Modelagem de Tópicos para a Tarefa de Recuperação de Casos Legais

Luisa Pereira Novaes [1], **Daniela Vianna** [1], **Altigran da Silva** [1]

[1]Instituto de Computação, Universidade Federal do Amazonas (UFAM)
Manaus - AM - Brazil

{luisa.novaes, dvianna, alti}@icomp.ufam.edu.br

Abstract. *This article presents a topic-based approach to the problem of legal case retrieval. The method consists of two phases: filtering and ranking. In the first phase, a topic modeling technique is applied to the entire dataset to select an initial set of candidate cases for each query. In the second phase, a ranking function is used to produce an ordered list of relevant cases for the given query. Experimental results obtained using three different ranking functions and data collections in different languages indicate that the proposed approach is competitive. This is due to the strong correlation observed in our experiments between the topics of a query document and the topics of relevant legal cases. In fact, our approach achieved higher precision values than the ones reported from the recently held Competition on Legal Information Extraction/Entailment (COLIEE) 2023, competing with groups from around the world.*

Resumo. *Este artigo descreve uma abordagem baseada em tópicos para o problema de recuperação de casos jurídicos (*legal case retrieval*). O método consiste em duas fases: filtragem e ordenação. Na primeira fase, uma técnica de modelagem de tópicos é aplicada em todo o conjunto de dados para selecionar um conjunto inicial de casos candidatos para cada consulta. Na segunda fase, uma função de ordenação é usada para produzir uma lista ordenada de casos relevantes para a consulta fornecida. Resultados experimentais obtidos utilizando três diferentes funções de ordenação, com coleções de dados em diferentes idiomas, indicam que a abordagem proposta é competitiva, o que se deve à forte correlação, verificada em nossos experimentos, entre os tópicos de um documento-consulta e os tópicos dos casos jurídicos relevantes. De fato, nossa abordagem obteve melhores valores de precisão do que os reportados na recém-realizada Competition on Legal Information Extraction/Entailment (COLIEE) 2023, concorrendo com grupos de todo o mundo.*

1. Introdução

O problema de recuperação de documentos jurídicos (*legal case retrieval*) é uma especialização da área de recuperação da informação (*Information Retrieval*). O objetivo é, dado um documento jurídico, encontrar um ou mais documentos em uma dada coleção que deem suporte ao documento original. O uso de casos jurídicos similares é o pilar do sistema de *Commom Law* adotado por países como os EUA e o Canadá, onde decisões judiciais precisam ser baseadas em casos semelhantes decididos anteriormente por um tribunal de justiça. Mesmo países baseados na *Civil Law* ou lei estatutária, como é o

caso do Brasil, se apoiam fortemente em precedentes, ou decisões anteriores, durante a tomada de decisão. Em qualquer sistema jurídico, precedentes se tornam uma ferramenta fundamental na busca por um sistema mais justo e padronizado.

Considerando o número cada vez maior de novos casos jurídicos, a busca por casos similares torna-se uma tarefa exaustiva e ineficiente. Por exemplo, em 2021 o Brasil apresentou um total de 27.7 milhões de novos casos, sendo 97.2% dos casos submetidos eletronicamente[1]. A digitalização dos sistemas judiciários ao redor do mundo tem permitido a criação e aprimoramento de ferramentas inteligentes baseadas em aprendizado de máquina e processamento de linguagem natural para atender diversas demandas dos sistemas judiciários, incluindo a busca por precedentes ou casos similares. Embora a recuperação de documentos seja um campo bem estudado, a recuperação de casos jurídicos semelhantes impões uma série de novos desafios, como é o caso da complexidade da linguagem jurídica e o fato de documentos jurídicos serem normalmente documentos longos, muitas vezes trazendo discussões sobre vários tópicos.

Nesse artigo propomos uma abordagem baseada em tópicos [Grootendorst 2022] para o problema de recuperação de documentos jurídicos. Nossa solução consiste em duas fases: *filtragem* e *ordenação*. Na primeira fase, *filtragem*, aplicamos uma técnica de modelagem de tópicos a todos os documentos da coleção, incluindo os documentos-consulta (*query*) e todos os documentos candidatos. Importante ressaltar que durante essa fase inicial todos os documentos da coleção são considerados candidatos. O objetivo dessa primeira fase é identificar tópicos similares entre o documento-consulta e os candidatos, reduzindo consideravelmente o número de documentos a serem analisados na próxima fase. Duas técnicas de filtragem por tópicos são avaliadas nessa fase. Na segunda fase, *ordenação* ou *ranking*, três funções de ordenação são avaliadas com o intuito de selecionar o conjunto final de documentos relevantes dado um documento-consulta. A primeira função de ordenação se baseia na similaridade de cosseno entre documentos candidatos e o documento-consulta. A segunda função de ordenação, baseia-se no resultado da modelagem de tópicos executada na etapa de filtragem para identificar documentos candidatos com alta probabilidade de pertencerem ao mesmo tópico principal de um documento-consulta. Finalmente, a terceira função de ordenação é uma combinação das duas primeiras funções, ou seja, ela combina a similaridade de cosseno e a probabilidade de documentos pertenceram aos mesmo tópicos para definir a lista final de documentos relevantes para um documento-consulta.

Diversos experimentos foram realizados com duas coleções de documentos jurídicos. A primeira coleção foi disponibilizada pela recém-realizada *Competition on Legal Information Extraction/Entailment* (COLIEE) 2023. Essa coleção contém 4400 documentos jurídicos em Inglês. A segunda coleção foi extraída pelos autores do site do STJ (Supremo Tribunal de Justiça) e contém 1139 documentos jurídicos em Português. Essa coleção será compartilhada com a comunidade científica possibilitando o avanço de pesquisas envolvendo ao sistema jurídico brasileiro. Os experimentos mostram que existe uma forte correlação entre os tópicos de um documento-consulta e seus documentos relevantes, o que torna a etapa de uma estratégia eficaz para reduzir o número de documentos candidatos. Para a fase de ordenação, a função baseada em similaridade de cosseno produziu os melhores resultados, com oportunidades de melhorar ainda mais o resultado

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

explorando outras funções alternativas.

Esse artigo está organizado da seguinte forma: Na Seção 2, discutimos trabalhos relacionados. Na Seção 3, apresentamos a abordagem proposta nesse trabalho para o problema de recuperação de documentos jurídicos. A Seção 4 apresenta os conjuntos de dados usados em nossos experimentos, técnicas de pré-processamento e as métricas utilizadas para avaliação da nossa abordagem. Uma avaliação experimental do nosso método é apresentado na Seção 5. Por fim, conclusões e trabalhos futuros são discutidos na Seção 6.

2. Trabalhos Relacionados

Nos últimos anos, a Recuperação de Informações Jurídicas (RIJ) tem sido amplamente estudada em diferentes pesquisas [Sansone and Sperlí 2022]. Diferentes competições sobre tarefas de RIJ estão surgindo para analisar, por exemplo, a identificação de casos semelhantes ou sequências deles para apoiar a resolução de novos casos, como é o caso da Competition on Legal Information Extraction/Entailment (COLIEE).

No contexto de recuperação de casos jurídicos, Rabelo et al. [Rabelo et al. 2022] venceram a competição da Tarefa 1 do COLIEE 2022 ao propor uma abordagem baseada na similaridade entre parágrafos de casos jurídicos e na geração de vetores de características (*feature vectors*) com base nessas similaridades, seguida pelo uso de um classificador para determinar se os casos devem ser sinalizados como relevantes ou não. Na edição de 2017, Nanda et al. [Nanda et al. 2017] utilizaram modelos de tópicos para categorizar documentos em *clusters* de tópicos como parte de seus esforços para medir a similaridade entre consultas e documentos. Isso envolveu a atribuição de um vetor de tópicos a cada documento, que poderia então ser comparado ao vetor de tópicos da consulta, resultando na seleção dos n documentos iniciais mais similares à consulta. Os documentos mais relevantes foram posteriormente processados usando um modelo de similaridade semântica para identificar os documentos mais relevantes para a consulta dada.

Sansone & Sperlí [Sansone and Sperlí 2022] apresentam o estado da arte em IA para o domínio jurídico, com foco em sistemas de Recuperação de Informações Jurídicas baseados em técnicas de Processamento de Linguagem Natural, Aprendizado de Máquina e Extração de Conhecimento. Os autores destacam que abordagens baseadas em aprendizado profundo estão se tornando cada vez mais prevalentes no domínio jurídico.

Por fim, é interessante destacar também o trabalho de Mandal et al.[Mandal et al. 2021], no qual os autores observam que os métodos mais tradicionais como o TF-IDF [Jalilifard et al. 2021] e LDA [Park et al. 2009], que dependem de uma representação do tipo *bag-of-words*, têm um desempenho melhor do que os métodos mais avançados com reconhecimento de contexto, como BERT [Devlin et al. 2019] e Law2Vec[2] para calcular a similaridade em nível de documento.

Já no contexto de modelagem de tópicos, Silveira et al. [Silveira et al. 2021] usam o modelo LEGAL-BERT [Chalkidis et al. 2020] para gerar representações vetoriais densas ou *embeddings* de texto específicas do domínio legal, e o BERTopic [Grootendorst 2022] para modelar tópicos em documentos legais. Os resultados são animadores: 84,6% dos tópicos selecionados pelo modelo correspondem ao tema

[2]Law2Vec: Legal Word Embeddings. `https://archive.org/details/Law2Vec`

principal do documento. Em estudos mais recentes, técnicas de modelagem de tópicos são utilizadas para organizar coleções de documentos jurídicos [Vianna and Moura 2022, Vianna et al. 2023]. Inspirados nestes trabalhos, utilizamos a técnica proposta pelos autores em uma das etapas do nosso método.

3. Visão Geral

Em nossa abordagem, a recuperação dos documentos relevantes acontece em duas fases: *filtragem* e *ordenação*. A fase de filtragem, baseada em uma técnica de modelagem de tópicos, apresentada na Seção 3.1, é responsável por reduzir o conjunto de documentos candidatos para cada documento-consulta. Em seguida, na Seção 3.2, introduzimos três diferentes métodos de ordenação alternativos responsáveis por gerar a lista final de documentos relevantes por consulta.

3.1. Fase 1: Filtragem

Vianna & Moura [Vianna and Moura 2022, Vianna et al. 2023] apresentaram uma solução baseada em modelagem de tópicos para organizar coleções de documentos jurídicos. Nesse trabalho adotamos uma estratégia semelhante para etapa de filtragem. Nessa fase, um modelo de modelagem de tópicos, chamado *BERTopic* [Grootendorst 2022], é treinado utilizando todos os documentos da coleção, incluindo os documentos-consulta. A ideia do *BERTopic* é encontrar *clusters* densos de documentos a partir das representações vetoriais (*embeddings*) desses documentos e da combinação do algorítimo para redução de dimensionalidade UMAP [McInnes and Healy 2018] e do algorítimo de clusterização HDBSCAN [McInnes et al. 2017]. Para a extração de termos representativos de um tópico, *BERTopic* utiliza uma variação do clássico TF-IDF baseada em classes, denominado *c-TF-IDF*. *BERTopic* assume que a similaridade semântica entre documentos é um forte indicativo da existência de um tópico comum a eles. Partindo desse princípio, aplicamos *BERTopic* à coleção de documentos jurídicos com o intuito de reduzir o conjunto de documentos candidatos para uma determinada consulta, assumindo que existe uma forte relação entre os tópicos dos documentos-consulta e os tópicos dos seus respectivos candidatos. É importante ressaltar, que além de apontar o tópico dominante associado a um documento, *BERTopic* também retorna uma lista com os k tópicos que mais contribuem para a formação de um documento, além das probabilidades de cada um desses tópicos. Com isso, assumimos que um documento pode "pertencer" a diversos tópicos com diferentes níveis ou probabilidades de relevância . Durante a fase de filtragem, temos a flexibilidade de experimentar diferentes valores de k ao definir se um documento deve ou não fazer parte da lista reduzida de documentos candidatos para um determinado documento-consulta. Com base na lista de tópicos atribuída a cada documento, propomos duas abordagens distintas para a fase de filtragem:

Tipo 1: Tópico dominante da consulta. Se o tópico dominante de uma consulta aparecer nos principais k tópicos atribuídos a um documento, então este documento é considerado um candidato. O tópico dominante de uma consulta é o tópico dominante atribuído ao documento-consulta pelo *BERTopic*.

Tipo 2: Tópico dominante dos candidatos. Se o tópico dominante de um documento aparecer nos principais k tópicos atribuídos a um documento-consulta, então este documento é considerado um candidato para este documento-consulta.

Nessa fase, buscamos pelo valor ideal de k de maneira que o conjunto de documentos candidatos seja consideravelmente reduzido, acelerando e facilitando a fase de ordenação. Ao mesmo tempo temos que garantir que documentos relevantes não sejam erroneamente descartados nessa fase de filtragem. Diversos valores de k serão explorados de forma exaustiva na Seção 5.

Como mencionado anteriormente, *BERTopic* utiliza a representação vetorial dos documentos na busca por *clusters* de documentos semanticamente similares com base nos tópicos desses documentos. Essa representação vetorial (*embeddings*) dos documentos pode ser obtida utilizando diversas modelos, incluindo os modelos de linguagem baseados em *transformers*, como é o caso do BERT, e modelos neurais como o Doc2Vec [Le and 2014]. Considerando que documentos jurídicos costumam ser documentos longos, e que modelos como o BERT tem o tamanho da entrada restrito a um número pequeno de tokens, nesse trabalho optamos por usar o modelo Doc2Vec combinado com o *BERTopic* para gerar as representações vetoriais dos documentos da nossa coleção. O modelo Doc2Vec é treinado durante o treinamento do modelo de tópicos *BERTopic* usando os mesmos dados de treinamento (Seção 4.1).

3.2. Fase 2: Ordenação

Na fase de filtragem conjuntos de candidatos são gerados para cada documento-consulta. Em seguida, durante a fase de ordenação, funções de classificação são aplicadas aos conjuntos candidatos gerando uma lista ordenada de documentos relevantes para cada uma das consultas. Mais especificamente, nesse trabalho, três funções de ordenação alternativas foram avaliadas:

1. **Similaridade de Cosseno.** Baseia-se no modelo Doc2Vec treinado na fase anterior para gerar representações vetoriais (*embeddings*) para documentos-consulta e documentos-candidato. Em seguida, similaridade de cosseno é calculada para cada par de *embeddings* do documento-consulta e do documento-candidato, resultando em uma lista ordenada de documentos relevantes para cada documento-consulta.
2. **Contribuição dos Tópicos.** Leva em consideração as probabilidades dos tópicos obtidas pelo modelo *BERTopic* treinado na fase de filtragem. Especificamente, usamos a probabilidade do tópico dominante de um documento-consulta ser atribuído a um documento-candidato. Assumimos que um documento-candidato é relevante a um documento-consulta quando o tópico dominante do documento-consulta é também atribuído a um documento-candidato com alta probabilidade.
3. **Híbrida.** Combinação das duas abordagens anteriores: adiciona o valor obtido a partir da similaridade do cosseno à probabilidade obtida pela contribuição do tópico dominante do documento-consulta na formação do documento-candidato.

A aplicação de uma das três funções de ordenação à lista de candidatos gerada na fase de filtragem, resulta em uma lista ordenada de documentos-candidatos de acordo com sua relevância para o documento-consulta. Em seguida um corte final, baseado em um valor de corte ($corte$) pré-definido experimentalmente, pode ser aplicado a lista de candidatos resultando em uma lista final de documentos relevantes por consulta. Os valores ideais de k e $corte$ são definidos experimentalmente e serão detalhados na Seção 5.

4. Metodologia Experimental

Nesta seção apresentamos os dois conjuntos de dados usados em nossos experimentos, técnicas aplicadas para pré-processar os dados, além das métricas usadas para avaliação da nossa abordagem.

4.1. Coleção de Dados da COLIEE

Nosso trabalho foi inspirado pela Tarefa 1 da competição COLIEE, que envolve a leitura de um novo caso (documento) jurídico, denominado documento-consulta, e a recuperação de documentos que representam casos relevantes para o documento-consulta. Diferentemente das ferramentas tradicionais de recuperação de informação, nas quais a consulta consiste em um conjunto de palavras-chave, na tarefa de recuperação de documentos jurídicos explorada neste trabalho, a consulta caracteriza-se por ser um documento de texto.

O conjunto de dados da COLIEE 2023 é composto principalmente por casos do Tribunal Federal do Canadá fornecidos pela empresa *Compass Law*, todos em Inglês. Ela é dividida em um conjunto de treinamento com 4.400 documentos e um conjunto de teste com 1.334 documentos. Dentre os 4.400 documentos de treino, 959 documentos são apenas consultas (documentos-consulta), 4.110 atuam como documentos-candidatos e/ou documentos relevantes a uma consulta e 290 documentos não são nem consulta e nem relevantes a uma consulta. Em média, existem 4,7 documentos relevantes por consulta, podendo o número máximo de relevantes por consulta chegar a 34 e o número mínimo a 1 documento por consulta. No caso do conjunto de teste, dentre os 1.334 documentos disponíveis, 319 documentos-consultas foram disponibilizados. Ou seja, nosso objetivo é encontrar, dentre os 1.334 documentos, todos os documentos relevantes para cada um dos 319 documentos-consulta presentes no conjunto de testes.

4.2. Coleção de Dados do STJ

Um segundo conjunto de dados foi utilizado para avaliar nossa abordagem proposta em documentos jurídicos em português. Foi compilado pelos autores, tendo como base os documentos da ferramenta *Pesquisa Pronta*[3] do Supremo Tribunal de Justiça (STJ), que permite consultar entendimentos da corte sobre questões relevantes. Decisões consideradas relevantes para um determinado tema são agrupadas e disponibilizadas no site do tribunal, ou seja, para cada decisão disponibilizada no site, uma lista de casos relevantes a essa decisão é listada.

Para cada tema em questão, foram selecionados todos os casos jurídicos similares disponibilizados pelo tribunal, sendo um destes considerado como documento-consulta, e os demais casos foram considerados como documentos relevantes para tal consulta. O conjunto de dados apresenta 1139 documentos, sendo 225 consultas, 408 itens relevantes e outros 506 casos adicionais coletados do site do tribunal, os quais foram utilizados para auxiliar na fase de treinamento e para tornar o problema mais interessante e próximo ao coleção de dados da COLIEE. Entretanto, esses 506 casos não foram utilizados como consulta ou caso relevante. O número máximo de casos relevantes por consulta foi de 25, e o mínimo foi de 1. Em média, há aproximadamente 2,8 casos relevantes por documento-consulta.

[3] `https://scon.stj.jus.br/SCON/pesquisa_pronta/tabs.jsp`

4.3. Pré-processamento e Métricas de Avaliação

Realizamos uma etapa de pré-processamento nos dados de ambas as coleções usando diversos procedimentos para limpar e normalizar o texto. Primeiro, removemos todos os sinais de pontuação para evitar qualquer interferência que possam causar. Segundo, removemos todas as "stopwords", que são palavras que ocorrem comumente e não são relevantes, podendo distorcer nossa análise. Por fim, convertemos todo o texto para minúsculas para garantir consistência na representação das palavras e evitar a criação de múltiplas representações da mesma palavra. Esses procedimentos foram aplicados para garantir uma representação mais limpa e precisa dos dados de texto, o que, por fim, melhorou o desempenho de nossa abordagem proposta. É importante observar que essa etapa de pré-processamento foi aplicada antes de treinar os modelos Doc2Vec e *BERTopic* em ambas as coleções.

Para esta tarefa, seguimos as métricas de avaliação sugeridas pela COLIEE: precisão, revocação e a F-measure. F-measure foi calculada da seguinte forma: F-measure $=$ $(2 \times \text{Precisão} \times \text{Revocação})/(\text{Precisão} + \text{Revocação})$

5. Avaliação Experimental

5.1. Resultados com a Coleção COLIEE

Fase 1 – Filtragem: A primeira etapa da fase de filtragem é o treinamento dos modelos Doc2Vec e BERTopic usando os 4.400 documentos presentes no conjunto de treinamento da COLIEE 2023 (Seção 4.1). Após o treino, 119 tópicos foram encontrados na coleção. Além disso, usamos a probabilidade de contribuição de cada tópico estimada pelo BERTopic para identificar o(s) tópico(s) dominante(s) de cada documento para a filtragem (Seção 3.1). Para avaliar a eficácia da filtragem, medimos a *Revocação da Topificação*, razão de documentos relevantes que aparece no conjunto candidato resultante da etapa de filtragem e *Revocação Média da Topificação*.

Para exemplificar: Considerando *k=5* e a existência de 6 documentos relevantes para um determinado documento-consulta, se o tópico relevante desde documento-consulta está presente entre os 5 top tópicos de um documento candidato, consideramos esse documento relevante. Sendo assim, se o tópico relevante do documento-consulta é encontrado em apenas 2 dos 6 documentos candidatos, a revocação da topificação será de 2/6 = 33%. Aqui usamos a fórmula de revocação tradicional. Calculamos a Revocação da Topificação para todos os documentos-consulta do conjunto de dados, para diferentes valores de *k*, e depois calculamos a média desses resultados, que chamamos de *Média da Revocação da Topificação*.

Os resultados obtidos para a abordagem Tipo 1 da etapa de filtragem podem ser vistos na Figura 1. A *Revocação Média da Topificação* mostra que existe uma forte correlação entre o tópico dominante do documento-consulta e os tópicos dos documentos relevantes a essa consulta, tanto para o conjunto de treino quanto para o conjunto de teste. Por exemplo, para um valor de $k = 5$ (treino) já garantimos uma revocação média de 62%, ou seja, quando buscamos o tópico dominante do documento-consulta entre os cinco ($k = 5$) primeiros tópicos dos demais documentos, garantimos uma revocação de 62% em média. Este resultado mostra a eficácia da abordagem de modelagem de tópicos na redução do tamanho do conjunto de candidatos durante a fase de filtragem.

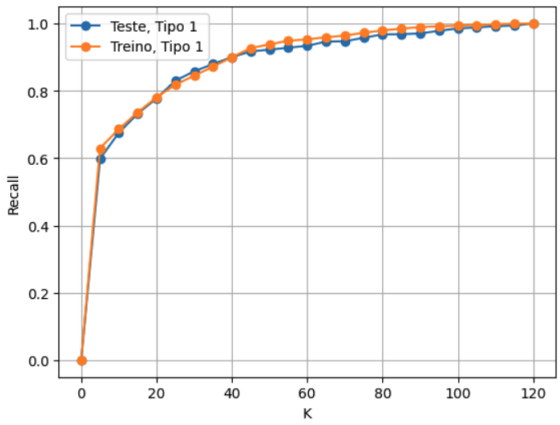

Figura 1. Revocação Média da Topificação para diferentes valores de k com as coleções de treino e teste da COLIEE.

Também foi avaliada a eficácia da abordagem Tipo 2 na redução dos conjuntos de candidatos de cada consulta, ou seja, dados os k tópicos dominantes do documento-consulta, buscamos entre os demais documentos aqueles que possuem um desses k tópicos como tópico dominante. Essa abordagem apresentou desempenho pior que o da abordagem Tipo 1 e, portanto, os resultados foram omitidos por questão de espaço.

Uma extensão dessas duas abordagens também foi considerada durante a avaliação da fase de filtragem. Além de filtrar os documentos com base na existência ou não do tópico dominante da consulta entre os k tópicos dos demais documentos, foi também considerado o impacto desses k tópicos na formação desses documentos. Esse impacto é medido pela porcentagem de contribuição de um tópico na construção de um determinado documento, ou seja, para ser considerado candidato, o tópico dominante do documento-consulta precisa aparecer nos k tópicos do possível candidato com um valor de contribuição acima de um valor limite pré-definido. No entanto, os resultados obtidos usando tal limite não se mostrou eficiente e consequentemente omitimos os resultados para poupar espaço no artigo.

Ao final da etapa de filtragem temos um modelo Doc2Vec e um modelo BERTopic treinado usando o conjunto de treinamento da coleção COLIEE. Esses dois modelos podem ser aplicados aos 1.334 documentos do conjunto de teste para inferirmos o percentual de contribuição dos 119 tópicos para cada um desses documentos e então realizar a filtragem dos documentos candidatos para cada um dos 319 documentos-consulta. Lembrando que o modelo Doc2Vec é utilizado para gerar uma representação vetorial de cada documento que, em seguida, é servido de entrada para a modelagem de tópicos BERTopic, responsável por calcular o percentual de contribuição dos tópicos para cada um dos documentos de entrada.

Fase 2 – Ordenação: Como apresentado na Seção 3.2, três funções de ordenação foram exploradas neste trabalho: similaridade de cossenos, porcentagem de contribuição do tópico dominante dos documentos-consulta com relação aos tópicos dos demais documentos e função híbrida que combina tanto a similaridade de cosseno quanto a contribuição percentual de tópicos. Após a aplicação de uma das três funções de ordenação, um valor de corte ($corte$) é aplicada a lista ordenada de candidatos reduzindo ainda mais a lista

final de documentos relevantes por consulta. Valores de k (fase de filtragem) e $corte$ (fase de ordenação) são estimados experimentalmente durante o treinamento.

Experimentos foram realizados com o conjunto de treino da COLIEE para definir valores de k e $corte$. Os melhores resultados foram obtidos para $k = 50$ e $corte$ no intervalo entre $0,32$ e $0,41$. Usando esses parâmetros, a Tabela 1 apresenta os resultados obtidos no conjunto de teste após a etapa de ordenação usando a função que tem como base a similaridade de cosseno. O melhor resultado para F-measure no conjunto de teste foi obtido com $corte = 0,33$, seguido pelas soluções com $corte = 0,34$ e $corte = 0,32$, respectivamente.

Considerando como baselines as 5 melhores submissões da COLIEE 2023 (Tabela 2), nossos melhores valores de F-measure foram inferiores. O F-measure do vencedor foi de 0,3001, o que evidencia a complexidade do problema. No entanto, nossos melhores valores ficaram acima da média das 22 submissões[4] que foi de 0,255. Por outro lado, os melhores valores de precisão obtidos em nossos experimentos foram superiores aos valores obtidos em todas as submissões. Esses resultados indicam a eficácia da modelagem de tópicos na seleção do conjunto de documentos candidatos. Consideramos esta uma das grandes vantagens da nossa abordagem, uma vez que para o operador do direito é de suma importância que os documentos retornados sejam relevantes.

Tabela 1. F-measure, Precisão e Revocação do conjunto de teste, com k = 50, Similaridade de Cosseno e variando $corte$.

corte	0,32	0,33	0,34	0,35	0,36	0,37	0,38	0,39	0,4	0,41
F-measure	**0,2625**	**0,2685**	**0,2655**	0,2618	0,2545	0,2424	0,2345	0,2156	0,2050	0,1973
Precisão	0,2284	0,2500	0,2644	0,2819	0,2975	0,3071	0,3199	0,3182	0,3342	0,3578
Revocação	0,3085	0,2899	0,2666	0,2445	0,2224	0,2002	0,1851	0,1630	0,1478	0,1362

Tabela 2. Resultados das 5 melhores submissões da Tarefa 1 na COLIEE 2023.

Submissão	F-measure	Precisão	Revocação
THUIR	**0,3001**	0,2379	0,4063
THUIR	0,2907	0,2173	**0,4389**
IITDLI	0,2874	**0,2447**	0,3481
THUIR	0,2771	0,2186	0,3783
NOWJ	0,2757	0,2263	0,3527

Os valores relativamente mais baixos de F-measure obtidos em nossos experimentos são devidos à baixa revocação em algumas consultas. De fato, observamos que para algumas consultas, nossas soluções não geram uma lista final de documentos relevantes. Isso ocorre porque, mesmo com a melhor função de ordenação, similaridade de cosseno, o $corte$ acaba eliminando todos os possíveis candidatos devido à baixa similaridade entre consulta e candidatos. Apesar disso consideramos que o uso do corte é importante para evitar a recuperação de documentos irrelevantes, pois isso afeta negativamente a precisão.

5.2. Resultados com a Coleção STJ

Fase 1 – Filtragem: Diferentemente da coleção da COLIEE onde tínhamos conjuntos de treino e teste pré-definidos, para aumentar a confiança nos resultados obtidos com a coleção do STJ optamos por usar uma estratégia de validação cruzada, criando 10 variações

[4] `https://sites.ualberta.ca/~rabelo/COLIEE2023/task1_results.html`

aleatórias do conjunto original, onde 70% dos 1.139 documentos foram usados para treino e 30% para teste. Em seguida, a abordagem proposta é aplicada a cada um dos 10 conjuntos gerados. Quando a modelagem de tópicos é aplicada a estes conjuntos em média um total de 32,5 tópicos distintos foram identificados. Estes valores variam a depender dos documentos-consulta aleatoriamente selecionados para o conjunto de treino. Com as porcentagens de contribuição de tópicos computadas para cada documento, pudemos discernir o(s) tópico(s) dominante(s) para cada documento. Para avaliar a eficácia desses tópicos como indicadores de relevância, usamos as duas abordagens de filtragem introduzidas na Seção 3.1. Novamente, usamos a *Revocação da Topificação* e a *Revocação Média da Topificação* para medir o desempenho dessas duas abordagens. Os resultados da *Revocação Média da Topificação* para diferentes valores de k é apresentado na Figura 2, tanto para o treino quanto para o teste. Neste caso, os valores correspondem às médias com cada conjunto aleatório.

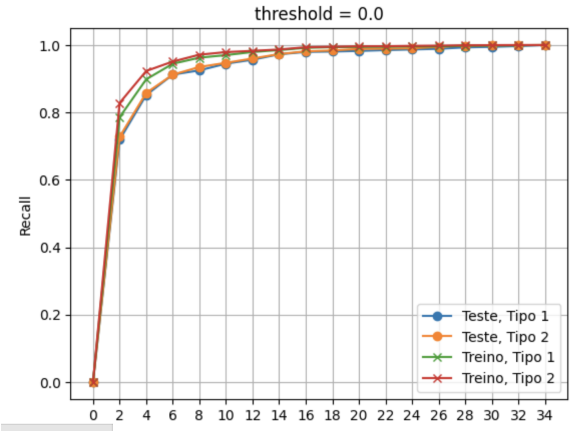

Figura 2. Revocação Média da Topificação para diferentes valores de k e diferentes funções de filtragem com a coleção do STJ.

Assim como no caso da COLIEE, observa-se uma correlação, nesse caso ainda mais forte, entre os tópicos de um documento-consulta e os tópicos de seus documentos relevantes. Diferentemente do que ocorreu com a COLIEE, para essa coleção em Português a abordagem de filtragem Tipo 2 identificou uma correlação ainda mais forte para determinados valores de k que a abordagem Tipo 1. Se olharmos para o conjunto de treino, já podemos encontrar mais de 78% dos casos relevantes para uma consulta ao procurar o tópico dominante da consulta nos primeiros 2 tópicos dominantes dos candidatos ($k = 2$, função de filtragem Tipo 1), em média. Ao procurar o tópico dominante da resposta nos 2 primeiros tópicos dominantes da consulta ($k = 2$, função de filtragem Tipo 2), já encontramos 82% dos casos relevantes para uma consulta. Este resultado reafirma a eficácia da abordagem de modelagem de tópicos na redução do tamanho do conjunto de candidatos durante esta fase. Por essa razão, optamos por utilizar a função Tipo 2 para o restante do experimento com a coleção do STJ.

Fase 2 – Ordenação: Após a fase de filtragem, a fase de ordenação é aplicada à lista de candidatos, ordenando todos os documentos com base em alguma medida de similaridade (função de ordenação) entre documentos e a consulta. Como introduzido na Seção 3.2, as mesmas três funções de ordenação foram exploradas para avaliar essa coleção de da-

dos. Como no caso da COLIEE, para os dados de treino do STJ os resultados obtidos evidenciam que a função baseada na similaridade de cosseno apresentou resultados muito superiores em relação às outras funções. Por este motivo, escolhemos a função baseada na similaridade de cosseno como a melhor para etapa de ordenação.

Após a aplicação da função de ordenação, que produz uma lista ordenada de documentos candidatos em relação a similaridade com o documento-consulta, um $corte$ (valor de corte) é aplicado para selecionar a lista final de candidatos por consulta. Para estimar os valores de k (fase de filtragem) e o intervalo de $corte$ (fase de ordenação), realizamos um conjunto de experimentos com os dados de treino. Os melhores resultados foram obtidos para $k = 2$ e $corte$ no intervalo entre $0,36$ e $0,48$. Usando esses parâmetros, a Tabela 3 apresenta a média dos resultados nos 10 conjuntos de teste após a etapa de ordenação usando a função que tem como base a similaridade de cosseno. A melhor média dos resultados para F-measure no conjunto de teste foi obtido com $corte = 0,36$, seguido pelas soluções com $corte = 0,38$ e $corte = 0,37$, respectivamente.

Tabela 3. F-measure, Precisão e Revocação no conjunto de teste do STJ, com $k = 2$, Similaridade de Cosseno, Filtragem Tipo 2 e variando $corte$

Corte	0,36	0,37	0,38	0,39	0,40	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48
F-measure	**0,3937**	**0,3782**	**0,3681**	0,3593	0,3450	0,3338	0,3303	0,3225	0,3241	0,3135	0,2979	0,2900	0,2771
Precisão	0,2858	0,2726	0,2599	0,2512	0,2410	0,2265	0,2226	0,2119	0,2096	0,2016	0,1899	0,1841	0,1743
Revocação	0,6460	0,6404	0,6695	0,6884	0,7169	0,7250	0,7308	0,7172	0,7575	0,7549	0,7627	0,7712	0,7712

Em média, os resultados obtidos são consistentes com os resultados obtidos com a coleção COLIEE, com uma melhoria nos valores de precisão. Por outro lado, houve uma melhoria bastante expressiva em termos de revocação. Isso pode ser explicado pelo fato dessa coleção ser composta a partir de documentos curados manualmente por especialistas do STJ. Isso faz com que os documentos tenha tópicos mais bem definidos, característica que é explorada pela modelagem de tópicos. Uma evidência desta hipótese é o baixo valor de k, ou seja, o número de tópicos considerados, que é 2 no caso da coleção STJ e 50 no caso da COLIEE. É também possível que o tamanho da coleção tenha alguma relevância o resultado, mas se os documentos não fosse bastante relacionados, a questão do tamanho não seria determinante. Deixamos como trabalho futuro, uma investigação mais aprofundada destes e outros possível fatores críticos para a nossa abordagem.

6. Conclusões e Trabalhos Futuros

Neste artigo, apresentamos uma abordagem baseada em tópicos para a tarefa de recuperação de casos jurídicos. A abordagem proposta tem duas fases: filtragem e ordenação. Na fase de filtragem, um método de descoberta de tópicos é aplicado para selecionar um conjunto inicial de candidatos para a consulta. Na fase de ordenação, uma função de ordenação é aplicada ao conjunto inicial de candidatos, seguidas por um corte usando um limite definido experimentalmente, gerando a lista final de documentos relevantes.

Dois conjuntos de dados foram utilizados em nossos experimentos. O primeiro é composto por documentos em Inglês usados na COLIEE (Competition on Legal Information Extraction/Entailment), e o segundo é formado por documentos jurídicos, em Português, da ferramenta *Pesquisa Pronta*, que foi coletado pelos autores do site do STJ e será disponibilizado para a comunidade. Experimentos foram realizados com o conjunto

de teste de cada coleção para definir os parâmetros adequados (k e $corte$) para cada conjunto de dados. Os resultados obtidos com ambas as coleções demonstraram a eficácia da modelagem de tópicos na seleção do conjunto de documentos candidatos, em ambas as coleções.

Com a coleção da COLIEE, os melhores valores de precisão obtidos em nossos experimentos são superiores aos valores obtidos em todas as submissões da competição. Sabendo que, para o operador do direito, é de suma importância que os documentos retornados sejam relevantes, esta é uma grande vantagem da nossa abordagem. Além disso, nossos melhores valores de F-measure estão acima da média das 22 submissões. Para a coleção do STJ, os resultados obtidos reforçam a eficácia do modelo, com melhorias nos valores de precisão e outra bastante expressiva em termos de revocação. Em ambas as coleções, os resultados mostram ainda que há espaço para melhorias na fase de ordenação, o que pode levar a resultados mais expressivos no futuro.

Dados os resultados obtidos, como próximo passo, planejamos investigar diferentes funções de ranqueamento a serem usadas durante a fase de ordenação, além de explorar a hierarquia dos tópicos com o BERTopic para melhor agrupar estes documentos entre tópicos relacionados e tentar chegar construir uma nova função de ordenação. Iremos também explorar diferentes modelos de linguagem para serem aplicados na geração dos embeddings, na fase de filtragem e na fase de ordenação. Por fim, temos planos de investigar com mais profundidade fatores críticos para a nossa abordagem, como tamanho da coleção, idioma, relacionamento entre os documentos, entre outros. Ter um domínio maior sobre estes fatores nos auxiliará a entender melhor os resultados encontrados e como aperfeiçoar a abordagem proposta.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001 financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD, e pelo CNPq através de uma bolsa PQ para Altigran da Silva (Proc. 307248/2019-4).

Referências

Chalkidis, I. et al. (2020). Legal-bert: The muppets straight out of law school.

Devlin, J. et al. (2019). 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, NAACL-HLT*, pages 4171–4186.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Jalilifard, A. et al. (2021). Semantic sensitive tf-idf to determine word relevance in documents. In *Advances in Computing and Network Communications: Proceedings of CoCoNet*, pages 327–337.

Le, Q. and , T. M. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - ICML*, page II–1188–II–1196.

Mandal, A. et al. (2021). Unsupervised approaches for measuring textual similarity between legal court case reports. *Artif. Intell. Law*, 29(3):417–451.

McInnes, L. et al. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2:205.

McInnes, L. and Healy, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction.

Nanda, R. et al. (2017). Legal information retrieval using topic clustering and neural networks. In *4th Competition on Legal Information Extraction and Entailment (COLIEE)*, pages 68–78.

Park, L. A. et al. (2009). The sensitivity of latent dirichlet allocation for information retrieval. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*, pages 176–188.

Rabelo, J. et al. (2022). Semantic-based classification of relevant case law. In *New Frontiers in Artificial Intelligence - JSAI-isAI*, pages 84–95.

Sansone, C. and Sperlí, G. (2022). Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Silveira, R. et al. (2021). Topic modelling of legal documents via legal-bert1. In *Proceedings http://ceur-ws org ISSN, 1613:0073*.

Vianna, D. and Moura, E. (2022). Organizing portuguese legal documents through topic discovery. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 3388–3392.

Vianna, D., Moura, E., and Silva, A. (2023). A topic discovery approach for unsupervised organization of legal document collections. *Artificial Intelligence and Law*, pages 1–30.