Classificação Multi-rótulo de Normas Jurídicas Estaduais

Duílio Campos Sasdelli

Introdução

- Normas jurídicas muitas vezes versam sobre mais de um assunto;
- Algumas bases de dados jurídicas possuem centenas ou milhares de assuntos;
- Nem sempre é trivial realizar a classificação manual sem erros;
- A principal **motivação** é **auxiliar o trabalho de Assessores e Consultores Legislativos;**

LEI 23631, DE 02/04/2020

Dispõe sobre a adoção de medidas para o enfrentamento do estado de calamidade pública decorrente da **pandemia de Covid-19**, causada por coronavírus.

Assuntos:

Agropecuária.
Calamidade Pública.
Indústria, Comércio e Serviços.
Saúde Pública.
Defesa do Consumidor.
Transporte Coletivo.
Finanças Públicas.
Assistência Social.
Municípios e Desenvolvimento Regional.
Direitos Humanos.
Povos e Comunidades Tradicionais.
Administração Estadual.
Pessoal.
Cultura.
Idoso.
Mulher.
Negro.

https://www.almg.gov.br/consulte/legislacao/completa/completa.html?tipo=LEI&num=23631&comp=&ano=2020

Conteúdo de uma lei

LEI 23631, DE 02/04/2020
(...)

Art. 3º – Para o enfrentamento da pandemia de Covid-19, poderão ser adotadas pela autoridade competente as seguintes medidas, entre outras:

I – isolamento;
II – quarentena;
III – determinação de realização compulsória dos seguintes procedimentos:

a) exames médicos;
b) testes laboratoriais;
c) coleta de amostras clínicas;
d) vacinação e outras medidas profiláticas;
e) tratamentos médicos específicos;
IV – estudo ou investigação epidemiológica;
V – exumação, necropsia, cremação e manejo de cadáver;
VI – requisição de bens e serviços de pessoas naturais e jurídicas, hipótese em que será garantido o pagamento posterior de indenização justa, em dinheiro;
VII – autorização excepcional e temporária para importação de produtos sujeitos à vigilância sanitária sem registro na Agência Nacional de Vigilância Sanitária – Anvisa –, desde que registrados por autoridade sanitária estrangeira e previstos em ato do Ministério da Saúde;

(...)

Referencial Teórico

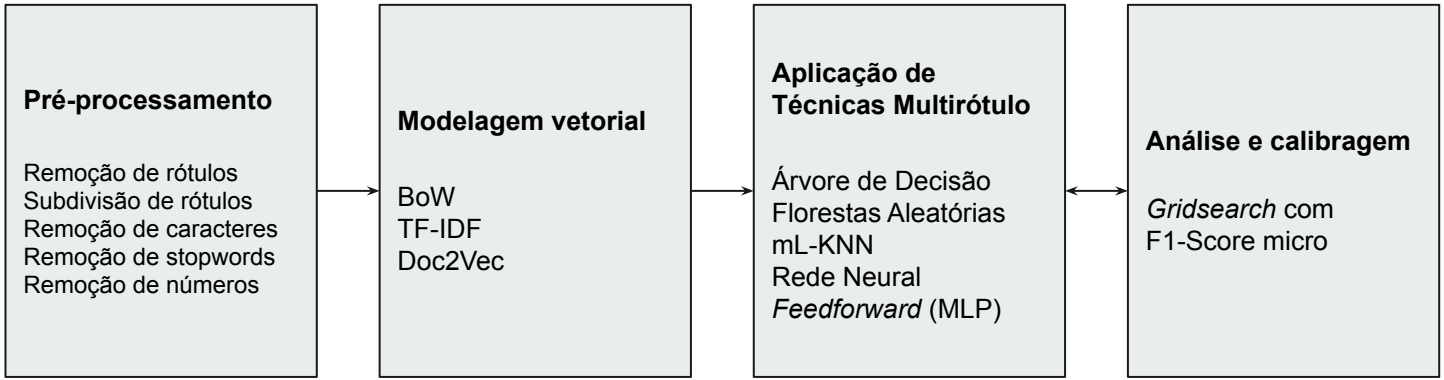
- Não há muitos trabalhos sobre esse tema no Brasil e na legislação brasileira;
- A maioria dos trabalhos (como monografias e TCCs) concentra-se em classificação de jurisprudência;
- O trabalho “Extreme Multi-Label Legal Text Classification: A Case Study in EU Legislation”:
    - Aborda diferentes tipos de redes neurais complexas;
    - Utiliza modelos de atenção BIGRU (*Bidirectional Gated Recurrent Unit*) e redes convolucionais (CNN);
    - Apresentou bons resultados inclusive em casos de aprendizado ***zero-shot***, ***one-shot***, **XMTC (Extreme multi-label text classification);**

Dados

- Será utilizada a base de dados de Normas Estaduais da ALMG;
- 28308 normas (24560 com assuntos);
- 646 assuntos distintos (217 assuntos ocorrem mais de 5 vezes)
- Apenas 4109 normas possuem mais de 1 assunto;

Assunto	Quantidade
utilidade pública	12193
próprio público	2183
imóvel	2038
estabelecimento de ensino	1899
crédito	1485
tributos	558
secretaria de estado de educação (see), pessoal	498
executivo, pessoal	368
divisão administrativa	346
saúde pública	324

Arquitetura



Pré-processamento

- Pré-processamento da base de dados:
  - Remoção de assuntos frequentes: **“Utilidade Público”** e **“Próprio Público”**;
  - Particionamento de assuntos: **“secretaria de estado de educação (see), pessoal”** se torna **“secretaria de estado de educação (see)”** e **“pessoal”**;
  - Correção de assuntos com nomes despadronizados: **“tributos”** e **“tributo”** se torna **“tributos”**;
- Pré-processamento do texto:
  - Transformação para caixa baixa (*lower*);
  - Remoção de *stopwords*, números, pontuações, numerais romanos de acentos;
  - Foi utilizada a lista de *stopwords* da biblioteca NLTK;
  - Não foi realizada *lemmatizing* ou *stemming*;

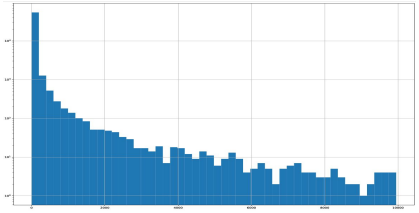
Resultado do pré-processamento

- 9963 normas com 80 assuntos:
  - 3379 normas possuem 2 ou mais assuntos;
  - 1300 normas possuem 3 ou mais assuntos;
  - 262 normas possuem 4 ou mais assuntos;
  - 1 norma possui 14 assuntos;
  - 36 assuntos em 100 ou mais normas;
- 57293 palavras ao todo;
  - 43 palavras aparecem mais de 10.000 vezes;
  - 41433 aparecem menos de 10 vezes;
  - 19919 aparecem apenas 1 vez;

Assunto	Qt
imóvel	2038
pessoal	1731
crédito	1486
executivo	549
secretaria de educação	543

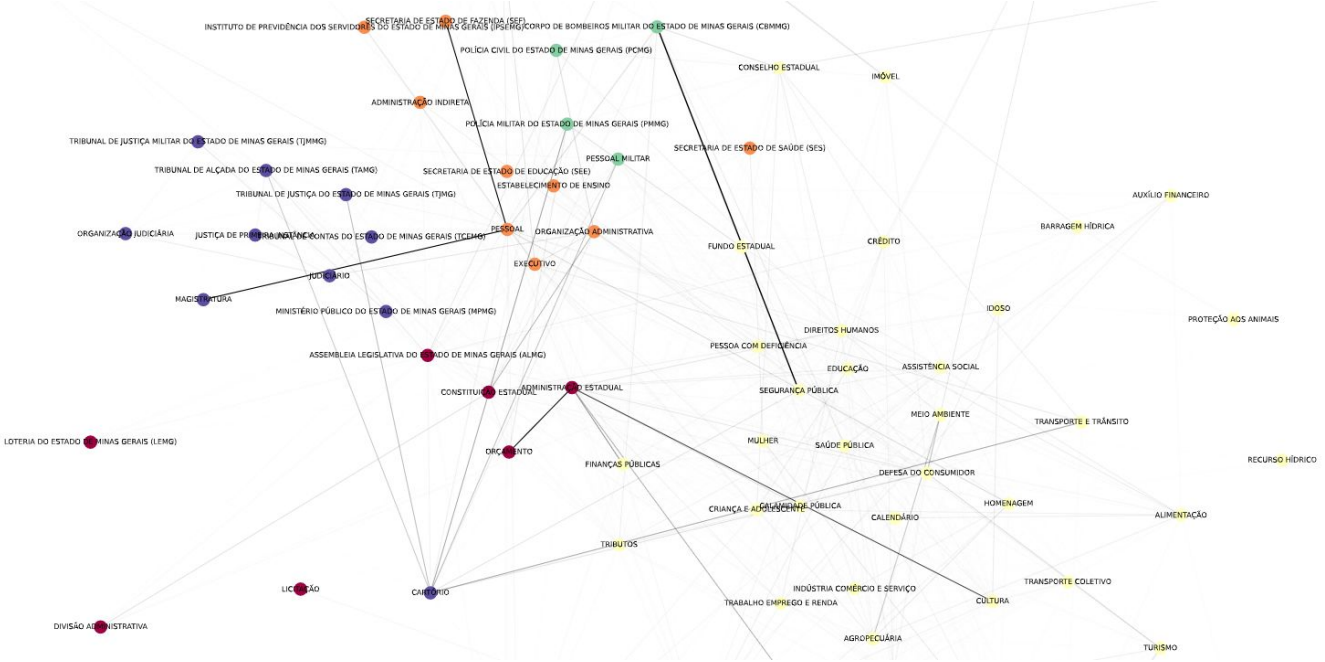
Assunto	Qt
pessoal, secretaria de educação	522
estabelecimento de ensino, pessoal	412
estabelecimento de ensino, secretaria de educação	408
executivo, pessoal	394
organização administrativa, pessoal	277

Histograma de palavras com c < 10000



Palavra	Qt
cr	97941
lei	70448
estado	59146
minas	34151
gerais	29654

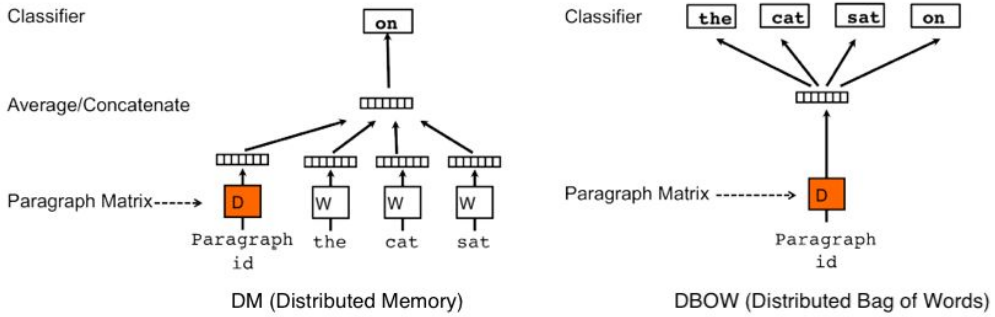
Resultado do pré-processamento



Modelagem Vetorial

- Foram utilizadas 3 abordagens;
- **Bag of words:** contagem simples das palavras da base de dados;
- **TF-IDF:** leva em conta a importância relativa da palavra para determinado texto;
- **Doc2Vec:** gera um vetor para um texto baseado nas palavras nele presentes utilizando modelos de redes neurais (similar ao word2vec)

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{\mathrm{N}}{\mathrm{df}_i}\right)$$

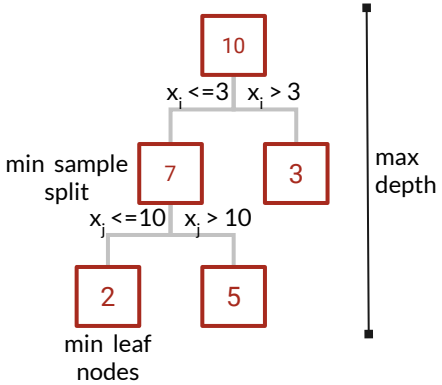


Árvore de decisão

- Modelo baseado em particionamentos sucessivos nas amostras;
- Pode ser utilizado em classificação multirótulo diretamente;
- Duas formas de particionamento:
  - Gini;
  - Entropia;

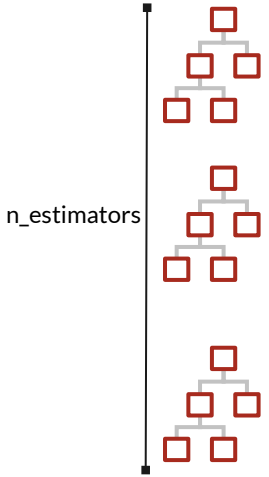
$$I_G = 1 - \sum_{j=1}^{c} p_j^2 \qquad I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

- Parâmetros utilizados:
  - min_samples_split: mínimo de amostras para se dividir um nós;
  - max_depth: máxima profundidade;
- Implementação da biblioteca *scikit-learn*;



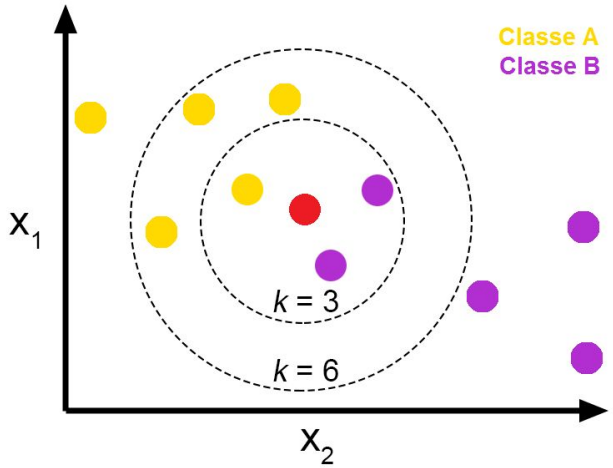
Florestas aleatórias

- Utiliza-se de diferentes árvores de decisão (estimadores);
- Cada árvore tem acesso apenas a um subconjunto aleatório de amostras (*bootstraping*) e de características (*feature boostraping*);
- Ao final é utilizado uma combinação de resultados, por exemplo por votação da maioria;
- Utiliza alguns dos mesmos parâmetros das árvores de decisão;
    - Parâmetro número de estimadores define o número de árvores de decisão;
- Utilizada implementação a biblioteca *scikit-learn*;



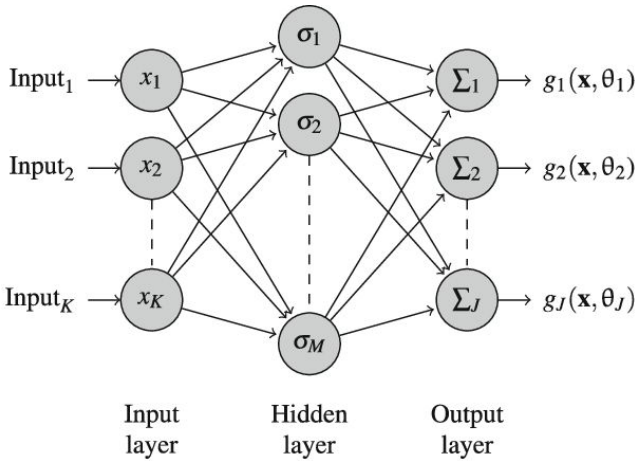
mL-kNN

- Extensão do algoritmo kNN para lidar com classificação multi-rótulo;
- Utiliza o princípio do **máximo a posteriori;**
- Parâmetros utilizados:
  - **k:** número de vizinhos a serem considerados;
  - **s:** fator de suavização
- Apresentou bons resultados em problemas de classificação de textos;
- Utilizada implementação da biblioteca *scikit-multilearn*;



Perceptron Multicamada (MLP)

- Rede Neural *Feedforward* totalmente conectada;
- Possui uma ou mais camadas ocultas;
- Funções de ativação:
  - Camadas ocultas: ReLU ou tanh;
  - Camada de saída: logística (para multirótulo);
- Resolvedores (*Solvers*) para atualização de pesos por *backpropagation*:
  - SGD (Gradiente Descendente Estocástico) ou Adam (Adaptive Moment Estimation);
- Pode ser aplicado em diferentes tipos de problemas, tais con de visão computacional ou processamento de texto;
- Implementação na biblioteca *multiscikit-learn*;



Experimentos realizados

- Foi realizada uma busca em grade de parâmetros com validação cruzada (*GridSearchCV*);
- Foram utilizados 3 tipos de modelagem vetorial para 4 diferentes estimadores:
  - Bow, TF-IDF e Doc2Vec;
  - Árvore de decisão, Florestas Aleatórias, mL-kNN, MLP;
- Foi realizada validação cruzada com duas partições estratificadas;

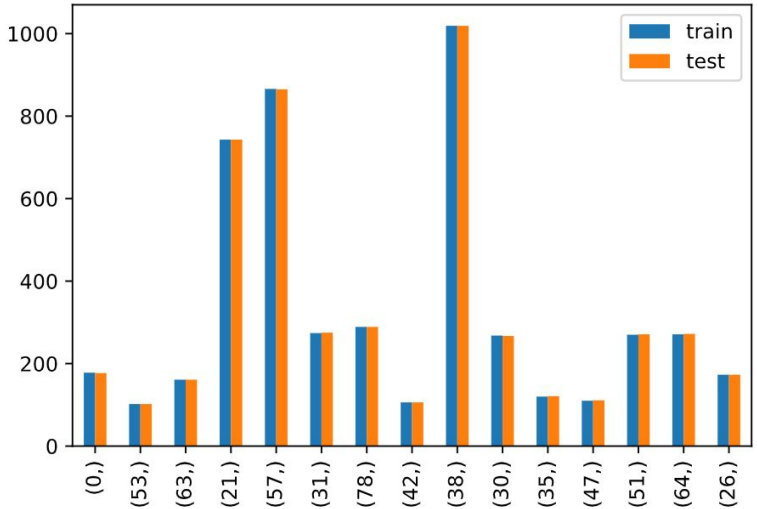
Table 1: Grade de parâmetros utilizada

Árvores de Decisão	Florestas Aleatórias	Ml-kNN	MLP
min samples split: $[0.005, 0.010, 2]$ **max depth:** $[16, 32, None]$	**min samples split:** $[0.005, 0.010, 2]$ **n estimators:** $[100, 110, 120]$ **min samples leaf:** $[5, 3, 1]$	**k:** $[6, 8, 10, 12]$ **s:** $[0.5, 1.0, 1.5, 2.0]$	**layer sizes:** $[(150), (100, 100), (50, 50, 50)]$ **activation:** $[tanh, relu]$ **solver:** $[sgd, adam]$
9 combinações	27 combinações	16 combinações	12 combinações
27 com modelos	81 com modelos	48 com modelos	36 com modelos

Total de 384 modelos com validação cruzada de duas partições

Experimentos realizados - Particionamento estratificado

- A biblioteca *scikit-multilearn* possui uma função de particionamento estratificado iterativo, útil para criação de divisões treino/testes representativas;
- Respeita combinações de rótulos, na medida do possível;



Experimentos realizados - Métrica de avaliação

- A métrica **f1-micro** consiste na média da **f1-score** para todas as classificações realizadas pelo algoritmo;
- Optou-se por utilizá-la porque pretende-se avaliar as classificações corretas independente da distribuição das classes;
- Os classificadores tenderão a se sobressair para as classes mais representadas;

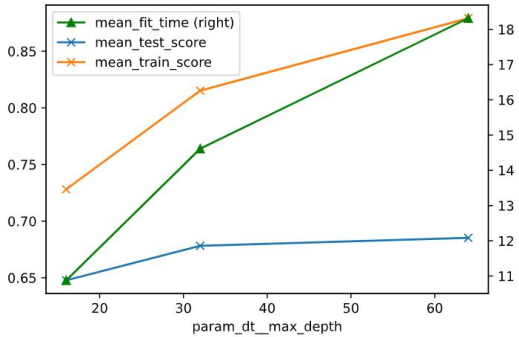
$$PrecisaoMicro = \frac{\sum_i^n TP_{C_i}}{\sum_i^n TP_{C_i} + \sum_i^n FP_{C_i}}, RecallMicro = \frac{\sum_i^n TP_{C_i}}{\sum_i^n TP_{C_i} + \sum_i^n FN_{C_i}}$$

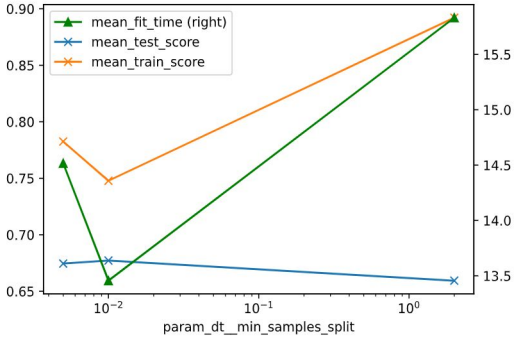
$$F1Micro = 2 \times \frac{PrecisaoMicro + RecallMicro}{PrecisoMicro \times RecallMicro}$$

Resultados - Árvore de decisão

- D2V não apresentou bons resultados;
- Houve overfitting para CV e TF-IDF em determinadas combinações de parâmetros;
- O tempo de treinamento é muito maior para D2V;
- TF-IDF apresentou o melhor resultado (gráficos abaixo):

origin	mean fit time	mean train score mean	mean train score max	mean test score	mean test score max
cv	6.6695	0.8001	1.0000	0.6623	0.6860
tfidf	14.6013	0.8075	1.0000	0.6704	0.6985
d2v	26.9666	0.6640	0.7329	0.5632	0.5785

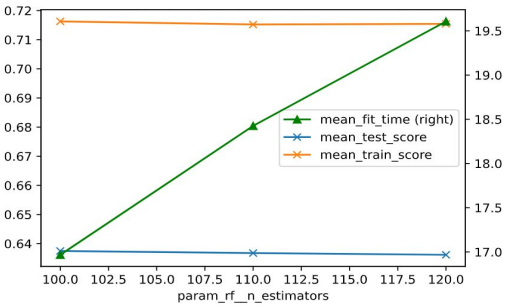


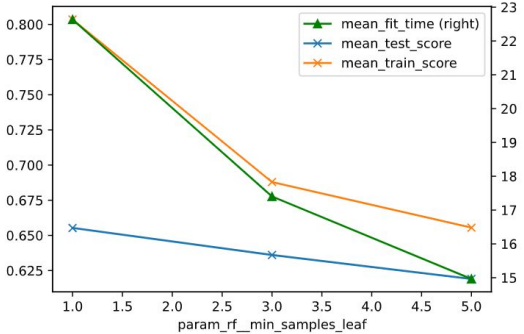


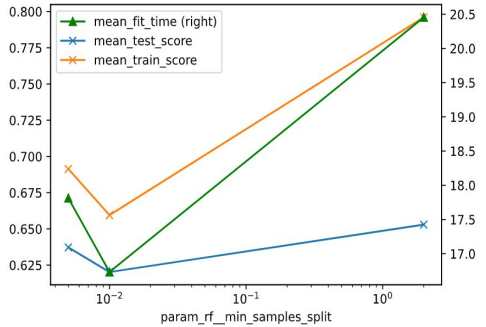
Resultados - Florestas Aleatórias

- D2V novamente não apresentou bons resultados;
- Houve overfitting para CV e TF-IDF em determinadas combinações de parâmetros;
- O tempo de treinamento é muito maior para D2V;
- TF-IDF apresentou o melhor resultado (gráficos abaixo):

origin	mean fit time	mean train score mean	mean train score max	mean test score	mean test score max
cv	11.6453	0.6940	0.9999	0.6241	0.6814
tfidf	18.3332	0.7156	1.0000	0.6368	0.6831
d2v	62.1731	0.6155	0.7479	0.5718	0.6231



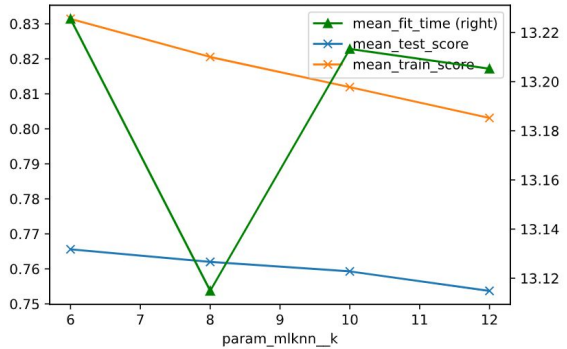


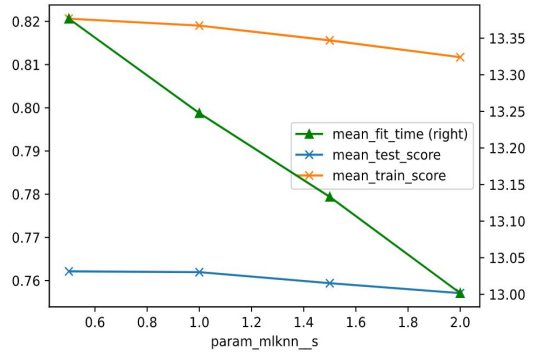


Resultados - ml-kNN

- D2V novamente não apresentou bons resultados;
- Não houve overfitting;
- O tempo de treinamento é maior para D2V;
- TF-IDF apresentou o melhor resultado (gráficos abaixo):

origin	mean fit time	mean train score mean	mean train score max	mean test score	mean test score max
cv	13.4302	0.7707	0.7950	0.7025	0.7099
tfidf	13.1897	0.8167	0.8330	0.7601	0.7665
d2v	33.9638	0.7799	0.8017	0.7272	0.7343

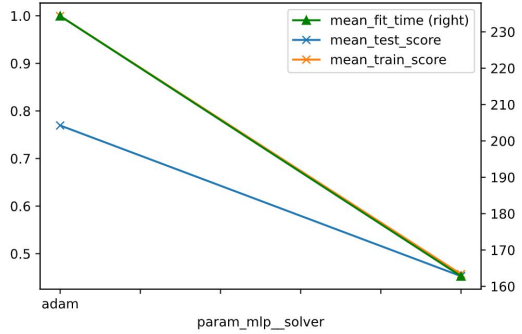


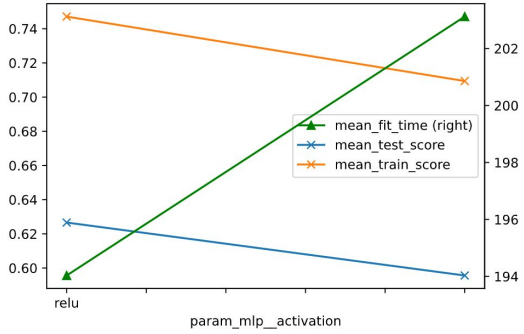


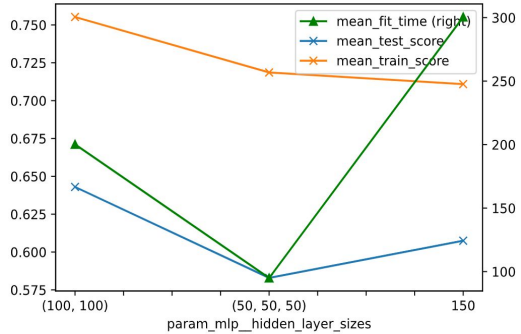
Resultados - Perceptron Multicamada (MLP)

- D2V novamente não apresentou bons resultados;
- Houve overfitting na rede para determinadas configurações de parâmetros;
- Não houve convergência para determinadas configurações de parâmetros;
- O tempo de treinamento é menor para D2V;
- TF-IDF apresentou o melhor resultado (gráficos abaixo):

origin	mean fit time	mean train score mean	mean train score max	mean test score	mean test score max
cv	519.5058	0.9253	1.0000	0.7546	0.7907
tfidf	198.5768	0.7283	1.0000	0.6111	0.7969
d2v	26.6733	0.7236	0.8905	0.6685	0.7586







Resultados - Perceptron Multicamada (MLP)

- Os melhores resultados foram identificados para a seguinte combinação de parâmetros:
  - Modelagem **TF-IDF;**
  - Camadas ocultas: **(150)**
  - Função de ativação das camadas ocultas: **ReLU**;
  - Resolvedor: **Adam**;
- Investigará-se agora um exemplo de classificação fazendo uso dessas configurações para uma instância do problema;

activation	hidden layer sizes	solver	mean fit time	mean train score	mean test score
tanh	150	sgd	266.2043	0.4163	0.4137
tanh	150	adam	356.9774	0.9998	0.7941
tanh	(100, 100)	sgd	163.7233	0.5072	0.5031
tanh	(100, 100)	adam	252.1229	0.9999	0.7837
tanh	(50, 50, 50)	sgd	76.7089	0.3356	0.3329
tanh	(50, 50, 50)	adam	102.9910	0.9975	0.7463
relu	150	sgd	238.4223	0.4277	0.4252
relu	150	adam	340.3201	0.9998	0.7969
relu	(100, 100)	sgd	157.2616	0.5139	0.5114
relu	(100, 100)	adam	227.8457	1.0000	0.7738
relu	(50, 50, 50)	sgd	74.5037	0.5414	0.5303
relu	(50, 50, 50)	adam	125.8403	1.0000	0.7221

Resultados - Perceptron Multicamada (MLP)

Table 13: Resultados de precisão, recall e F1 para MLP com modelagem TF-IDF

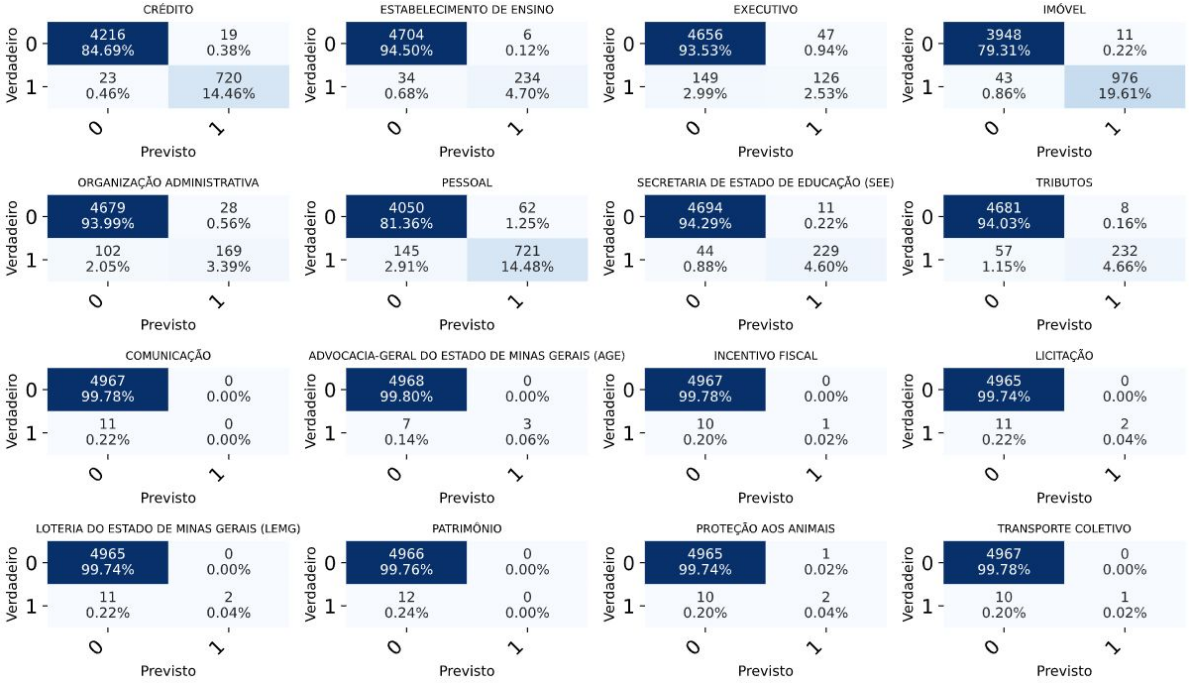
	precision	recall	f1-score	support
samples avg	0.8034	0.7541	0.7650	7552.0000
weighted avg	0.8985	0.6962	0.7652	7552.0000
macro avg	0.8082	0.4361	0.5294	7552.0000
micro avg	0.9229	0.6962	0.7937	7552.0000
IMÓVEL	0.9889	0.9578	0.9731	1019.0000
PESSOAL	0.9208	0.8326	0.8745	866.0000
CRÉDITO	0.9743	0.9690	0.9717	743.0000
TRIBUTOS	0.9667	0.8028	0.8771	289.0000
EXECUTIVO	0.7283	0.4582	0.5625	275.0000
SECRETARIA DE EDUCAÇÃO (SEE)	0.9542	0.8388	0.8928	273.0000
ORGANIZAÇÃO ADMINISTRATIVA	0.8579	0.6236	0.7222	271.0000
ESTABELECIMENTO DE ENSINO	0.9750	0.8731	0.9213	268.0000
ADMINISTRAÇÃO ESTADUAL	0.8532	0.5254	0.6503	177.0000
DIVISÃO ADMINISTRATIVA	0.9573	0.9075	0.9318	173.0000

Resultados - Perceptron Multicamada (MLP)

Table 13: Resultados de precisão, recall e F1 para MLP com modelagem TF-IDF

	precision	recall	f1-score	support
DEFENSORIA PÚBLICA DE MG (DPMG)	0.8750	0.5385	0.6667	13.0000
SECRETARIA DE SEGURANÇA (SSPMG)	0.0000	0.0000	0.0000	13.0000
LOTERIA DO ESTADO DE MG (LEMG)	1.0000	0.1538	0.2667	13.0000
LICITAÇÃO	1.0000	0.1538	0.2667	13.0000
PROTEÇÃO AOS ANIMAIS	0.6667	0.1667	0.2667	12.0000
PATRIMÔNIO	0.0000	0.0000	0.0000	12.0000
TRANSPORTE COLETIVO	1.0000	0.0909	0.1667	11.0000
COMUNICAÇÃO	0.0000	0.0000	0.0000	11.0000
INCENTIVO FISCAL	1.0000	0.0909	0.1667	11.0000
ADVOCACIA-GERAL DE MG (AGE)	1.0000	0.3000	0.4615	10.0000

Resultados - Perceptron Multicamada



Conclusão

- A utilização da métrica ***f1-micro*** valorizou as classes mais frequentes e deve ser evitada em que cada classe possui um mesmo peso;
- A modelagem com Doc2Vec não trouxe ganhos significativos, possivelmente em virtude do tamanho reduzido dos textos;
- Trabalhos futuros:
    - Reduzir dimensionalidade dos modelos vetoriais utilizando PCA por exemplo;
    - Utilizar outros modelos vetoriais, tais como LDA, Word2Vec médio, dentre outros;
    - Utilizar estimadores baseados em redes neurais mais complexos (LSTM, CNN, etc);

Referências

- Portal de Dados Abertos da Assembléia Legislativa de Minas Gerais.Disponível em:http://dadosabertos.almg.gov.br/ws/ajuda/sobre
- Classificação multi-rótulo no SKLEarn. Disponível em:https://scikit-learn.org/stable/modules/multiclass.html
- Chalkidis. I., et. al. Extreme Multi-Label Legal Text Classification: A case study in EU Legislation. Proceedings of the Natural Legal Language Processing Workshop 2019, pages 78–87. Minneapolis, Minesotta, June 7, 2019.