A Textual Representation Based on Bag-of-Concepts and Thesaurus for Legal Information Retrieval

Wagner M. Costa, Glauco V. Pedrosa

Programa de Pós-Graduação em Computação Aplicada (PPCA)
Universidade de Brasilia (UnB), Brazil
wagnermc@gmail.com, glauco.pedrosa@unb.br

Abstract. The retrieval of similar textual documents is a challenging task for the legal area due to its peculiar language with unique characteristics. This paper presents a new approach, called BoC-Th, proposed to represent legal documents based on the Bag-of-Concept (BoC) approach, which generates concept through clustering word vectors generated from a basic neural network model, and compute the frequencies of these concept clusters to represent document vectors. The novel contribution of the BoC-Th is to generate weighted histograms of concepts defined from the distance of the word to its respective similar term within a thesaurus. The idea is to emphasize those words that have more significance for the context, thus generating more discriminative vectors. Experimental evaluations were performed by comparing the proposed approach with the traditional BoW and BoC approaches, both popular techniques for document representation. The proposed method obtained the best result among the evaluated techniques for retrieving judgments and jurisprudence documents. The BoC-Th increased the mAP (mean Average Precision) in 51% compared to the traditional BoC approach, while being up to 3.4 times faster than the traditional BoW representation.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: textual representation, bag of concepts, text mining, word embeddings

1. INTRODUÇÃO

A quantidade de informação jurídica produzida diariamente em meio digital está crescendo enormemente. Esta é uma área que é caracterizada por possuir estruturas implícitas e uma linguagem peculiar com características únicas, e que tem despertado muito interesse tanto por suas peculiaridades quanto pelos novos recursos, atualmente, disponíveis para definir novas metodologias de gestão de documentos legais, dentre elas a área de Inteligência Artificial [Martins and Silva 2021]. De fato, nos últimos anos, os modelos de aprendizado de máquina e aprendizado profundo têm atraído ampla atenção para o processamento de documentos legais [Solihin et al. 2021]. Por exemplo, os trabalhos de [Noguti et al. 2020], [de Campos et al. 2020], [Silva and Maia 2020] e [Dal Pont et al. 2020] mostram a aplicação de modelos de aprendizagem profunda, clássicos e estatísticos para classificação de documentos no contexto jurídico do Brasil. O artigo [de Araujo et al. 2020] também faz um estudo no contexto jurídico brasileiro, porém utiliza como proposta principal o ULMFiT.

Além da classificação de textos, uma outra tarefa importante dentro da área jurídica é a recuperação de informações similares a partir de um grande e diverso conjunto de dados. Essa é uma tarefa que permite, principalmente, fornecer acesso à lei para leigos e profissionais do direito, ampliando a transparência e legitimidade dos processos [Castano et al. 2022]. Por exemplo, através da recuperação de informações jurídicas é possível monitorar mudanças nas leis e regulamentos tributários ou usar o Raciocínio Jurídico a fim de verificar a conformidade e/ou validar a qualidade de documentos

legislativos.

Na esfera pública brasileira, algumas ferramentas de pesquisa foram criadas com o objetivo de apoiar a atividade dos operadores do Direito (magistrados, advogados, assessores jurídicos, dentre outros) quando da necessidade de localizar a jurisprudência sobre temas específicos. No Tribunal de Contas da União (TCU), por exemplo, a Pesquisa Integrada[1] disponibiliza acesso à toda a base de dados indexada do Órgão. Ferramentas com esse propósito, que estão presentes nos sítios de órgãos jurídicos – tais como Tribunais Superiores – utilizam operadores que localizam termos e expressões no conteúdo indexado. Esse tipo de busca pressupõe que o usuário que irá utilizar o sistema de busca tenha um certo nível de conhecimento sobre o texto a ser localizado. Um ponto de melhoria sobre esse tipo de busca é que o sistema consiga recuperar conteúdos com semânticas equivalentes ou aproximadas.

O passo fundamental e crucial em um sistema de recuperação de informação textual consiste em extrair uma representação vetorial compacta e discriminativa dos documentos. Para isso, a modelagem Bag-of-Words (BoW), por exemplo, é uma das técnicas mais populares na área para essa tarefa. Essa técnica representa um documento por suas frequências de palavras e, por sua simplicidade, permite uma interpretabilidade intuitiva. No entanto, esse método sofre com a maldição da dimensionalidade e desconsidera o impacto de palavras semanticamente similares. Para superar tal limitação, alguns modelos semanticamente mais enriquecidos, tais como o modelo Word2Vec [Le and Mikolov 2014] - que é baseado em redes neurais profundas - utilizam informações contextuais de cada palavra, capturando informações ao seu redor. No entanto, os vetores gerados a partir dessas técnicas são de difícil interpretação, visto que seus valores indicam o peso da rede neural usada no treinamento.

O método Bag-of-Concepts (BoC) [Kim et al. 2017] surgiu como uma alternativa de representação vetorial de documentos que supera os pontos fracos das abordagens Bag-of-Words e Word2Vec. A abordagem BoC identifica "conceitos" através do agrupamento de vetores-de-palavras gerados a partir do Word2Vec, e utiliza as frequências desses agrupamentos de conceitos para representar os documentos. Por meio de conceitos, o método Bag-of-Concepts incorpora o impacto de palavras semanticamente semelhantes, enriquecendo a representação semântica de um documento textual sem aumentar drasticamente a dimensionalidade do vetor-de-característica necessário para representá-lo.

Este trabalho apresenta uma nova abordagem para representação de documentos textuais, denominada de BoC-Th (*Bag-of-Concepts based on Thesaurus*), que é uma extensão da abordagem BoC tradicional. O diferencial inovador da técnica BoC-Th é contemplar, no processo de sumarização dos conceitos, o vocabulário específico envolvido nos textos jurídicos. A abordagem BoC-Th utiliza-se de um *thesaurus*, que é uma lista de palavras com significados semelhantes, dentro de um domínio específico de conhecimento. Na abordagem tradicional BoC, cada palavra contribui igualmente na representação final do documento, porém na abordagem BoC-Th cada palavra será ponderada pela distância ao termo mais próximo do tesauro: quanto mais distante, menos peso essa palavra terá na representação do documento. A ideia é que ao se trabalhar com um vocabulário de palavras específico ao contexto é possível enfatizar as palavras e termos que são diretamente relacionados ao linguajar peculiar dos documentos jurídicos e, dessa forma, representar os documentos textuais de uma maneira mais discriminativa.

A partir de um estudo de caso com dados reais, o método proposto foi avaliado e comparado com as abordagens Bag-of-Words e a abordagem Bag-of-Concept (BoC) tradicional. Para os testes, foram utilizadas uma base de dados de jurisprudências do TCU e um conjunto-verdade (*ground truth*), definido com a ajuda de especialistas do Órgão, e realizados experimentos de recuperação de jurisprudências similares à um dado acórdão informado pelo usuário.

O texto deste artigo está organizado da seguinte forma: na Seção 2 são apresentados os trabalhos correlatos; na Seção 3 é apresentado o método proposto; a Seção 4 apresenta os resultados experimentais obtidos em um estudo de caso envolvendo a recuperação de acórdãos e jurisprudências do TCU e

[1] `https://pesquisa.apps.tcu.gov.br/\#/pesquisa/jurisprudencia`

a Seção 5 finaliza com as conclusões do estudo desenvolvido e trabalhos futuros.

2. TRABALHOS CORRELATOS

A área de Recuperação da Informação lida com o desenvolvimento de algoritmos e modelos para recuperar informações a partir de um repositório de documentos [Yan 2009]. Do ponto de vista computacional, para a implementação de um sistema de recuperação de informação, os documentos devem ser representados de forma que o computador possa interpretar e diferenciar cada documento dentro da coleção. Esta representação visa a representar numericamente os documentos de texto não estruturados para torná-los matematicamente computáveis. Para um dado conjunto de documentos de texto $D = \{d_1, d_2, d_3, ..., d_n\}$, onde cada d_i representa um documento, o problema da representação textual é representar cada d_i de D como um ponto s_i em um espaço numérico S, onde a distância entre cada par de pontos no espaço S é bem definida.

Assim, o uso de algoritmos de aprendizagem de máquina para inferências quantitativas e preditivas e execução de tarefas como classificação ou regressão, em dados textuais, depende da transformação desses dados por meio de alguma técnica de codificação, por exemplo, *word embeddings*, que por sua vez, podem ser agrupadas em duas categorias: *embedding* baseado em frequência e *embedding* baseado em predições [Analytics Vidhya 2017]. No primeiro os vetores são formados pela contagem de palavras, no segundo, a construção dos vetores é apoiada por mecanismos de redes neurais, que resultam em representações numéricas não-determinísticas.

2.1 *Bag-of-Words*

Uma abordagem comumente adotada e eficaz para a representação de documentos é o modelo *Bag-of-Words*. O modelo BoW atribui um vetor a um documento como $d = \{x_1, x_2, x_3, ..., x_l\}$, em que x_i denota o número normalizado de ocorrências do i-ésimo termo (palavra) no documento, e l é o tamanho da coleção de termos (palavras) dos documentos presentes na base de dados.

A abordagem BoW é um método simples, mas eficaz, para mapear um documento em um vetor de comprimento fixo. No entanto, a função de mapeamento no modelo BoW é *hard* ou binária, visto que ela representa apenas a presença ou ausência de um termo base no documento. A função de mapeamento rígido tem várias limitações. Primeiro, o vetor gerado para cada documento é extremamente esparso, pois um documento contém apenas uma porção muito pequena de todos os termos básicos de uma base. Em segundo lugar, as representações BoW podem não capturar efetivamente a semântica dos documentos, uma vez que documentos semanticamente semelhantes mas com diferentes conjunto de palavras serão mapeados para espaços vetoriais muito diferentes.

2.2 *Word Embeddings*

A ideia central por trás das técnicas de *Word Embeddings* é atribuir uma representação vetorial à cada palavra, de modo que palavras que são semanticamente similares fiquem próximas umas das outras no espaço vetorial. O mérito da abordagem *Word Embeddings* é que a semelhança semântica entre duas palavras pode ser convenientemente avaliada com base na medida de semelhança de cosseno entre as representações vetoriais correspondentes das duas palavras.

Uma das técnicas de *Word Embeddings* mais populares é a técnica Word2Vec [Mikolov et al. 2013], que é baseada em uma rede neural de duas camadas. A estrutura Word2Vec contém dois modelos separados, incluindo Bag-of-Words Contínuo (CBoW) e Skip-gram com dois objetivos de treinamento reversos. CBoW tenta predizer uma palavra considerando as palavras ao seu redor, enquanto Skip-gram tenta prever uma janela de palavras dada uma única palavra. Devido à sua arquitetura e ao treinamento não-supervisionado, o Word2Vec pode ser construído eficientemente em um corpus não anotado de grande escala. O Word2Vec é capaz de codificar relações linguísticas significativas entre

palavras em *embeddings* de palavras aprendidas. Normalmente, a medida de semelhança de cosseno entre incorporações de palavras é usada para medir a semelhança semântica entre duas palavras.

Embedding em nível de documento é utilizado em [Renjit and Idicula 2019] para a avaliação de similaridade de textos jurídicos. Usando o algoritmo *paragraph vector* [Le and Mikolov 2014], os autores consolidam os vetores de características do texto e estimam os demais parágrafos com o método *stochastic gradient descent*. Em seguida, obtém-se a orientação angular dos documentos, a partir do cálculo da similaridade de cosseno de seus vetores. Os documentos que possuírem orientações similares são considerados também semanticamente similares.

2.3 *Bag-of-Concepts* (BoC)

Na abordagem *Bag-of-Concepts*, as palavras de um texto são associados às suas representações semânticas, obtidas a partir de agrupamentos originados por técnicas de clusterização. Esse método busca tratar algumas limitações apresentadas pela abordagem BoW, como sua ausência de representatividade semântica, além de alta dimensionalidade e esparsidade [Kim et al. 2017]. A definição de "conceito" como unidade de significado foi apresentado por [Mourino Garcia et al. 2015], que estudou a utilização de bases de conhecimento - no caso, a *Wikipedia*. Como resultado, obteve-se um ganho de até 157% na modelagem de classificadores, em comparação com o BoW.

3. MÉTODO PROPOSTO

O objetivo do método proposto, denominado BoC-Th, é representar documentos textuais em histogramas ponderados de conceitos, que serão utilizados para a recuperação de informações similares, dado um documento fornecido pelo usuário. O método BoC-Th é uma extensão da abordagem BoC tradicional, que contempla a semântica envolvida na análise de documentos jurídicos usando um tesauro. A ideia é que, ao se utilizar um tesauro (que possui os termos com as palavras e seus sinônimos do domínio de aplicação), aumenta-se o poder discriminativo da representação do documento.

A Figura 1 mostra os passos da abordagem proposta, que está dividida em duas fases: codificação e sumarização. A fase de codificação atribui um conceito para cada palavra contida em um documento, e a fase de sumarização contabiliza de maneira ponderada a frequência de cada conceito no documento. A seguir, essas duas fases serão detalhadas: inicialmente, será descrita a geração do dicionário de conceitos considerando o tesauro e, em seguida, a contabilização ponderada dos conceitos em um único vetor, que será a representação final do documento.

3.1 Geração do Dicionário de Conceitos

Seja $W = \{w_1, w_2, ..., w_v\}$ o vocabulário que abrange todas as palavras existentes em um corpus, e v o tamanho desse vocabulário. Cada palavra $w_i \in W$ é representada por um vetor r-dimensional obtido por uma técnica de *word embeddings*, como word2vec, gerando uma matriz $S \in \mathbb{R}^{v \times r}$. Considere $T = \{t_1, t_2, ..., t_m\}$ um tesauro com m palavras, em que cada t_i corresponde à um termo de um domínio específico de conhecimento que, no caso deste trabalho, envolve as palavras e seus sinônimos utilizados nos textos de acórdãos e jurisprudências. O método proposto faz uso de um dicionário de conceitos $C = \{c_1, c_2, ..., c_k\}$, onde cada conceito $c_i \in \mathbb{R}^r$ é o centroide de um grupo $C_i \subset \{S \cap T\}$, tal que:

$$c_i = \frac{1}{|C_i|} \sum_{x_j \in \{S \cap T\}} x_j, \quad \forall i \in \{1, 2, ..., k\} \tag{1}$$

em que, $|C_i|$ é a quantidade de palavras associadas ao grupo C_i e k é a quantidade de conceitos.

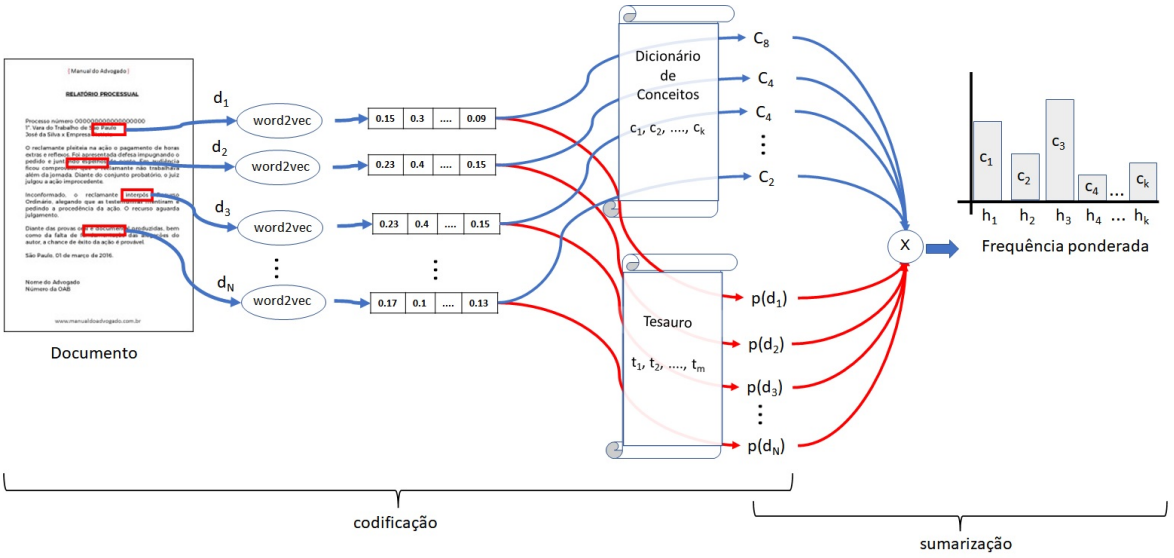

Fig. 1. Etapas do método proposto para representar um documento em um histograma ponderado de conceitos com base em um tesauro. O fluxo representado pelas linhas em azul indicam a geração do dicionário de conceitos baseado em termos presentes no tesauro. As linhas em vermelho assinalam o cálculo da função de ponderação para cada palavra do documento.

Como técnica de agrupamento, utilizou-se o algoritmo não-supervisionado *Spherical k-means*, indicado para o processamento de vetores esparsos e com grande dimensionalidade [Dhillon and Modha 2001], que vem a ser o caso dos dados analisados neste trabalho.

A abordagem proposta para a geração do dicionário de conceitos se difere da abordagem tradicional ao considerar apenas aquelas palavras do corpus que estão presentes no tesauro. A motivação para tal abordagem é que, ao se restringir a geração de conceitos às palavras estritamente relacionadas ao contexto da aplicação, evita-se a geração de conceitos não pertencentes ao domínio.

3.2 Vetorização ponderada de conceitos

Considere um documento $D = \{d_1, d_2, ..., d_N\}$, onde N é a quantidade de palavras existentes nesse documento e cada palavra $d_i \in D$ é representada em um por vetor r-dimensional obtido pela mesma técnica de *word embedding* utilizada para a geração do dicionário de conceitos. Cada palavra $d_i \in D$ é associada à cada um dos k conceitos existentes no dicionário de conceitos C pela função $\phi : \mathbb{R}^d \rightarrow \mathbb{N}^k$, definida como:

$$d_i \rightarrow \phi(d_i) = \{\alpha_1^i, \alpha_2^i, ..., \alpha_k^i\} \tag{2}$$

em que α_j^i pode ser vista como uma função de ativação, associando a i-ésima palavra do documento ao j-ésimo conceito do dicionário de conceitos. Essa ativação pode ser *hard* ou *soft*. A codificação clássica é baseada na codificação *hard* em que cada palavra d_i é atribuída à apenas um conceito, ou seja:

$$\alpha_j^i = \begin{cases} 1, & \text{se } j = \arg\min_{j \in \{1,...,k\}} ||d_i - c_j||_2^2 \\ 0, & \text{caso contrário.} \end{cases} \tag{3}$$

O diferencial inovador deste trabalho consiste em uma contabilização ponderada dos conceitos ao se vincular a distância de cada palavra $d_i \in D$ à sua palavra mais próxima $t_r \in T$, em que T é o tesauro. Em outras palavras, o vetor numérico final que irá representar o documento D será dado por um histograma ponderado de conceitos $H = [h_1, h_2, ...h_k]$, onde k é a quantidade de conceitos

existentes no dicionário de conceitos e cada $h_j \in H$ é definido como:

$$h_j = \sum_{i=1}^{N} p(d_i)\alpha_j^i, \qquad \forall j \in \{1, 2, ..., k\} \tag{4}$$

em que,

$$p(d_i) = 1 - \arg\min_{t_r \in T} ||d_i - t_r||_2^2 \tag{5}$$

A função de ponderação $p : \mathbb{R}^d \to \mathbb{R}$ retorna a distância da palavra d_i à sua palavra correspondente mais próxima no tesauro T. Se $d_i \in T$, então $p(d_i) = 1$ e, portanto, o conceito associado à palavra d_i será contabilizado integralmente no histograma. Caso contrário, quanto mais distante for a palavra d_i das palavras do tesauro, menor será a contribuição do conceito da palavra d_i no histograma final do documento.

4. RESULTADOS EXPERIMENTAIS

O método proposto BoC-Th foi comparado com duas outras abordagens tradicionais da literatura: BoW e BoC. O desempenho de cada uma das técnicas foi avaliado sobre um estudo de caso, envolvendo a busca e análise de similaridade de acórdãos e jurisprudências do Tribunal de Contas da União. A seguir, serão detalhados as configurações e os resultados obtidos.

4.1 Base de Dados e *Ground-Truth*

A base de dados para a realização dos experimentos comparativos é composta de um conjunto de 15.000 (quinze mil) enunciados de jurisprudências do Tribunal de Contas da União, disponível em seu Portal[2]. No processo de elaboração da jurisprudência, especialistas do TCU agrupam os acórdãos conforme as áreas de atuação do Controle Externo e consolidam o entendimento do Tribunal a respeito de determinado assunto. Esses textos recebem o nome de *enunciados*. Por sua vez, os acórdãos, que representam as decisões da Corte, têm seus textos sintetizados em *sumários*.

Com o objetivo de se estabelecer um *benchmark*, foi construído uma base de referência (*ground truth*) para comparação dos resultados obtidos. Esta base é composta pela seleção de 10 (dez) sumários de acórdãos prolatados pelo TCU. Para cada um desses sumários, especialistas do TCU manualmente indicaram os enunciados com maior similaridade semântica. Esse conjunto (sumários de acórdãos e seus respectivos enunciados) constitui o *ground truth* deste trabalho.

Complementarmente, empregou-se o tesauro do TCU, denominado Vocabulário de Controle Externo (VCE)[3], criado com o objetivo de uniformizar a terminologia usada nas atividades institucionais do TCU, além de apoiar o tratamento da informação no Órgão.

4.2 Melhores parâmetros

Para a implementação das abordagens BoC e BoC-Th, dois parâmetros precisam ser previamente definidos. São eles:

(1) o tamanho do dicionário de conceitos
(2) o tamanho do vetor-de-palavras gerado pela técnica word2vec.

[2] `https://www.tcu.gov.br/`
[3] `https://portal.tcu.gov.br/vocabulario-de-controle-externo//`

Empiricamente, realizou-se exaustivos testes com diferentes valores para esses dois parâmetros para as duas abordagens, investigando qual a combinação com melhor precisão. Os melhores parâmetros para as duas abordagens foi obtido com o vetor-de-palavras (word2vec) com dimensão igual a 100 (cem) e um dicionário com 50 (cinquenta) conceitos. Portanto, esses dois parâmetros foram utilizados, tanto pela abordagem BoC quanto pela BoC-Th, para a realização dos resultados comparativos de desempenho, que serão detalhados em seguida.

4.3 Resultados comparativos

A métrica utilizada para avaliação do desempenho entre as técnicas foi o *mean Average Precision* (mAP), que fornece a média das precisões médias para cada objeto de consulta (*query*). A Tabela I exibe os melhores valores obtidos para cada abordagem, bem como o tamanho do vetor utilizado para representar cada documento textual.

Nota-se que, a abordagem proposta BoC-Th foi 140% superior à abordagem BoW e 51% superior em relação à abordagem BoC. Além disso, o tamanho do vetor utilizado pela abordagem BoC-Th é 97.3% menor que a abordagem BoW. Isso se reflete em economia de espaço de armazenamento e, principalmente, no baixo custo computacional para realizar o cálculo da similaridade entre os vetores no processo de recuperação de documentos similares.

Abordagem	***mAP***	***Dimensionalidade do Vetor***	***Tempo de Processamento***
BoW	0,27	1850	0,435s
BoC	0,43	50	0,0092s
BoC-Th	0,65	50	0,0135s

Table I: Melhores valores mAP obtidos para cada técnica analisada, a dimensionalidade e tempo de processamento.

Um dos principais gargalos dos sistemas de recuperação de informação é o tempo de execução para a comparação do documento de consulta (*query*) com todos os outros documentos da base de dados. Por esse motivo, é importante também avaliar a eficiência das técnicas. A Tabela I mostra o tempo de processamento de cada técnica para recuperar os documentos similares. A técnica BoW, por possuir o vetor com a maior dimensionalidade entre as avaliadas, possui o maior tempo de processamento para recuperar documentos similares da base de dados. Ambas as abordagens, BoC e BoC-Th, possuem tempos de execução similares.

Além da precisão, outra vantagem da abordagem BoC-Th é a rápida execução no cálculo da similaridade entre os documentos da base de dados: por ter um vetor compacto, menos cálculos são necessários. Isso se reflete na escalabilidade da técnica proposta: a busca por documentos similares se torna mais eficiente em uma área cuja base de dados vêm crescente consideravelmente ao longo do tempo, devido a crescente digitalização de processos jurídicos.

5. CONCLUSÃO

Este trabalho apresentou uma nova abordagem, denominada BoC-Th, para representar documentos textuais jurídicos. O método proposto é uma extensão da abordagem tradicional BoC, ao contemplar uma frequência ponderada dos conceitos mais importantes na representação final dos documentos de acordo com o domínio da aplicação. A técnica BoC-Th faz uso de um tesauro (vocabulário especializado), com os termos/palavras do contexto da aplicação, o que permite enriquecer semanticamente o método BoC.

Resultados experimentais foram realizados com o objetivo de analisar o desempenho da abordagem proposta na recuperação de documentos jurídicos por meio do cálculo de similaridade semântica de

suas representações vetoriais. O método proposto BoC-Th foi avaliado comparativamente com as abordagens tradicionais BoW e BoC. O método proposto obteve desempenho 140% superior ao ser comparado ao modelo BoW e, em média, 51% mais eficaz ao obtido com o BoC, quando a clusterização de conceitos foi realizada com o algoritmo *Spherical K-means*. Dessa forma, demonstrou-se significativa vantagem com o uso da técnica BoC-Th para o estudo de caso analisado. Como trabalhos futuros, outras técnicas de agrupamento para a geração de conceitos deverão ser consideradas.

Em suma, a maior contribuição do método proposto neste trabalho é que ele incorpora as vantagens da abordagem BoC, tais como representação compacta, e oferece um desempenho de representação superior ao se considerar palavras/termos específicos de um domínio de aplicação, que no caso deste trabalho envolve textos jurídicos. É uma contribuição que enriquece uma área de aplicação com características peculiares, fornecendo um recurso de busca por informações textuais de forma mais precisa e com o menor tempo possível. Contudo, a técnica aqui apresentada também pode ser aplicada em domínios distintos do utilizado neste estudo.

REFERENCES

Analytics Vidhya, N. An intuitive understanding of word embeddings: From count vectors to word2vec, 2017.

Castano, S., Falduti, M., Ferrara, A., and Montanelli, S. A knowledge-centered framework for exploration and retrieval of legal documents. *Information Systems* vol. 106, pp. 101842, 2022.

Dal Pont, T. R., Sabo, I. C., Hübner, J. F., and Rover, A. J. Impact of text specificity and size on word embeddings performance: An empirical evaluation in brazilian legal domain. Springer-Verlag, Berlin, Heidelberg, pp. 521–535, 2020.

de Araujo, P. H. L., de Campos, T. E., and aes Silva de Sousa, M. M. Inferring the source of official texts: Can svm beat ulmfit? In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings*. Springer-Verlag, Berlin, Heidelberg, pp. 76–86, 2020.

de Campos, T. E., de Araujo, P. H. L., Braz, F. A., and da Silva, N. C. VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pp. 1449–1458, 2020.

Dhillon, I. S. and Modha, D. S. Concept decompositions for large sparse text data using clustering. *Machine learning* 42 (1): 143–175, 2001.

Kim, H. K., Kim, H., and Cho, S. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing* vol. 266, pp. 336–352, 2017.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara (Eds.). Proceedings of Machine Learning Research, vol. 32. PMLR, Bejing, China, pp. 1188–1196, 2014.

Martins, V. and Silva, C. Text classification in law area: a systematic review. In *Anais do IX Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, pp. 33–40, 2021.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Mourino Garcia, M. A., Perez Rodriguez, R., and Anido Rifon, L. E. Biomedical literature classification using encyclopedic knowledge: a wikipedia-based bag-of-concepts approach. *PeerJ (San Francisco, CA)* vol. 3, pp. e1279–e1279, 2015.

Noguti, M. Y., Vellasques, E., and Oliveira, L. S. Legal document classification: An application to law area prediction of petitions to public prosecution service. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*. IEEE, pp. 1–8, 2020.

Renjit, S. and Idicula, S. M. Cusat nlp@ aila-fire2019: Similarity in legal texts using document level embeddings. In *FIRE (Working Notes)*. pp. 25–30, 2019.

Silva, A. C. and Maia, L. C. G. The use of machine learning in the classification of electronic lawsuits: An application in the court of justice of minas gerais. In *Intelligent Systems*, R. Cerri and R. C. Prati (Eds.). Springer International Publishing, Cham, pp. 606–620, 2020.

Solihin, F., Budi, I., Aji, R. F., and Makarim, E. Advancement of information extraction use in legal documents. *International Review of Law, Computers & Technology* 35 (3): 322–351, 2021.

Yan, J. pp. 3069–3072. In L. LIU and M. T. ÖZSU (Eds.), *Text Representation*. Springer US, Boston, MA, pp. 3069–3072, 2009.