IX Semana de Ciência e Tecnologia do ICE
SECT - ICE
Jubileu de Ouro
50 ANOS
INSTITUTO DE CIÊNCIAS EXATAS
Cinco Décadas de Ciência, Tecnologia e Inovação no Estado do Amazonas
18 a 21 de outubro de 2022

Aplicação do algoritmo *CatBoost* em textos jurídicos e na Agenda 2030 da ONU

Lucas José Gonçalves Freitas[1] (PG)*, Pamella Sada Dias Edokawa[2] (PQ), Ariane Hayana Thomé de Farias[2] (PQ), Euler Rodrigues de Alencar[2] (PQ), Thaís Carvalho Valadares Rodrigues[3] (PQ)

* lucas.freitas@stf.jus.br.

1.Escritório de Estatística, STF/Departamento de Estatística, UnB. 2.Secretaria de Gestão Estratégica, STF. 3.Departamento de Estatística, UnB.

Palavras Chave: Machine Learning, Agenda 2030.

Introdução

O Supremo Tribunal Federal (STF), instância máxima do sistema judiciário brasileiro, produz muitos dados organizados em forma de texto, por meio de decisões, petições, liminares, recursos e outros documentos legais. Neste contexto, uma ferramenta de aprendizagem supervisionada centrada na classificação de documentos jurídicos nos 17 objetivos (ODS) da Agenda 2030 da ONU pode ser de grande utilidade para o tribunal, uma vez que essa tarefa é realizada de forma manual por um grande grupo de funcionários da corte. O objetivo geral deste trabalho é utilizar o algoritmo *CatBoost*, que permite combinar features em texto e variáveis explicativas, para sugerir etiquetas de ODS da Agenda ONU 2030 de forma automática para peças jurídicas. As variáveis explicativas escolhidas foram: quantidade de palavras-chave de cada ODS na peça em tela, classe processual e ramo do direito. As palavras-chave são coleções de palavras e tokens relacionados à cada ODS da Agenda 2030, produzidos por especialistas da área jurídica do próprio tribunal. Feminicídio, por exemplo, é uma palavra-chave relacionada ao ODS 5 – Igualdade de Gênero. Classe processual e ramo do direito são metadados típicos de registro em autuação processual.

Material e Métodos

A base de dados é composta por cerca de 2100 petições e acórdãos de processos devidamente rotulados com ODS da Agenda 2030 da ONU. Esta quantidade de processos é substancialmente menor do que o conjunto de processos rotulados no Supremo Tribunal Federal e isso ocorre porque nem todos os processos etiquetados tem petições iniciais ou acórdãos disponíveis e legíveis. O adjetivo legível se aplica a processos com peças em PDF nato digital ou digitalizados via OCR (*Tesseract*), tarefa que extrai, através de redes neurais LSTM, textos de imagens. Para aplicação dos algoritmos da família *CatBoost* pode-se utilizar o fluxo básico de limpeza em textos jurídicos, como na lista a seguir:

- Remoção de stopwords e pontuações;
- Remoção de caracteres especiais;
- Criação de tokens com leis;
- Lowercase de todas as palavras e remoção de números desnecessários.

Aos passos básicos de limpeza de texto, soma-se uma etapa de pos tag (do inglês part of speech), atividade que realiza uma espécie de classificação morfológica das palavras de um texto. O objetivo é tornar os textos limpos mais densos, mantendo apenas verbos, substantivos e adjetivos.

Os modelos *CatBoost* que serão utilizados no presente trabalho são algoritmos de ensemble em machine learning, que usam gradient boosting e árvores simétricas de decisão, criados pela Yandex em 2017 [1]. Por possuírem, nativamente, suporte para variáveis categóricas e utilizar métodos de embedding centrados em abordagens de dicionário, tais algoritmos podem ser utilizados largamente em contextos jurídicos. As principais diferenças entre os algoritmos da família *CatBoost* e outros modelos de *boosting* (como XGBoost e LightGBM) são dadas pela natureza das árvores de decisão, que podem ser simétricas ou assimétricas, e pela implementação do gradiente, que no caso do *CatBoost* é ordenado. Na prática, a ordenação do gradiente permite que apenas dados não utilizados na etapa de treinamento sejam utilizados para avaliar localmente a performance do algoritmo.

O presente trabalho pretende comparar a performance do algoritmo *CatBoost* em cenários diversos, a saber: com ou sem pos-tag e com todas as combinações possíveis de textos e variáveis categóricas. A lista a seguir mostra todos os cenários:

- cenário 1: classe processual;
- cenário 2: ramo do direito;
- cenário 3: contagem de palavras-chave;
- cenário 4: classe processual e ramo do direito;
- cenário 5: classe processual e contagem de palavras-chave;
- cenário 6: ramo do direito e contagem de palavras-chave;
- cenário 7: classe processual, ramo do direito e contagem de palavras-chave.

Considerando ainda a presença ou não da atividade de pos tag, são 14 avaliações ao todo. O objetivo maior deste trabalho é verificar se o algoritmo *CatBoost* é capaz de atingir uma boa acurácia na predição de etiquetas ODS, mas também merecem destaque a avaliação do passo de pos tag na limpeza dos textos e análise descritiva de cada uma das variáveis explicativas escolhidas. Como as peças jurídicas podem ser grandes e o tamanho dos textos aumenta de maneira significativa a complexidade dos algoritmos de processamento de linguagem natural (NLP),

ter mais uma opção viável para tornar os registros densos e menores pode ser interessante, ao passo que observar padrões nos registros de classe processual e ramo do direito pode ser interessante para especialistas na Agenda 2030. Alguns Objetivos de Desenvolvimento Sustentável (ODS) ainda não possuem processos etiquetados em número suficiente para aprendizagem de máquina. A distribuição das etiquetas pode ser consultada no painel gerencial disponível no hotsite Agenda 2030 no STF [2]. Por este motivo, o presente trabalho se ocupa de aplicar a metodologia proposta apenas para o ODS 16, cujo tema é Paz, Justiça e Instituições Eficazes, dado que tal etiqueta possui muitas entradas - cerca de 56% dos processos avaliados. Este Objetivo de Desenvolvimento Sustentável é, também, o de maior complexidade para classificação pelos especialistas do tribunal, pelo caráter constitucional do Supremo Tribunal Federal (STF).

Resultados e Discussão

Os resultados iniciais apontam para aumento do desempenho usando a abordagem pos tag, o que faz enorme sentido, uma vez que os textos diminuem entre 25% e 30% do seu tamanho original (em média). Foram separados 85% do conjunto de dados inicial para treinamento e os 15% restantes para testes. O recorte com a abordagem pos tag, contagem de palavras-chave e classe processual foi o de melhor performance segundo a métrica acurácia (86%), como mostra a tabela a seguir:

Tabela 1: Cenários - Performance

	Texto Integral	Texto - Pos Tag
Cenário 1	0.76	0.79
Cenário 2	0.65	0.66
Cenário 3	0.75	0.79
Cenário 4	0.75	0.75
Cenário 5	0.83	0.86
Cenário 6	0.82	0.83
Cenário 7	0.81	0.81

Pode-se verificar que nos cenários 2, 4, 6 e 7 não há muita diferença entre usar ou não a abordagem pos tag. O ramo do direito se mostrou a variável de pior performance sozinha. No geral, as acurácias em quase todos os cenários se mostram boas, o que torna o algoritmo *CatBoost* um bom candidato para classificação textual neste contexto jurídico específico. Da análise descritiva surgiu a ideia de avaliar um recorte mais agressivo, mantendo apenas processos da classe ADI e ramo Direito Administrativo e outras Matérias de Direito Público. Diminuindo o universo de análise e, consequentemente, a base de dados, a abordagem pos tag se mostrou ainda mais forte, uma vez que diminui drasticamente o tamanho dos textos sem necessariamente os tornar menos densos. A acurácia, neste caso, mantendo os textos integrais é de 68% e com a abordagem pos tag ultrapassa 85%. Trata-se da maior diferença observada neste estudo. As demais métricas populares de *Machine Learning* foram avaliadas e se mostraram boas (acima de 75% em quase todos os casos).

Os algoritmos foram ajustados em *Python,* num ambiente de desenvolvimento dedicado. O consumo básico de RAM se manteve em próximo de 15% do total disponível (32gb) e apenas para o processamento dos textos brutos foram utilizadas abordagens de processamento paralelo. O pacote *Spacy*, também construído em linguagem *Python,* é amplamente utilizado para pos tag, contando com corpus etiquetados em quase todos os idiomas mais falados com alfabeto latino, a exemplo da língua inglesa, espanhola e francesa, além do português.

Conclusões

É possível observar que o algoritmo *CatBoost* se comportou bem neste contexto jurídico específico, de classificação de peças em ODS da Agenda 2030 da ONU. As métricas de acurácia se mostraram boas em quase todos os cenários e isso indica que, com algum esforço de balanceamento dos dados, outros ODS além do Objetivo 16 podem ser utilizados no futuro em estudos parecidos. A ideia de reunir features em texto e variáveis categóricas é bastante forte, uma vez que os textos jurídicos são naturalmente densos e possuem linguagem extremamente formal, o que aumenta a complexidade da limpeza. A abordagem de limpeza usando pos tag (part of speech) também se mostrou forte e trabalhos futuros podem considerar utilizá-la em textos de decisões judiciais e em tribunais com maiores fluxos processuais, de primeira e segunda instâncias. Também é possível notar que as palavras-chave criadas pelos especialistas do tribunal geraram melhoria nos modelos avaliados.

Agradecimentos

Agradecimentos especiais são dedicados aos servidores do Supremo Tribunal Federal (STF) lotados na Secretaria Geral da Presidência, Secretaria de Gestão Estratégica (SGE), Secretaria de Gestão de Precedentes, Núcleo de Gerenciamento de Precedentes (NUGEP) e Secretaria de Tecnologia da Informação (STI).

---

[1]Documentação do Algoritmo CatBoost. Disponível em: https://catboost.ai/en/docs/. Acesso em: jul, 03. 2022.
[2]Agenda 2030 no Supremo Tribunal Federal. Disponível em: https://bit.ly/STFAgendaONU2030. Acesso em: set, 02. 2022.