria.

Ainda seguindo a metodologia de avaliação proposta, para a amostra de 163 tópicos foram avaliados entre 10 e 20 documentos de cada tópico, resultando em um total de 1882 documentos avaliados. Dentre eles, 1797 (95%) documentos foram considerados relevantes pelos especialistas. Os resultados indicam que além da clareza expressa pelas palavras-chave na descrição dos tópicos, a qualidade dos agrupamentos de documentos se mostra relevante, o que pode ter aplicação direta em diversos problemas da área, como a recuperação de documentos jurídicos semelhantes, um trabalho que quando feito de forma manual pode ser lento e custoso.

O Quadro 2 ilustra alguns dos tópicos avaliados pelos especialistas. Na primeira coluna, apresentamos as palavras-chave que representam o tópico e, na segunda coluna, apresentamos comentários deixados pelos especialistas sobre cada um desses tópicos.

Concluimos que os agrupamentos foram organizados conforme os documentos tratam dos mesmos temas de cauda longa, ou seja, assuntos mais específicos dentre os existentes em um grupo temático mais amplo. Por exemplo, todos os temas do Quadro 2 pertencem ao ramo do Direito Tributário, mais especificamente, ao tema mais específico "contribuições previdenciárias", que são os tributos pagos ao INSS. Dentro deste tema de mais alto nível, conseguimos separar agrupamentos identificados por tipos de contribuições previdenciárias, como demonstrado na linha 1. Também identificamos agrupamentos ainda mais específicos, como no caso da linha 4, em que os documentos tratavam nomeadamente da Lei estadual de Minas Gerais que instituiu contribuições

previdenciárias para custeio de serviços de assistência médica, ou das linhas 6 e 7, que discutem a incidência dessas contribuições sobre verbas remuneratórias de natureza salarial polêmica, quais sejam, o auxílio alimentação e o décimo terceiro. Neste caso específico [das linhas 6 e 7], é muito importante ressaltar a especificidade dos agrupamentos, já que os documentos discutem teses advocatícias muito similares, mas foi capaz de segregá-los de acordo com a parcela remuneratória discutida.

Ou seja, a modelagem de tópicos foi capaz de, diante de um conjunto pré-selecionado de documentos, sem qualquer direcionamento, identificar assuntos específicos que eram discutidos em documentos jurídicos.

Quadro 2 – Alguns exemplos de tópicos e palavras-chave encontradas pelo algoritmo de modelagem de tópicos.

Tópico: Palavras-chave	Comentário
incra, destinada incra, destinada, extinta, funrural, agrária, domínio econômico, reforma agrária, domínio, social, seguridade, econômico, empresas, contribuição, urbana, seguridade social, reforma, empresa urbana, sesc, intervenção	Diferentes assuntos relacionados à contribuição para o Incra.
renúncia, desaposentação, aposentadoria, devolução valores, renúncia aposentadoria, devolução, jubramento, valores recebidos, benefício, recebidos, concessão novo, segurado, posterior jubramento, concessão, novo benefício, valores, previdenciário, posterior, novo posterior, novo	Devolução dos valores recebidos em razão de desaposentação
rural, prova material, prova, início prova, testemunhal, início, material, prova testemunhal, atividade, idade, atividade rural, rurícola, autora, aposentadoria, exercício atividade, aposentadoria idade, período, carência, familiar, autor	Tempo de serviço rural/contribuições não recolhidas
saúde, minas gerais, minas, estado minas, gerais, mg, serviços saúde, custeio, custeio serviços, ipsemg, serviços, agrg mg, estado, indébito, mg mín, repetição, assistência saúde, contribuição custeio, servidores estado, indevida	Lei estadual que instituiu CP para custeio de serviços de assistência médica
rural, prova material, prova, início prova, testemunhal, início, material, prova testemunhal, atividade, idade, atividade rural, rurícola, autora, aposentadoria, exercício atividade, aposentadoria idade, período, carência, familiar, autor	Tempo de serviço rural/contribuições não recolhidas

alimentação, in natura, natura, auxílio alimentação, pat, auxílio, programa alimentação, alimentação trabalhador, trabalhador pat, in, pago, alimentação pago, programa, trabalhador, fornecida, empregador, salário in, transporte, pecúnia, alimentação	Incidência de CP sobre auxílio alimentação
separado, natalina, gratificação natalina, gratificação, salário, cálculo separado, separado gratificação, tributação separado, incidente gratificação, mês dezembro, aplicação separado, terceiro salário, décimo terceiro, décimo, mediante aplicação, cálculo, dezembro, terceiro, separado tabela, tabela	CP sobre 13º salário (gratificação natalina)

Fonte: autoria própria.

8. CONCLUSÕES

Neste artigo, apresentamos um caso de uso de inteligência artificial e processamento de linguagem natural aplicado ao contexto do Direito no Brasil. Precisamente, mostramos que a utilização de modelos de linguagem combinados com a modelagem não supervisionada de tópicos pode ser aplicada de forma eficiente e precisa na organização de grandes acervos de texto jurídico. Experimentos foram realizados com uma coleção de dados composta por aproximadamente 14 mil acórdãos do STJ sobre o assunto "Contribuições Previdenciárias". Os resultados foram cuidadosamente avaliados por um grupo de especialistas (Vianna et al., 2023) que concluiu que 86% dos 163 tópicos avaliados eram de boa qualidade (notas 4 e 5), o que significa que, no geral, as palavras-chave dos tópicos representam de forma clara e satisfatória o conteúdo de cada um desses tópicos. Além disso, os especialistas também avaliaram a qualidade e coerência entre os documentos de um mesmo tópico. Os resultados obtidos foram expressivos, com 95% dos documentos avaliados considerados relevantes com relação a seus tópicos.

Com os avanços na área de processamento de linguagem natural, mais especificamente, com os grandes modelos de linguagem, é possível vislumbrar possíveis melhorias principalmente em relação à descrição dos tópicos gerados. Por exemplo,

modelos gerativos podem ser usados para enriquecer o conjunto de palavras-chave que descrevem um tópico, trazendo mais contexto sobre o conteúdo dos tópicos e automaticamente facilitando o entendimento do mesmo por parte não apenas dos especialistas do direito, mas também do cidadão comum. Uma outra possibilidade é o uso de modelos de linguagem para a geração de resumos que descrevam os tópicos em uma linguagem mais próxima da linguagem humana.

Além do caso de uso apresentado, esse estudo traz uma revisão abrangente das tecnologias e métodos de inteligência artificial e processamento de linguagem natural, com foco em aplicações no contexto do Direito do Brasil.

REFERÊNCIAS

ARORA, S. et al. **Contextual embeddings: When are they worth it?** arXiv preprint arXiv:2005.09117 (2020).

BARTHOLOMEW, J. et al. **How the media is covering ChatGPT.** Disponível em: <https://www.cjr.org/tow_center/media-coverage-chatgpt.php> Acesso em 12/08/2023

BIANCHI, Federico; TERRAGNI, Silvia; HOVY, Dirk. **Pre-training is a hot topic: Contextualized document embeddings improve topic coherence.** In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. [S.l.]: Association for Computational Linguistics, 2021. p. 759-766.

BIEVER, C. **ChatGPT broke the Turing test – the race is on for new ways to assess AI.** Disponível em: <<https://www.nature.com/articles/d41586-023-02361-7>> Acesso em 12/08/2023

BROWN, T. B. et al. **Language models are few-shot learners.** In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020. (NIPS'20). ISBN 9781713829546.

CHALKIDIS, I. et al. **Extreme multi-label legal text classification: A case study in EU legislation.** In: *Proceedings of the Natural Legal Language Processing Workshop 2019*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 78-87. Disponível em: <<https://aclanthology.org/W19-2209>>.

CHEN, H. et al. **A comparative study of automated legal text classification using random forests and deep learning.** *Information Processing Management*, v. 59, n. 2, p. 102798, 2022. ISSN 0306-4573. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0306457321002764>>.

BRASIL. Conselho Nacional de Justiça. **Painel de projetos com inteligência artificial no Poder Judiciário**, 2020. Disponível em: <<https://paineisanalytics.cnj.jus.br/single/?appid=29d710f7-8d8f-47be-8af8-a9152545b771&sheet=b8267e5a-1f1f-41a7-90ff-d7a2f4ed34ea&lang=pt-BR&opt=ctxmenu,currsel>>. Acesso em: 14/08/2023.

BRASIL. Tribunal Eleitoral Regional do Espírito Santo. **Bel, a assistente virtual do TRE-ES, vence a categoria "Inovação Tecnológica" do Prêmio de Inovação Judiciário Exponencial**. Comunicação, 06 out. 2021. 2021a. Disponível em: <<https://www.tre-es.jus.br/comunicacao/noticias/2021/Outubro/bel-a-assistente-virtual-do-tre-es-vence-a-categoria-inovacao-tecnologica-do-premio-de-inovacao-judiciario-exponencial>>. Acesso em: 14 out. 2023.

BRASIL, Tribunal de Justiça do Amazonas. **TJAM automatiza classificação de petições intermediárias no Portal e-SAJ**. Imprensa, 19 dez. 2019. Disponível em: <<https://www.tjam.jus.br/index.php/menu/sala-de-imprensa/2387-tjam-automatiza-classificacao-de-peticoes-intermediarias-no-portal-e-saj>>. Acesso em: 14 ago. 2023.

BRASIL. Tribunal de Justiça do Distrito Federal e Territórios. **TJDFT lidera número de projetos de Inteligência Artificial no Poder Judiciário**. Janeiro de 2021b. Disponível em: <<https://www.tjdft.jus.br/institucional/imprensa/noticias/2021/janeiro/tjdft-e-o-tribunal-com-mais-projetos-de-inteligencia-artificial>>. Acesso em: 14 ago. 2023.

BRASIL. Supremo Tribunal Federal. **Ministra Rosa Weber lança robô Vitória para agrupamento e classificação de processos**, 17 mai. 2023. Disponível em: <<https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=507426&ori=1>>. Acesso em: 14 ago. 2023.

DEVLIN, J. et al. **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)*. [S.l.: s.n.], 2019. p. 4171–4186.

DEVLIN, J. et al. **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)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <<https://aclanthology.org/N19-1423>>.

DIENG, Adj. B.; RUIZ, Francisco J. R.; BLEI, David M. **Topic modeling in embedding spaces**. *Transactions of the Association for Computational Linguistics*, MIT Press, Cambridge, MA, v. 8, p. 439–453, 2020.

ETHAYARAJH, K. BERT, ELMO, & GPT-2: How Contextual are Contextualized Word Representations? Disponível em: <<http://ai.stanford.edu/blog/contextual/>> Acesso em 12/08/2023

GROOTENDORST, Maarten. **Bertopic: Neural topic modeling with a class-based tf-idf procedure**. *CoRR*, abs/2203.05794, 2022. Citado na página 1.

HOFMANN, V. et al. **Dynamic contextualized word embeddings**. arXiv preprint arXiv:2010.12684 (2020).

HUANG, Z. et al. **Context-aware legal citation recommendation using deep learning**. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. New York, NY, USA: Association for Computing Machinery, 2021. (ICAIL '21), p. 79–88. ISBN 9781450385268. Disponível em: <<https://doi.org/10.1145/3462757.3466066>>.

JOJRI, P. et al. **Natural language processing: History, evolution, application, and future work**. In: *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020*, pp. 365-375. Springer Singapore, 2021.

LE, Quoc; MIKOLOV Tomas. **Distributed representations of sentences and documents**. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - ICML*. [S.l.: s.n.], 2014. p. II-1188-II-1196.

LIU, Q. et al. **A survey on contextual embeddings**. arXiv preprint arXiv:2003.07278, 2020

LUND, K.; BURGESS C. **Producing high-dimensional semantic spaces from lexical co-occurrence**. In: *Behavior research methods, instruments, & computers* 28.2 (1996): 203-208. Disponível em: <<https://doi.org/10.3758/BF03204766>>

MIKOLOV, T. et al. **Efficient Estimation of Word Representations in Vector Space**. *Proceedings of the First International Conference on Learning Representations*, 2013.

NOVAES, Luisa Pereira et al. **Modelagem de Tópicos para a Tarefa de Recuperação de Casos Legais**. *Brazilian Symposium on Databases* (2023).

RAHMAN, M. F. et al. **Hdbscan: Density based clustering over location based services**. *ArXiv*, abs/1602.03730, 2016.

RUSSEL, S. J. **Artificial intelligence a modern approach**. 3. ed. Pearson Education, Inc., 2010.

SALOMÃO, Luis Felipe (coord.). **Tecnologia aplicada à gestão dos conflitos no âmbito do Poder Judiciário Brasileiro**. Rio de Janeiro: Editora FGV Conhecimento, 2023.

SANSONE, Carlo; SPERLÍ, Giancarlo. **Legal information retrieval systems: State-of-the-art and open issues**. *Inf. Syst.*, Elsevier Science Ltd., GBR, v. 106, n. C, may 2022. ISSN 0306-4379. Disponível em: <<https://doi.org/10.1016/j.is.2021.101967>>.

SCHWARCZ D. et al. **AI Tools for Lawyers: A Practical Guide**. In: *Minnesota Law Review Headnotes*. Disponível em: <<http://dx.doi.org/10.2139/ssrn.4404017>>

SHAO, Y. et al. **Bert-pli: Modeling paragraph-level interactions for legal case retrieval**. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 2021. (IJCAI'20). ISBN 9780999241165.

SRIVASTAVA, Akash; SUTTON, Charles. **Autoencoding Variational Inference for Topic Models**. *arXiv e-prints*, p. arXiv:1703.01488, mar. 2017.

SULIS, E. et al. **Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts.** *Information Systems*, v. 106, p. 101821, 2022. ISSN 0306-4379. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0306437921000648>>.

VASWANI, A. et al. **Attention is all you need.** In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

VIANNA, D. et al. **A topic discovery approach for unsupervised organization of legal document collections.** *Artificial Intelligence and Law* (2023): 1-30.

WALDRON, Jeremy. The Concept and the Rule of Law. In: No. 08–35; *Public Law & Legal Theory Research Paper Series*, Issue November, 2008.

WIEDMANN, G. et al. **Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings.** arXiv preprint arXiv:1909.10430, 2019

YANG, J. et al. **Legalgnn: Legal information enhanced graph neural network for recommendation.** *ACM Trans. Inf. Syst.*, Association for Computing Machinery, New York, NY, USA, v. 40, n. 2, sep 2021. ISSN 1046-8188. Disponível em: <<https://doi.org/10.1145/3469887>>.