Técnicas de sumarização de textos jurídicos para suporte à classificação de documentos de decisões judiciais

Hellen Harada[1] , Fabíola Pereira[1] , Alex Almeida[2], Daniela Freire[3] , Márcio Dias[4], Nádia Silva[5] , Pedro Andrade[5] , André Carvalho[3]

[1]Faculdade de Computação
Universidade Federal de Uberlândia

[2]Faculdade de Tecnologia de Ourinhos

[3]Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

[4]Departamento de Ciência da Computação
Universidade Federal de Catalão

[5]Instituto de Informática
Universidade Federal de Goiás

{hellen.harada,fabiola.pereira}@ufu.br, alex.marino@fatecourinhos.edu.br,

{danielalfreire,andre}@usp.br, marciodias@ufcat.edu.br,

{andrdpedro,nadia.felix}@ufg.br

Resumo. *Acórdãos são documentos de texto que contêm decisões judiciais referentes a um certo processo jurídico. No contexto de um Tribunal de Justiça, os acórdãos possuem uma classificação por temas bem definida, que auxilia juristas na organização e agilidade de suas tarefas diárias. Devido ao alto volume diário de novos acórdãos produzidos, faz-se necessária a adoção de técnicas capazes de automatizar a classificação temática de um novo acórdão. Algoritmos de aprendizado de máquina supervisionado, para tarefas de classificação, não têm se saído bem diante de textos extensos, em português, com a linguagem usada no domínio jurídico. Este trabalho propõe a adoção de sumários de acórdãos para classificação temática. A hipótese levantada é que textos mais curtos, sumarizados, podem melhorar a classificação de tais documentos nos temas corretos. Este é um trabalho em andamento, que pretende desenvolver uma nova abordagem de classificação a partir de sumários. Resultados parciais indicam que algoritmos de sumarização melhoram a classificação de acórdãos.*

1. Introdução

Ao julgar um processo, o magistrado pode realizar despachos, decisões interlocutórias e sentenças. Os despachos, as decisões e sentenças são redigidos, datados e assinados pelos juízes, enquanto os acórdãos são feitos pelos desembargadores. Neste trabalho, os documentos judiciais abordados serão os acórdãos. Os acórdãos possuem uma classificação por temas do Superior Tribunal de Justiça (STJ), que são categorias baseadas nos fundamentos das decisões, bem como na legislação utilizada para embasar tais fundamentos. A estrutura de tematização é hierárquica, com a existência de subtemas, e dinâmica, com a

possibilidade de surgimento de novos temas. Seis temas gerais compõem o primeiro nível da árvore de hierarquia.

A classificação por temas é uma prática que auxilia os juristas na organização e agilidade do dia a dia, facilitando futuras decisões por similaridade entre os processos. O volume diário é de aproximadamente 100 novos acórdãos e, por isso, faz-se necessária a automatização do processo por meio de técnicas de Inteligência Artificial como um todo.

Os documentos dos acórdãos são, em geral, extensos e não possuem uma estrutura padrão. A classificação de acórdãos em temas pode ser modelada como um problema de aprendizado supervisionado. Entretanto, utilizar representações de texto que consideram o texto completo na construção dos modelos pode reduzir a eficácia da classificação, gerando indesejáveis instâncias de falsos positivos/negativos [Wang et al. 2021]. Sumarizar os textos em busca de seus segmentos mais representativos antes do processo de classificação pode ser uma promissora abordagem.

Dessa forma, a pergunta que se pretende responder neste trabalho é: *utilizar os acórdãos de maneira sumarizada pode aumentar a eficácia na classificação de tais documentos em relação aos seus temas?*

2. Fundamentação Teórica e Trabalhos Correlatos

Sumarização é a escrita de um texto mais curto comparado ao texto original e que permanece com a mesma ideia. Um resumo. Em termos de formação, sumários podem ser classificados como extrativos ou generativos.

Sumários extrativos são sumários compostos por partes inalteradas do texto original, de forma que o sumário seja composto das partes mais importantes do texto, sem haver modificações. Sumários generativos, também conhecidos como abstrativos são sumários feitos por meio da reescrita, havendo alteração dos seus trechos em comparação com o original, de forma que um pequeno texto passe toda a ideia principal do texto sem se preocupar com o modo de escrita original.

A sumarização automática de texto permite que os usuários compreendam e comparem rapidamente temas em determinados *corpora*. Tornou-se cada vez mais importante com o acúmulo crescente de documentos de texto em todos os campos [Wang et al. 2021].

Neste trabalho o foco é na sumarização extrativa por documento (monodocumento), ou seja, aquela que seleciona os segmentos mais importantes de um documento e os concatena para formar um sumário [El-Kassas et al. 2021].

Existem diversas técnicas bem estabelecidas para sumarização automática de textos, como por exemplo técnicas baseadas em grafos, em semântica e em centralidade de sentenças [El-Kassas et al. 2021]. Da mesma forma, existem muitas técnicas bem estabelecidas para classificação de textos, em especial textos jurídicos [Chen et al. 2022].

Entretanto, pouco ainda foi explorado acerca da classificação de textos com o auxílio de técnicas de sumarização. Em [Rahamat Basha et al. 2019] é proposto um novo método de seleção de características para o classificador KNN (*K-nearest neighbor*) resumindo os documentos de treinamento originais com base na medida de importância da sentença. A abordagem para sumarização de documento único usa duas medidas para similaridade de sentenças: a frequência dos termos em uma sentença e a similaridade dessa

sentença com outras sentenças.

Em [Jeong et al. 2016] os autores propõem um interessante *framework* que utiliza tanto informações de resumo quanto das categorais. Um modelo de língua é utilizado para combinar distribuições de recursos em cada categoria e texto, e um modelo para classificação de texto faz as pontuações de importância de sentença estimadas a partir da sumarização de texto.

Em [Du et al. 2021] a sumarização é utilizada para construção de exemplos em uma abordagem de *text augmentation*, para resolver o problema de limitações de anotações.

Nenhum destes trabalhos possui foco em textos jurídicos.

3. Metodologia

A metodologia de desenvolvimento do trabalho está organizada em 4 etapas, descritas a seguir.

3.1. Desenvolvimento de abordagens de sumarização

O primeiro passo do trabalho foi o desenvolvimento de abordagens de sumarização existentes na literatura. O objetivo foi reunir 4 técnicas bem estabelecidas na literatura (KL-Soma, LexRank, LSA, Luhn) [El-Kassas et al. 2021], todas elas considerando a abordagem de sumarização extrativa monodocumento. Esta etapa foi importante para o entendimento do problema e percepção do impacto dos diferentes algoritmos.

3.2. Anotação do *corpus* por especialistas

Está sendo preparado um *corpus* para que especialistas realizem a anotação que, neste caso, consiste em gerar manualmente sumários em uma amostra de decisões judiciais, formando o *gold standard dataset*. Será utilizada a ferramenta Inception [Klie et al. 2018]. É parte também desta etapa o desenvolvimento de um plano de anotação que guiará a metodologia a ser seguida pelos anotadores.

3.3. Avaliação dos sumários

Existem métricas supervisionadas e não supervisionadas para avaliação de sumários extrativos. As métricas não supervisionadas estão mais voltadas a uma avaliação quantitativa, com foco nas separações de segmentos encontradas pelo algoritmo de sumarização. Elas serão aplicadas primeiro aos sumários obtidos.

Em posse do *gold standard dataset* gerado por especialistas, será possível tanto uma avaliação qualitativa quanto uma avaliação quantitativa considerando métricas supervisionadas. É esperado que o conjunto de sumários anotados enriqueça a avaliação e seja um diferencial no trabalho.

3.4. Execução e avaliação de classificadores sobre sumários

Os algoritmos de classificação: árvores de decisão, SVM, MLP e CNN) serão executados e avaliados sobre os sumários e sobre os textos originais. O objetivo será comparar com os resultados das abordagens de classificação que não envolvem sumários. A Figura 1 ilustra a solução a ser desenvolvida.

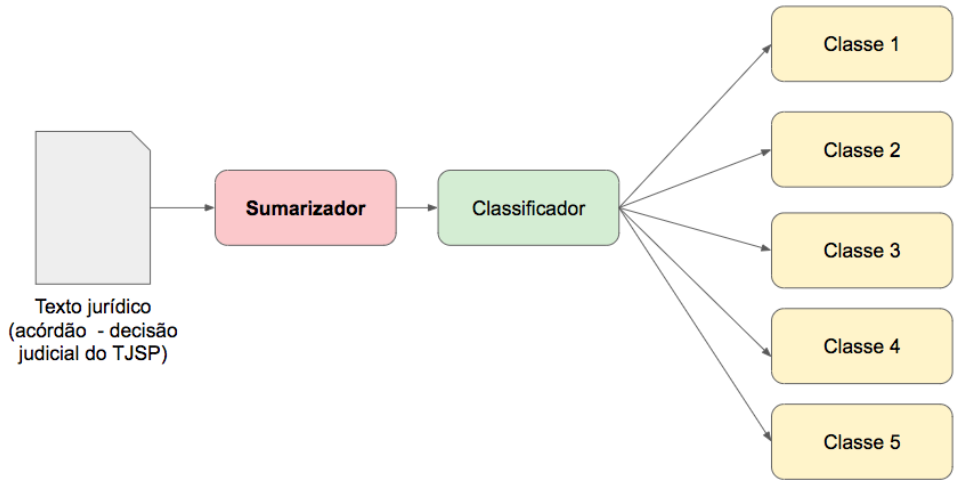

Figura 1. Etapas da solução sendo desenvolvida neste trabalho.

4. Experimentos iniciais

Foram selecionados aleatoriamente 3 acórdãos para execução dos algoritmos de sumarização selecionados. Os algoritmos LexRank, Luhn, LSA e KL-SUM foram utilizados para cada texto. Em cada execução, foi passado o parâmetro k=5, indicando o número de sentenças a comporem os sumários resultantes.

Observou-se uma divergência grande entre sumários gerados por diferentes algoritmos. A tabela 4 ilustra a similaridade de Jaccard obtida entre os sumários gerados pelos respectivos algoritmos, considerando os 3 acórdãos selecionados.

Tabela 1. Índice Jaccard obtido entre os sumários gerados pelos algoritmos LexRank, Luhn, KL-Sum e LSA.

	LexRank & Luhn	LexRank & KL-Sum	LexRank & LSA	Luhn & KL-Sum	Luhn & LSA	KL-Sum & LSA
Acórdão 1	0	0.25	0.25	0.1	0	0.1
Acórdão 2	0.25	0.25	0.428	0.25	0.1	0.1
Acórdão 3	0.1	0.1	0	0.1	0	0

5. Conclusão e Trabalhos Futuros

A implementação e execução dos sumários nos textos jurídicos mostrou-se efetiva e factível. Trata-se de um trabalho em andamento. Atualmente, está sendo elaborado um *corpus* anotado por especialistas para se tornar o *dataset* de referência para avaliação dos sumários.

6. Agradecimentos

Os autores agradecem o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (CNPq/MCTI/SEMPI Nº 56/2022).

Referências

Chen, H., Wu, L., Chen, J., Lu, W., and Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Inf. Process. Manage.*, 59(2).

Du, Y., Ma, T., Wu, L., Xu, F., Zhang, X., and Ji, S. (2021). Constructing contrastive samples via summarization for text classification with limited annotations. In *EMNLP*.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Jeong, H., Ko, Y., and Seo, J. (2016). How to improve text summarization and classification by mutual cooperation on an integrated framework. *Expert Syst. Appl.*, 60(C):222–233.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Rahamat Basha, S., Keziya Rani, J., and Prasad Yadav, J. J. C. (2019). A novel summarization-based approach for feature reduction enhancing text classification accuracy. *Engineering, Technology amp; Applied Science Research*, 9(6):5001–5005.

Wang, F., Zhang, J. L., Li, Y., Deng, K., and Liu, J. S. (2021). Bayesian text classification and summarization via a class-specified topic model. *J. Mach. Learn. Res.*, 22(1).