

# Processamento de linguagem natural aplicado à classificação de decretos administrativos brasileiros

André Luís Ribeiro, Othávio Ruddá da Cunha  
Araújo, Leonardo B. Oliveira  
andre.ribeiro, othaviorudda, leob@dcc.ufmg.br  
Departamento de Ciência da Computação - UFMG  
Belo Horizonte, MG

Magna Maria Inácio  
magna.inacio@gmail.com  
Departamento de Ciência Política - UFMG  
Belo Horizonte, MG

## ABSTRACT

Decretos administrativos são emitidos pelo Poder Executivo com o objetivo de controlar despesas, agências, ministérios e planos do governo. Apesar de serem textos de importância estratégica, no Brasil não existem classificações agregadas que explicitam as áreas de influência de cada um desses documentos. Sendo assim, neste trabalho apresentamos duas metodologias de classificação e sua automatização através de algoritmos clássicos e do estado da arte em aprendizado de máquina para modelagem textual. Nossos resultados explicitam desempenho de cerca de 80% de F1-score para alguns desses modelos, além de apresentarem uma visão geral desse tipo de dado e suas possibilidades e culminarem com a ampliação da transparência das ações e decisões do Poder Executivo Federal.

## KEYWORDS

Machine learning, neural networks, automatic document classification, political science, administrative decrees.

## 1 INTRODUÇÃO

O Poder Executivo é agente determinante no processo de decisão governamental e na produção de políticas públicas. Embora atuando de forma independente com o Parlamento e o Judiciário, o Executivo tem fortalecido sua posição estratégica por meio da expansão dos seus poderes administrativos. São esses poderes que definem, concretamente, a execução de políticas públicas, a regulação de setores econômicos e a alocação de recursos orçamentários. Esses poderes administrativos permitem ao presidente agir unilateralmente, ainda que suas decisões possam ser revogadas pelos demais Poderes. É crescente o interesse no estudo dessas decisões administrativas e seus impactos na dinâmica institucional e política.

A fim de avaliar este fenômeno nas mãos de um chefe do executivo, estudos sobre poderes unilaterais são amplamente conduzidos em países como os EUA, que buscam entender a importância do presidente como chave nas diretrizes do país [4, 7]. No Brasil, decretos administrativos são os documentos que definem os rumos do plano de governo referentes às burocracias federativas, a manutenção da máquina estatal e o financiamento de políticas públicas.

Essa configuração política brasileira permite ampla discricionariedade ao chefe do Poder Executivo, capacitando-o de defender seus interesses políticos através da distribuição de cargos, competências e recursos financeiros entre agências e ministérios. Dessa

In: XVII Workshop de Trabalhos de Iniciação Científica (WTIC 2020), São Luís, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2020.  
© 2020 SBC – Sociedade Brasileira de Computação.  
ISSN 2596-1683

forma, tem-se uma moeda de troca exclusiva do presidente para angariar recursos e manter seu plano de governo. Este fato pode ser observado através de análises políticas, no entanto, a quantidade e complexidade desse conjunto de documentos se apresenta como um problema para seu uso por cientistas políticos, legisladores, especialistas, grupos afetados e interessados. Ainda assim, não encontramos trabalhos que utilizam do estado da arte computacional para expor e tornar essas ações unilaterais mais acessíveis no Brasil.

O sistema classificatório brasileiro indexa a organização temática por ramos de direito, porém a área política, marco regulatório e a natureza da decisão do presidente, que são essenciais para análise e mapeamento do impacto da decisão, não são apresentados. Tais informações podem ser inferidas pela referenda ou por outras classificações existentes, como a disponibilizada pelo Senado Federal, entretanto os tópicos apresentados por essas sistematizações são amplos e desagregados. Além disso, no contexto governamental classificações humanas podem ser suscetíveis a viés político e, portanto, o ideal seria que esse processo fosse automatizado. Nesse contexto, o objetivo principal deste trabalho é desenvolver modelos computacionais para classificação de decretos administrativos brasileiros, através das seguintes contribuições específicas:

- Coleta e organização de decretos administrativos do período de 2000 a 2019;
- Desenvolvimento de classificações agregadas referentes à temática e às áreas de interesse dos decretos;
- Aplicação e comparação de abordagens de aprendizado de máquina para modelagem das classificações criadas, automatizando o processo de categorização de novos textos.

Este trabalho se encontra dividido da seguinte forma. A Seção 2 apresenta a fundamentação teórica dos algoritmos e técnicas utilizadas na automatização classificativa. A Seção 3 apresenta trabalhos relacionados. A Seção 4 apresenta a metodologia proposta para modelagem textual. A Seção 5, por sua vez, sistematiza os resultados e descobertas feitas no decorrer dos experimentos. Finalmente, a Seção 6 pontua nossas conclusões e apresenta sugestões de trabalhos futuros a serem realizados com os dados em questão.

## 2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção apresentamos os fundamentos em processamento de linguagem natural e em algoritmos de classificação textual.

### 2.1 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) é um conjunto de técnicas da Ciência da Computação que possibilitam a interação

humano computador através da modelagem de textos em uma sequência lógica computável. Com abordagens do estado da arte em aprendizado de máquina e aprendizado profundo, PLN é utilizado para modelagem, análise e classificação no campo linguístico.

## 2.2 Classificação Supervisionada de Textos

Abordagens de classificação supervisionadas se pautam na representação de estruturas textuais em formas numéricas e na modelagem da distribuição e comportamento dos dados a partir de um conjunto de treino categorizado. Após obtidas as representações numéricas, classificadores como *Naive Bayes*, árvores de decisão, SVMs podem ser utilizados para modelagem [3]. Em particular, a representação de palavras via Tf-Idf e a classificação através de SVMs representaram o estado da arte em alguns problemas há alguns anos. Entretanto, atualmente abordagens supervisionadas baseadas em redes neurais chamam atenção pelo desempenho superior em diversos contextos.

## 2.3 Redes Neurais para Classificação de Documentos

Redes neurais são modelos inspirados pelo funcionamento cerebral e compostos por agrupamentos de neurônios artificiais. Em uma rede neural, neurônios são organizados em diferentes camadas, responsáveis por definir o comportamento e o tipo de operação efetuada em cada nível da rede, além de quanta e em que forma a informação será propagada para os próximos níveis e neurônios conectados. Em particular à modelagem textual, tipos de redes neurais comuns incluem redes convolucionais (CNNs) e redes recorrentes (RNNs).

## 3 TRABALHOS CORRELATOS

O uso de PLN para classificação de texto é estudado em diferentes contextos. Especificamente para textos governamentais, os autores de [8] e [1] compararam os resultados de algumas arquiteturas de redes neurais para classificação textual. A base utilizada contém documentos da União Europeia classificados tanto em tópicos mais agregados quanto em tópicos de maior especificidade.

Não encontramos trabalhos que envolvem classificação de decretos administrativos através de redes neurais. Entretanto, os autores de [7] abordam modelagem de tópicos para ordens executivas através de algoritmos não supervisionados tradicionais e da interpretação do significado semântico e da coerência dos tópicos obtidos a partir de aspectos da ciência política. Eles também apresentam as vantagens desse método científico e defendem que, para uma amostragem extensa de textos, o uso de máquinas é mais eficiente. Nessa mesma linha, [4] destaca as promessas do uso de abordagens computacionais para textos do Poder Executivo.

No contexto brasileiro, [5] apresenta a base de dados do Supremo Tribunal Federal (STF) e aborda o problema de classificação dos documentos provenientes desta sob duas perspectivas: classificação por tipo de documento e multi rotulagem de temas. Preocupados com o alto gasto de recursos para análise manual desses documentos, os autores utilizam de métodos computacionais, abordando estratégias baseadas em *bag-of-words* e Tf-Idf associadas a classificadores como SVM, além de métodos baseados em redes neurais. Em [2], a mesma base de dados é utilizada, entretanto com foco na aplicação específica de uma CNN para esse problema.

## 4 METODOLOGIA

A Figura 1 apresenta o fluxograma de obtenção e tratamento dos dados utilizados neste trabalho. Exploramos classes criadas a partir do agrupamento de referendas e classes criadas a partir do agrupamento dos subtópicos provenientes do Senado. Essa abordagem de agrupamento permite considerar mudanças organizacionais e nominais ao longo dos anos e criar classes com maior amostragem, entretanto, resulta também em um aumento da generalidade dos tópicos considerados. Os subtópicos utilizados pelo Senado para descrever os decretos dos 20 anos em questão foram agrupados em 15 categorias. As referendas, por sua vez, foram agrupadas em 13. Ambas classificações procuram levar em conta a consistência dos textos envolvidos e a relevância das áreas abordadas.

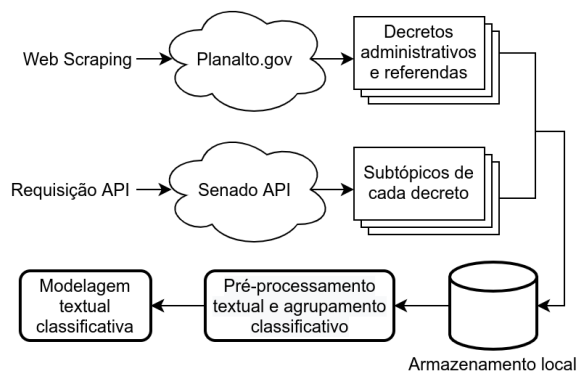


Figura 1: Fluxograma metodológico

As duas classificações apresentam tópicos com interesses semelhantes, entretanto apresentam lógicas conceituais distintas. As classes obtidas a partir do agrupamento de subtópicos provenientes do Senado representam uma visão diretamente atrelada à área de interferência de cada decreto. Por outro lado, os agrupamentos de referendas estão associadas às responsabilidades atribuídas a cada órgão do governo e à autoria de cada documento. Ainda assim, ambas são influenciadas pelo conteúdo textual dos decretos.

Destacamos ainda que decretos podem envolver múltiplas jurisdições ministeriais e requerer a co-autoria de diferentes membros do Executivo, sendo indicados pelo Senado como pertencentes a um ou mais subtópicos simultaneamente e referendados por mais de um órgão. Sendo assim, ambos problemas classificativos são tratados como problemas *multi-class* e *multi-label*, dado que existem  $n$  classes e um decreto pode se enquadrar em uma a  $n$  classes.

Neste trabalho comparamos os resultados de três estratégias supervisionadas para classificação textual previamente apresentadas na literatura. Para isso, o conjunto total de decretos obtido foi dividido em três *datasets*: treino (70% dos dados), teste (15% dos dados) e validação (15% dos dados). As Figuras 2 e 3 apresentam as classes obtidas e a distribuição dos dados nos subconjuntos mencionados, tanto do ponto de vista da classificação baseada nos dados do Senado quanto na baseada em referenda. Apesar do claro desbalanceamento entre as classes, os três *datasets* contemplam a distribuição apresentada na realidade e incluem amostras proporcionais de cada classe representada.

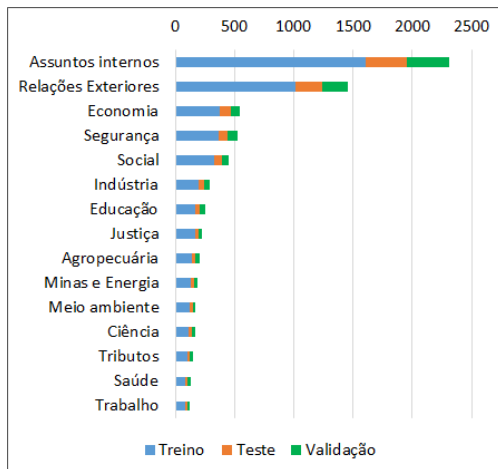


Figura 2: Distribuição classificativa - Senado.

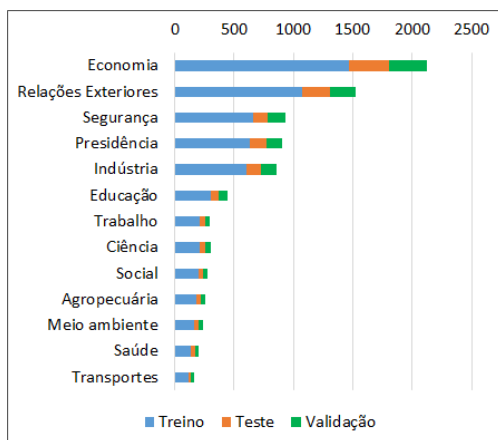


Figura 3: Distribuição classificativa - Referenda.

#### 4.1 Abordagem de classificação via SVM

A primeira abordagem é baseada na utilização de Tf-Idf e SVMs. Dado o conjunto de decretos, associamos cada palavra à sua raiz linguística utilizando lematização. Além disso, removemos *stopwords* da língua portuguesa, i.e. palavras de uso comum na língua que, para um classificador baseado em *bag of words*, não contribuem com a constatação de um tópico. Para conversão dos textos em representações numéricas, utilizamos o método de Tf-Idf, inicializado somente com palavras do *dataset* de treino. A seguir, treinamos o algoritmo SVM de kernel linear aplicando o método de relevância binária para obter classificações *multi-class* e *multi-label*. Nesse processo, otimizamos ainda o hiperparâmetro de regularização do SVM (C) através de 35 iterações de otimização bayesiana [6].

#### 4.2 Abordagens baseadas em redes neurais

Os métodos baseados em redes neurais, por sua vez, foram uma rede convolucional e uma rede recorrente com mecanismo de atenção hierárquico. Nessas abordagens, as palavras são representadas através de *embeddings*. Otimizamos ainda os principais hiperparâmetros

dessas redes através de 100 iterações de otimização bayesiana [6]. Após obtidos, utilizamos os melhores hiperparâmetros para treinar o modelo final e testá-lo. O tamanho da camada de saída das redes varia de acordo com a classificação em questão, porém todas as camadas de saída possuem função de ativação sigmoid. Cada rede foi compilada com função de perda *binary cross-entropy*, de forma a garantir probabilidades independentes para cada uma das classes por se tratar de um problema *multi-label*.

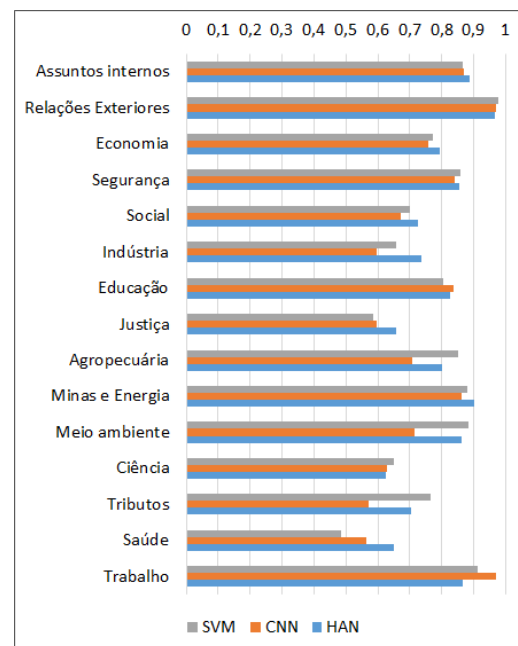
**4.2.1 Rede convolucional.** A rede convolucional utilizada parte do trabalho de [2], que demonstrou sucesso na classificação de documentos jurídicos em língua portuguesa com essa arquitetura.

**4.2.2 Rede com mecanismo de atenção hierárquico.** A rede recorrente com mecanismo de atenção hierárquico (HAN) utilizada tem arquitetura apresentada em [9]. Para classificação textual, o mecanismo de atenção adotado permite considerar diretamente a importância de cada palavra e sentença para uma dada classificação, ao contrário do que acontece com a rede convolucional utilizada e com outros mecanismos existentes na literatura.

## 5 RESULTADOS

Nesta seção, apresentamos os resultados obtidos com a condução deste projeto, evidenciando o tipo de modelagem efetuada e a performance dos classificadores testados. Para a avaliação dos algoritmos, utilizamos a métrica F1-score individualmente em cada classe.

A Figura 4 indica os resultados obtidos por classe obtidos por cada modelo para a classificação baseada nos subtópicos do Senado. Nesse caso, constata-se maior estabilidade no desempenho entre classes para os modelos HAN e SVM, dado que a abordagem CNN tem desempenho consideravelmente inferior à média das demais nos tópicos meio ambiente e tributos.

Figura 4: F1 por classe (*dataset* de teste) - Senado.

A Figura 5 ilustra o desempenho obtido para automatização da classificação pautada em referendas. Por sua vez, ambos os algoritmos SVM e HAN demonstram relativa estabilidade entre classes, i.e., não ficam muito abaixo da média para nenhum tópico, ao passo que novamente a CNN demonstra resultados notoriamente inferiores para um tópico, no caso, social.

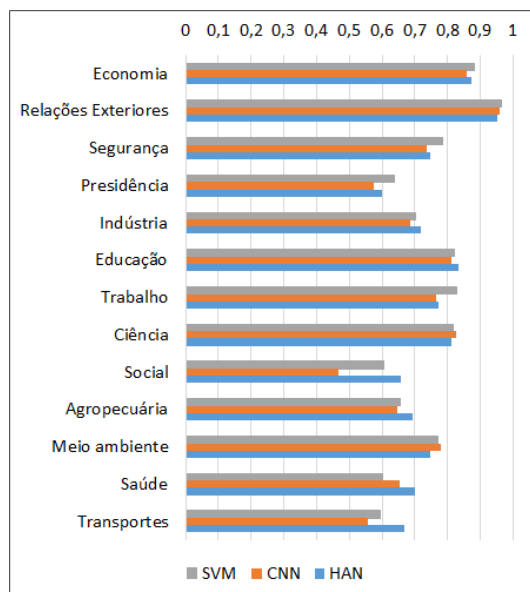


Figura 5: F1 por classe (dataset de teste) - Referenda.

A Tabela 1 sumariza os resultados dos classificadores através da média macro do F1-score dos tópicos, aplicável a problemas *multi-label* com desbalanceamento entre classes. Em ambas classificações, constata-se desempenho superior da HAN em nosso conjunto de testes, seguida pelo SVM e pela CNN. Destaca-se ainda que os resultados de todos os modelos é superior na classificação baseada nos dados do Senado. Infere-se que tal observação se deva a tópicos que não são consistentemente referendados pelas mesmas estruturas organizacionais devido a potenciais conflitos de interesse, o que indica uma dificuldade extra para os modelos estudados.

Tabela 1: Resultados (dataset de teste) - Macro F1.

Classificação/Modelo	SVM	CNN	HAN
Senado	77,7%	74,4%	79,1%
Referenda	74,3%	71,8%	75,2%

Dado que este é um problema *multi-label*, de uma perspectiva prática é ainda interessante observar com que frequência os classificadores avaliados não erram nenhuma das classes esperadas para uma dada amostra. A Tabela 2 apresenta os resultados obtidos sob essa ótica. Destaca-se que esse é um ponto de vista altamente influenciado pelo desbalanceamento entre classes e pela quantidade de textos que de fato apresentam rótulos *multi-label*, portanto, esse aspecto não deve ser considerado individualmente. Em geral, esse resultado acompanha os demais apresentados, entretanto, dessa

perspectiva o algoritmo SVM para classificação baseada em referenda supera os resultados obtidos pelo modelo HAN.

Tabela 2: Frequência com que os modelos acertam todas as classes de uma amostra de teste.

Classificação/Modelo	SVM	CNN	HAN
Senado	75,2%	71,4%	77,5%
Referenda	68,8%	64,0%	66,4%

## 6 CONCLUSÃO

Este trabalho apresentou o desenvolvimento de duas classificações agregadas para decretos administrativos do Executivo Federal no Brasil e algumas abordagens de modelagem computacional para automatização classificativa. Do aspecto da classificação automática, em geral constata-se desempenho superior do modelo HAN para ambas classificações, seguido pelo SVM e pela CNN. Até onde sabemos, este é o primeiro trabalho a aplicar abordagens desse tipo a decretos administrativos.

A metodologia e os resultados apresentados podem ser utilizados para facilitar o acesso a decretos administrativos pela população e para acelerar a categorização e consequente divulgação de novos textos, culminando com a ampliação da transparência do Poder Executivo Federal. Por fim, destacamos ainda as amplas possibilidades de estudos que essa base oferece. Como trabalho futuro, sugerimos a condução de novos estudos do conteúdo textual, áreas de interesse e do relacionamento entre esses textos de importância estratégica para o país.

## ACKNOWLEDGMENTS

Agradecemos à PRPq/UFMG, à FAPEMIG e ao CNPq pelo apoio.

## REFERENCES

- [1] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192* (2019).
- [2] N Correia Da Silva, FA Braz, TE de Campos, D Gusmao, F Chaves, D Mendes, D Bezerra, G Ziegler, L Horinouchi, M Ferreira, et al. 2018. Document type classification for Brazil's supreme court using a convolutional neural network. In *10th International Conference on Forensic Computer Science and Cyber Law (ICoFCS)*, Sao Paulo, Brazil. 29–30.
- [3] Mita K Dalal and Mukesh A Zaveri. 2011. Automatic text classification: a technical review. *International Journal of Computer Applications* 28, 2 (2011), 37–40.
- [4] Aaron R Kaufman. 2020. Measuring the Content of Presidential Policy Making: Applying Text Analysis to Executive Branch Directives. *Presidential Studies Quarterly* 50, 1 (2020), 90–106.
- [5] Pedro Henrique Luz de Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. 2020. VICTOR: a Dataset for Brazilian Legal Documents Classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1449–1458. <https://www.aclweb.org/anthology/2020.lrec-1.181>
- [6] Luigi Nardi, David Koepflinger, and Kunle Olukotun. 2019. Practical design space exploration. In *2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*. IEEE, 347–358.
- [7] JB Ruhl, John Nay, and Jonathan Gilligan. 2018. Topic Modeling the President. (2018).
- [8] Samir Undavia, Adam Meyers, and John E Ortega. 2018. A comparative study of classifying legal documents with neural networks. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 515–522.
- [9] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.