Domain-specific datasets for document classification and named entity recognition

Pedro Henrique Luz de Araujo

Departamento de Ciência da Computação, Universidade de Brasília, Brasília – DF, Brazil

pedro.luz@aluno.unb.br
`https://cic.unb.br/~teodecampos/peluz/`

Banca:
Teófilo E. de Campos (Orientador - UnB)
Alexandre Rademaker (IBM Research, FGV)
Thiago de Paulo Faleiros (UnB)
Luís Paulo F. Garcia (Suplente - UnB)

29 de julho de 2021

Sumário

Introdução

Motivação

- Dados textuais estão em constante produção.
  - Posts de redes sociais, livros, notícias, publicações oficiais, processos judiciais.
- Dados precisam ser estruturados para gerar conhecimento.
  - Aprendizado de Máquina e Processamento de Linguagem Natural permitem análise de texto em uma escala inatingível por humanos.

Aprendizado Profundo em NLP

- O uso de redes neurais profundas representou avanços em uma gama de tarefas de processamento de texto.
  - Análise de sentimento;
  - Tradução por máquina;
  - Inferência em textos naturais.
- Problema: Generalização para diferentes domínios.
  - Transferência de aprendizado ajuda; mas dados rotulados de domínios específicos ainda são necessários—ajuste fino e avaliação.

Exemplos

DODF

O GOVERNADOR DO DISTRITO FEDERAL, no uso das atribuições que lhe confere o artigo 100, incisos XXVI e XXVII, da Lei Orgânica do Distrito Federal, resolve [...]

Acórdão STF

HABEAS CORPUS 110.260 SÃO PAULO RELATOR : MIN. LUIZ FUX PACTE.(S) :LAERCIO BRAZ PEREIRA SALES IMPTE.(S) :DEFENSORIA PÚBLICA DA UNIÃO PROC.(A/S)(ES) :DEFENSOR PÚBLICO-GERAL FEDERAL COATOR(A/S)(ES) :SUPERIOR TRIBUNAL DE JUSTIÇA

Objetivos

- Propor três bases de dados e gerar benchmarks de avaliação para cada uma delas.
- Em comum:
  1. Tarefas de NLP.
  2. Em português do Brasil.
  3. Específicas a algum domínio (jurídico ou de publicações oficiais).

Contribuições I

- Como principais contribuições, os conjuntos de dados:
  1. LeNER-Br dataset.
  2. VICTOR dataset.
  3. DODF dataset.
- Contribuições empíricas na forma de treinamento e avaliação de modelos em cada conjunto de dado.

Contribuições II

- Trabalho gerou seguintes publicações:
  - VICTOR: a dataset for Brazilian legal documents classification. [Luz de Araujo et al., 2020a]
  - Inferring the source of official texts: can SVM beat ULMFiT? [Luz de Araujo et al., 2020b]
  - Topic Modelling Brazilian Supreme Court Lawsuits[Luz de Araujo and de Campos, 2020].
  - LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text. [Luz de Araujo et al., 2018]
  - Sequence-aware multimodal page classification of Brazilian legal documents. Submetido ao International Journal on Document Analysis and Recognition.

LeNER-Br dataset

LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text

Entities

- Pessoa
- Tempo
- Organização
- Local

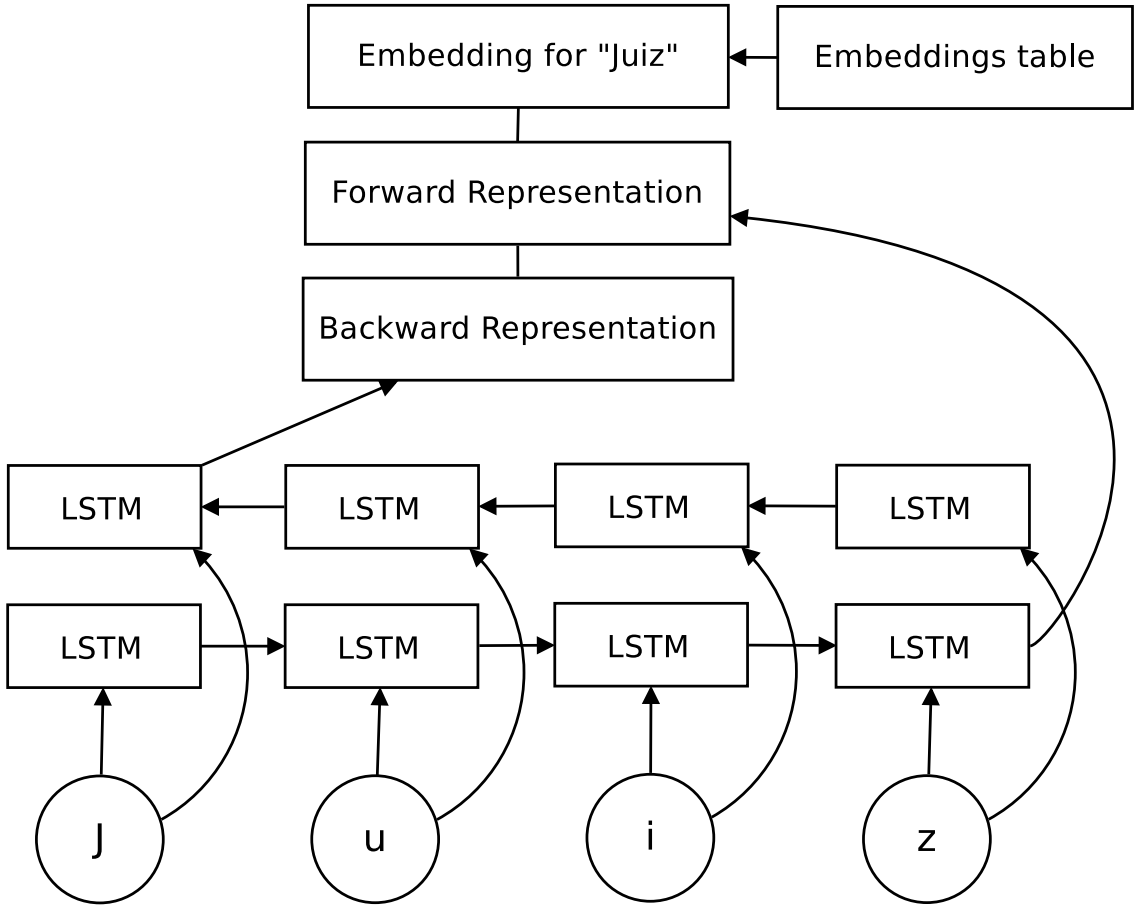
Pedro, estudante da Universidade de Brasília, foi para Canela em 2018.

Dados

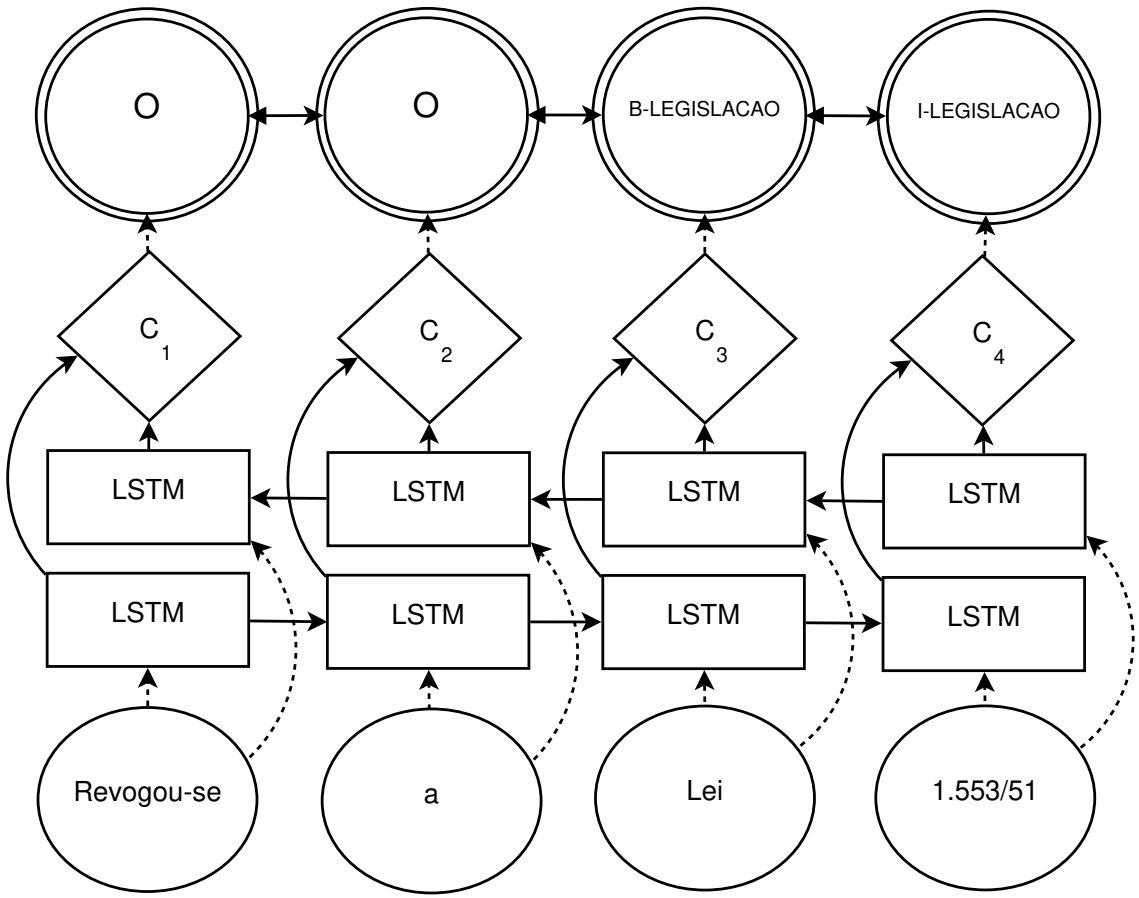
- Conjunto de 70 documentos jurídicos (acórdãos, leis e portarias) com anotação de entidades, totalizando 318.073 tokens; destas, 44.513 pertencentes a entidades.
- Entidades de 4 tipos genéricos (pessoa, tempo, organização e local) e 2 específicos ao domínio (legislação e jurisprudência).
- Documentos divididos entre mim e dois anotadores. Todos revisados por mim.

Modelagem I

- Treino de modelo char-biLSTM-CRF usando os dados.



Modelagem II



Resultados

Tabela: Resultados (em %) para reconhecimento de entidade no conjunto de teste.

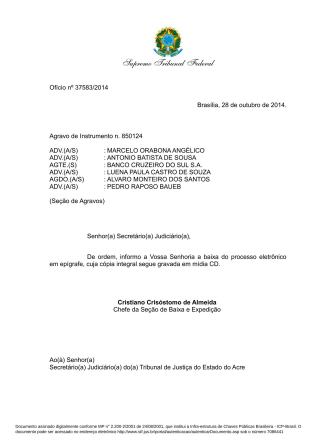
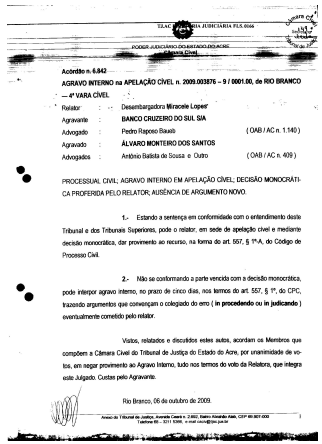
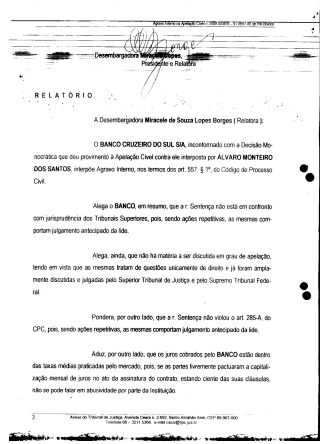
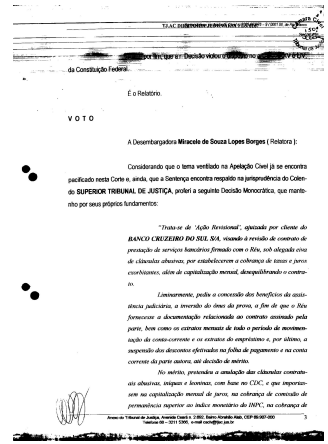
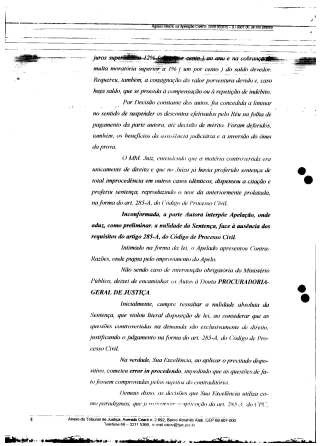
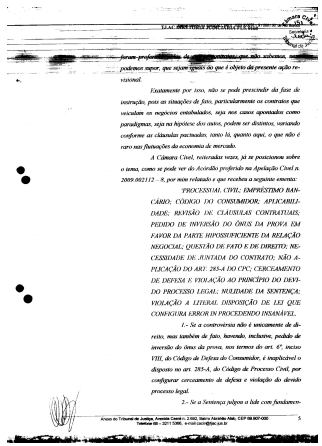
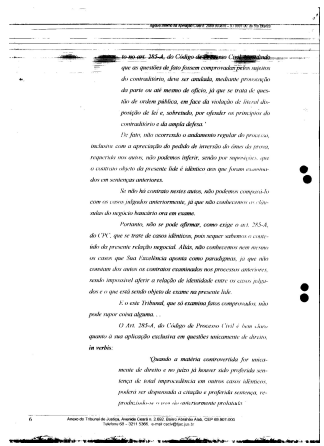
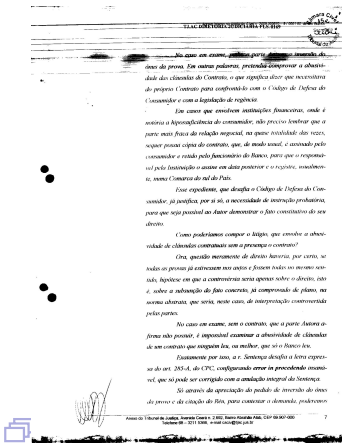
Entidade	Precisão	Revocação	F_1
Pessoa	85.58	78.97	82.14
Local	69.77	63.83	66.67
Organização	88.30	82.83	85.48
Tempo	91.30	87.50	89.36
Legislação	93.93	94.18	94.06
Jurisprudência	79.29	84.86	81.98
Média ponderada	87.98	85.29	86.61

VICTOR dataset

VICTOR: a dataset for Brazilian legal documents classification

VICTOR: a dataset for Brazilian legal documents classification

- Conjunto de dados com mais de 40.000 Recursos Extraordinários.
- Rótulos a nível de página (tipo de documento) e a nível de processo (tema de repercussão geral).

Dados I

- Três versões:
  1. Big VICTOR (BVIC), com 45.532 processos (692.966 documentos ou 4.603.784 páginas).
  2. Medium VICTOR (MVIc), com 44.855 processos (628.820 documentos ou 2.086.899 páginas).
  3. Small VICTOR (SVic), com 6.510 Extraordinary Appeals (94.267 documentos ou 339.478 páginas.

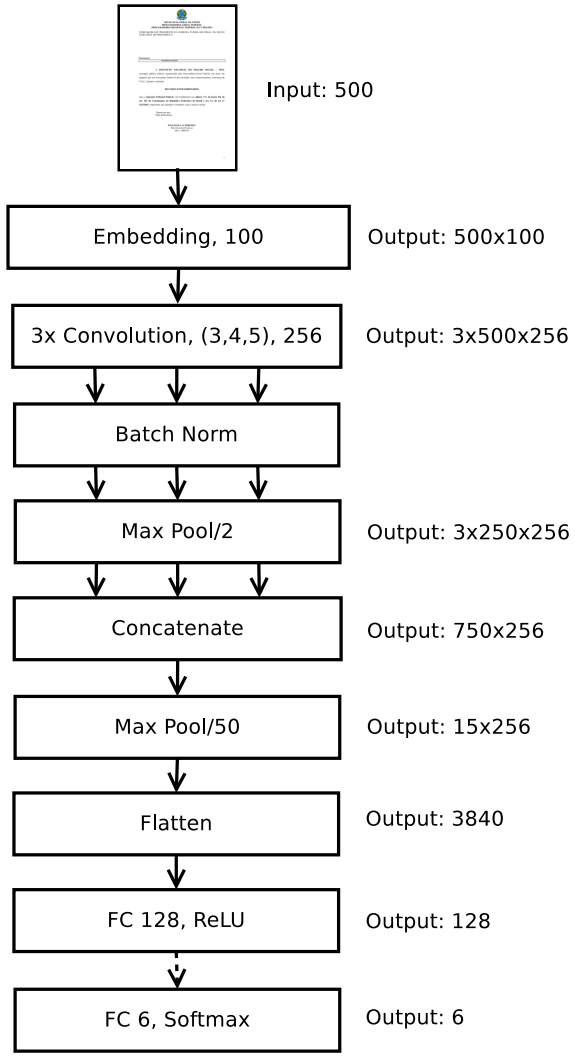
Dados II

- Tipo de documento (por página).
- Tema de repercussão geral (por processo, multi-rótulo).
  - "questões relevantes do ponto de vista econômico, político, social ou jurídico, que ultrapassem os interesses subjetivos da causa".
  - Tema 26 - Concessão de aposentadoria especial a policiais civis nos termos da Lei Complementar nº 51/1985.
  - Tema 33 - Relevância e urgência da medida provisória nº 2.170-36/2001 que disciplina a capitalização de juros com periodicidade inferior a um ano nas operações realizadas pelas instituições integrantes do Sistema Financeiro Nacional.

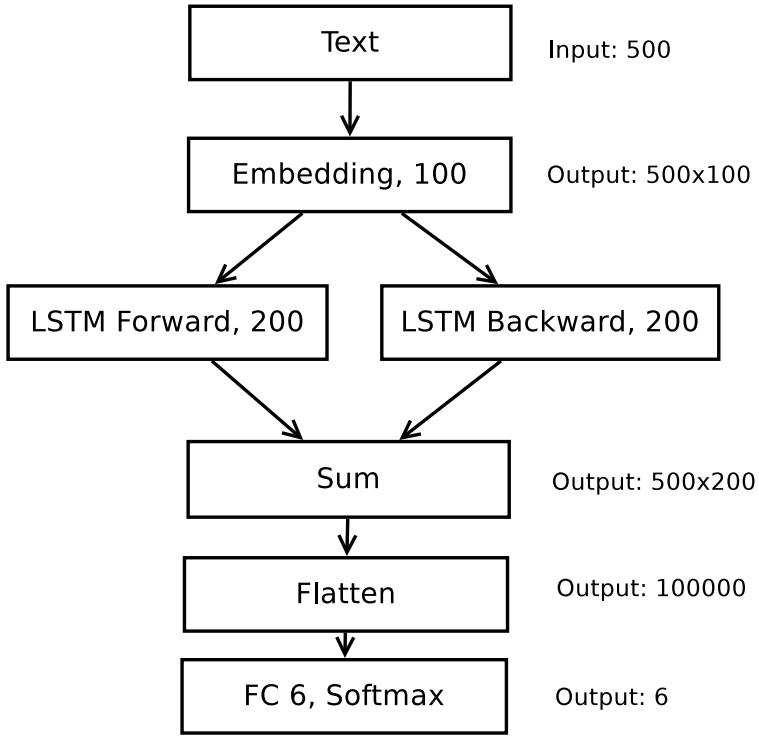
Modelagem I

- Classificação de tipo:
  - Saco-de-palavras + NB, SVM
  - CNN
  - LSTM
  - CNN + CRF
- Classificação de tema:
  - Saco-de-palavras + NB, SVM, XGBoost.
- Classificação de tema com conhecimento de domínio.

Modelagem II



Modelagem III



Resultados I

Tabela: Escores F_1 (em %) dos métodos para classificação de tipo de documento nos conjuntos de teste. Uma baseline de classe majoritária obtém escores F_1 com média ponderada de $87,06/84,41$ e aritmética de $15,90/15,73$ no MVic e SVic, respectivamente.

Conjunto	Modelo	Acórdão	ARE	Despacho	Outros	RE	Sentença	Ponderada	Artimética
MVic	NB	49.20	32.08	39.82	89.38	38.06	37.80	84.77	47.72
	SVM	65.41	52.62	59.34	95.85	64.52	69.75	92.88	67.92
	BiLSTM	**72.84**	57.82	**60.07**	97.11	67.74	69.96	94.33	**70.92**
	CNN	71.06	**58.11**	56.04	**97.37**	**68.71**	**72.35**	**94.64**	70.61
SVic	NB	66.40	36.07	51.15	93.24	55.89	55.99	88.93	59.79
	SVM	81.15	**58.06**	**67.88**	96.85	74.66	**79.30**	94.25	**76.32**
	BiLSTM	85.82	52.12	51.01	97.15	74.06	76.70	94.65	72.81
	CNN	**86.43**	55.92	59.88	**97.30**	**76.23**	79.29	**94.72**	75.84

Resultados II

Tabela: Escores F_1 (em %) antes e depois de processamento por CRF.

	MVic		SVic	
Classes	CNN	CNN-CRF	CNN	CNN-CRF
Acórd.	71.06	75.02 / +5.57%	86.43	90.60 / +4.82%
ARE	58.11	62.89 / +8.23%	55.92	59.54 / +6.47%
Desp.	56.04	62.55 / +11.62%	59.88	56.69 / -5.33%
Outros	97.37	97.66 / +0.30%	97.30	97.68 / +0.39%
RE	68.71	74.38 / +8.25%	76.23	78.77 / +3.33%
Sent.	72.35	77.77 / +7.49%	79.29	81.13 / +2.32%
M.p.	94.64	95.37 / +0.77%	94.72	95.33 / +0.64%
M.a.	70.61	75.05 / +6.29%	75.84	77.40 / +2.06%

Resultados III

Tabela: Escores F_1 (em %) dos métodos para classificação de tema nos conjuntos de teste. Uma baseline que sempre atribui todos os temas obtém escores F_1 com média ponderada de $41,17$ /$40,87$/$10,87$ e aritmética de $5,48$/$5,49$/$6,52$ em BVic, MVic e SVic, respectivamente.

	BVic			MVic			SVic		
Temas	NB	SVM	XGBoost	NB	SVM	XGBoost	NB	SVM	XGBoost
0	81.63	87.35	**90.70**	79.50	88.85	**92.41**	49.90	**72.29**	69.71
5	17.95	92.47	**94.15**	18.73	79.05	**85.50**	30.22	**84.79**	82.87
6	65.85	61.65	**77.84**	37.45	36.52	**76.81**	21.93	63.11	**77.03**
26	60.38	92.06	**93.33**	14.59	36.48	**94.74**	12.75	**97.44**	94.44
33	30.03	46.32	**77.17**	8.35	14.42	**78.62**	30.71	57.78	**74.65**
139	61.82	81.25	**90.57**	17.54	74.67	**92.59**	14.95	88.89	**94.34**
163	77.38	75.41	**86.09**	25.05	76.19	**88.00**	73.86	86.08	**94.67**
232	40.93	44.64	**69.33**	27.63	13.90	**55.12**	37.32	65.00	**65.08**
313	47.42	58.56	**72.55**	31.11	43.37	**80.77**	60.22	76.12	**82.69**
339	23.17	52.12	**74.47**	20.62	45.84	**77.04**	26.73	74.38	**86.06**
350	73.27	55.26	**86.96**	73.27	12.05	**89.58**	85.06	52.94	**90.11**
406	57.41	44.44	**85.71**	20.27	10.41	**85.71**	55.81	46.15	**84.93**
409	74.42	79.12	**86.25**	29.03	72.64	**90.68**	91.14	90.91	**95.48**
555	39.02	65.06	**83.33**	0.00	17.06	**84.75**	47.06	52.46	**88.89**
589	77.97	82.01	**88.00**	35.02	63.44	**88.71**	82.05	90.16	**90.76**
597	**96.77**	90.91	96.55	53.57	90.91	**96.55**	85.71	88.24	**96.77**
634	89.87	90.91	**95.48**	70.24	89.29	**94.19**	92.81	93.08	**95.42**

Resultados IV

Tabela: Escores F_1 (em %) dos métodos para classificação de tema nos conjuntos de teste. Uma baseline que sempre atribui todos os temas obtém escores F_1 com média ponderada de $41,17$ /$40,87$/$10,87$ e aritmética de $5,48$/$5,49$/$6,52$ em BVic, MVic e SVic, respectivamente.

	BVic			MVic			SVic		
Temas	NB	SVM	XGBoost	NB	SVM	XGBoost	NB	SVM	XGBoost
660	51.23	74.14	**89.00**	35.30	80.39	**90.07**	36.41	91.10	**93.51**
695	93.27	**97.65**	96.65	95.37	**98.13**	96.68	96.52	**98.49**	96.94
729	**100.00**	**100.00**	97.78	62.07	**95.65**	93.02	63.16	**100.00**	93.33
766	21.88	73.21	**77.65**	21.82	76.64	**82.61**	19.81	81.08	**86.67**
773	68.03	96.40	**97.06**	61.54	95.71	**98.55**	81.30	**94.03**	93.13
793	66.67	84.52	**92.96**	28.26	86.23	**91.43**	26.59	87.80	**90.79**
800	87.70	98.42	**98.73**	87.34	98.41	**98.62**	69.86	**92.71**	91.10
810	62.28	88.72	**95.32**	23.89	92.16	**94.87**	21.06	**95.62**	94.69
852	64.67	82.61	**87.34**	54.40	76.68	**89.74**	49.08	89.41	**92.31**
895	25.10	63.68	**89.66**	14.64	94.08	**98.32**	24.07	92.17	**95.93**
951	94.74	**100.00**	99.54	39.04	98.21	**98.62**	57.36	**99.50**	95.29
975	86.15	91.67	**94.44**	15.62	68.69	**91.43**	41.61	**89.74**	**89.74**
M.p.	69.55	82.35	**89.57**	60.62	81.37	**90.72**	48.75	82.31	**86.34**
M.a.	63.35	77.61	**88.43**	37.97	66.42	**88.82**	51.21	82.46	**88.87**

Resultados V

Tabela: Escores F_1 (em %) de um modelo XGBoost treinado com e sem páginas da classe *Outros* em um subconjunto de teste do BVic que inclui somente processos com ao menos uma página não rotulada como *Outros*.

Temas	Sem	Com
M.p.	84.55	**90.27**
M.a.	66.20	**74.42**

Topic modelling Brazilian Supreme Court lawsuits

Topic modelling Brazilian Supreme Court lawsuits

- Modelagem de tópicos
  - Não supervisionada.
  - Organização e exploração de quantidades massivas de dados.
  - Encontra tópicos latentes presentes em uma coleção de documentos.
    - ★ Tópico como distribuição de palavras.
- Análise qualitativa (10 e 30 tópicos) e quantitativa (10, 30, 100, 300 e 1.000 tópicos).
- Dados: BVic (45.532 processos).

Modelagem I

- Alocação de Dirichlet latente (LDA).
  - Cada documento tem uma distribuição de tópicos θ_i gerada por uma distribuição de Dirichlet Dir(α)
  - Cada tópico tem uma distribuição de palavras ϕ_j gerada por uma distribuição de Dirichlet Dir(β)
  - Cada palavra de cada documento é gerada: escolhendo um tópico de θ_i e uma palavra do tópico escolhido.
- Aprendizado consiste em inferir as distribuições de cada documento e de cada tópico.
- XGBoost para avaliação quantitativa.
  - Tópicos vs sacos-de-palavras.

Resultados I

Tabela: Rótulos atribuídos aos tópicos e as correspondentes palavras mais relevantes (10 tópicos).

Tópico	λ	Rótulo atribuído	Palavras
1	0,6	Remuneração de servidor público	servidores, servidor, prescrição, remuneração
2	0	Direito Penal	entorpecente, hidrômetro, clandestino, interrogatório,
3	0,6	Direito Previdenciário	benefício, evento, aposentadoria, previdenciário,
4	0,6	Direito Civil	banco, contrato, consumidor, projudi
5	0,6	Direito à Saúde	saúde, município, municipal, medicamentos
6	0,4	Erros de OCR	ento, não, ro, co
7	0,6	Direito Tributário	icms, ipi, imposto, receita
8	0	Entidades	econorte, rcte, pieter, bruyn
9	0,4	Questões trabalhistas	fgts, pss, horas, folha
10	0,6	Publicidade do processo	original, site, acesse, informe

Resultados II

Tabela: Rótulos atribuídos aos tópicos e as correspondentes palavras mais relevantes (30 tópicos).

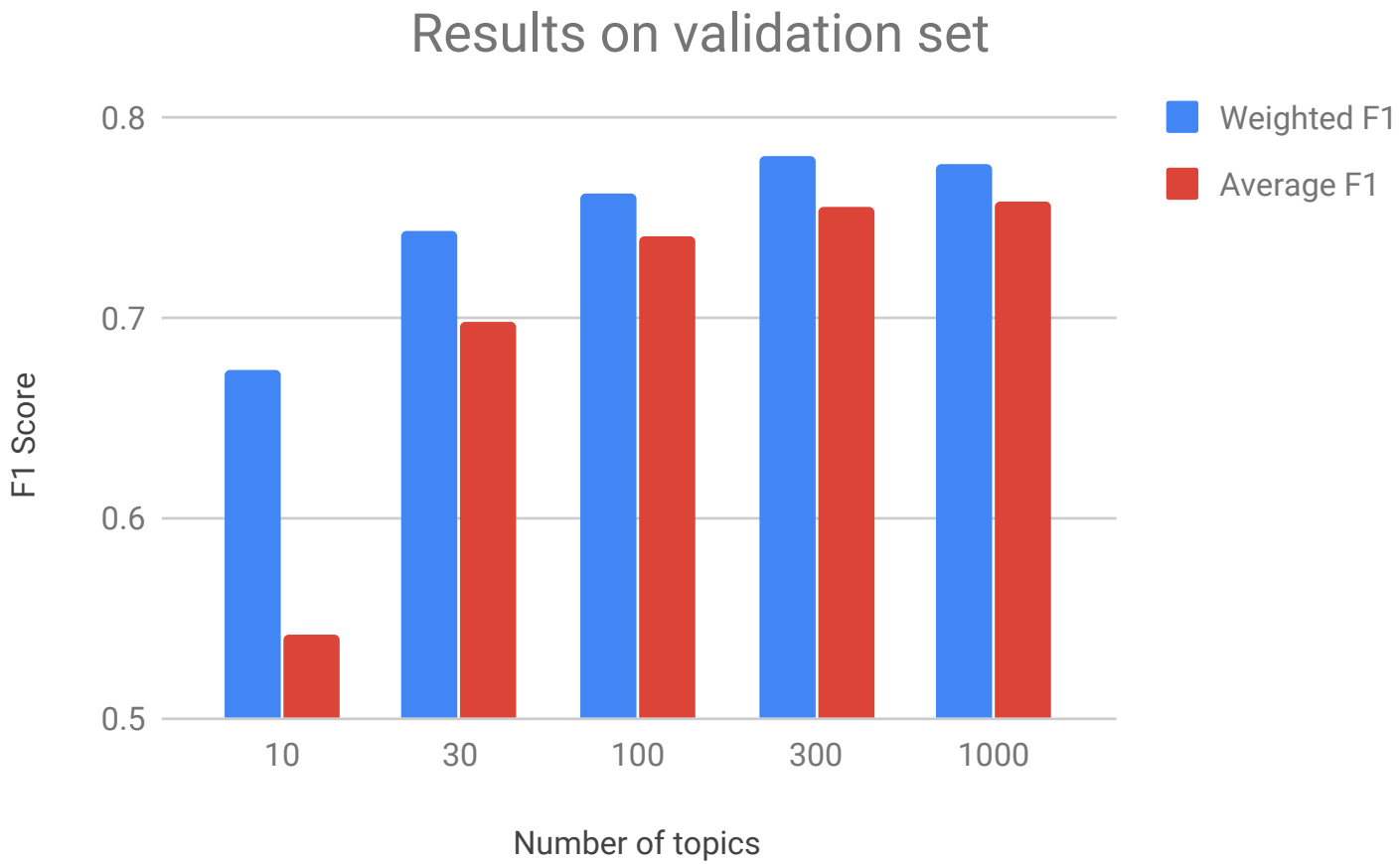
Tópico	λ	Rótulo atribuído	Palavras
1	0,6	Responsabilidade Civil	dano, danos, indenização, moral
2	0,22	Decadência benefício previdenciário	benefício, decadência, teto, previdenciário
3	0,6	Direito Tributário	fazenda, tributário, receita, imposto
4	0,1	Miscelânea - Vocabulário jurídico, entidades e leis	n^{o}série, pet, carimbo, itaperuna
5	0,4	Gratificação de servidores públicos	gratificação, desempenho, inativos, avaliação
6	0,4	Previdência rural	rural, contribuição, LEI_8212, aposentadoria
7	0,6	Reajuste de vencimento de servidor	reajuste, servidores, vencimentos, urv
8	0,4	Erros de OCR	ento, não, ro, ffl
9	0,6	Servidores militares	militar, servidor, militares, servidores
10	0	Direito Penal	clandestino, sepetiba, semiaberto, entorpecente
11	0,4	Direito Contratual/Comercial	contrato, contratos, taxa, contas
12	0,05	Conselhos técnicos	confea, crea, agronomia, LEI_6496
13	0,2	Concursos públicos	concurso, candidato, edital, vagas
14	0,4	Antecipação de reajuste de vencimento	upag, pccs, trabalhista, LEI_8460
15	0,6	Direito à Saúde	saúde, medicamentos, tratamento, medicamento
16	0,9	Poupança, juros e correção monetária	correção, monetária, poupança, mora
17	0,6	Publicidade do processo	original, site, acesse, informe
18	0,6	Reclamações trabalhistas	estran, tst, entidade, reclamante
19	0,4	Miscelânea- Direito do Consumidor e Bahia	consumidor, salvador, bahia, pdf

Resultados III

Tabela: Rótulos atribuídos aos tópicos e as correspondentes palavras mais relevantes (30 tópicos).

Tópico	λ	Rótulo atribuído	Palavras
20	0	Entidades - nomes	lauxen, tainá, heloise, soeli
21	0,7	Qualificações	num, normal, internamento, foz
22	0,5	Seguros	seguro, previd, instituto, dpu
23	0,4	Folhas de pagamento	horas, fgts, folha, extras
24	0	Miscelânea-Assembleias, estatutos e palavras estrangeiras	andaterra, peixer, funds, market
25	0,5	Documentos fiscais	ltda, ipi, nfe, icms
26	0,4	Processos relacionados ao Rio Grande do Sul	sul, grande, alegre, paese
27	0,4	Imposto de Renda	atualizado, meses, rra, irpf
28	0,2	Direito tributário-Circulação de mercadorias	compatível, issqn, saída, eireli
29	0,2	Miscelânea-Movimentação processual e Paraná	paraná, arq, curitiva, mov
30	0,4	Pagamentos	jam, vlr, recolhido, crédito

Resultados IV



Resultados V

Tabela: Escores F_1 (em %) obtido por cada método de representação de processo no conjunto de teste. Um classificador que sempre atribui todos os temas obtém um escore F_1 com média ponderada 41.17 e aritmética 5.48.

Tema	Contagem	Tf-idf	300 tópicos
0	**90.11**	89.63	88.12
5	94.12	**95.81**	93.36
6	68.00	**77.99**	70.79
26	**96.67**	91.53	75.47
33	**82.87**	79.55	67.42
139	86.27	**88.46**	72.00
163	84.35	**86.49**	81.33
232	65.28	**70.67**	52.86
313	70.00	**76.92**	75.93
339	**77.53**	76.29	19.31
350	**83.87**	79.57	82.22
406	84.06	**87.32**	78.26
409	86.79	**87.90**	83.13
555	**80.00**	70.37	50.00
589	**87.80**	86.40	85.94
597	**96.77**	**96.77**	92.86
634	92.72	**95.36**	90.91
660	88.81	**88.87**	52.45
695	**96.65**	**96.65**	96.62
729	95.45	95.45	**97.78**
766	75.61	**82.76**	48.72

Resultados VI

Tabela: Escores F_1 (em %) obtido por cada método de representação de processo no conjunto de teste. Um classificador que sempre atribui todos os temas obtém um escore F_1 com média ponderada 41.17 e aritmética 5.48.

Tema	Contagens	Tf-idf	300 tópicos
773	**96.35**	96.30	94.74
793	89.36	**92.31**	80.00
800	**98.74**	98.41	95.20
810	**94.58**	93.42	83.77
852	84.77	**85.91**	80.00
895	97.33	**97.67**	18.65
951	**99.54**	**99.54**	97.67
975	94.29	**98.55**	92.96
M.p.	**89.29**	89.22	78.07
M.a.	87.54	**88.37**	75.81

Sequence-aware multimodal page classification of Brazilian legal documents

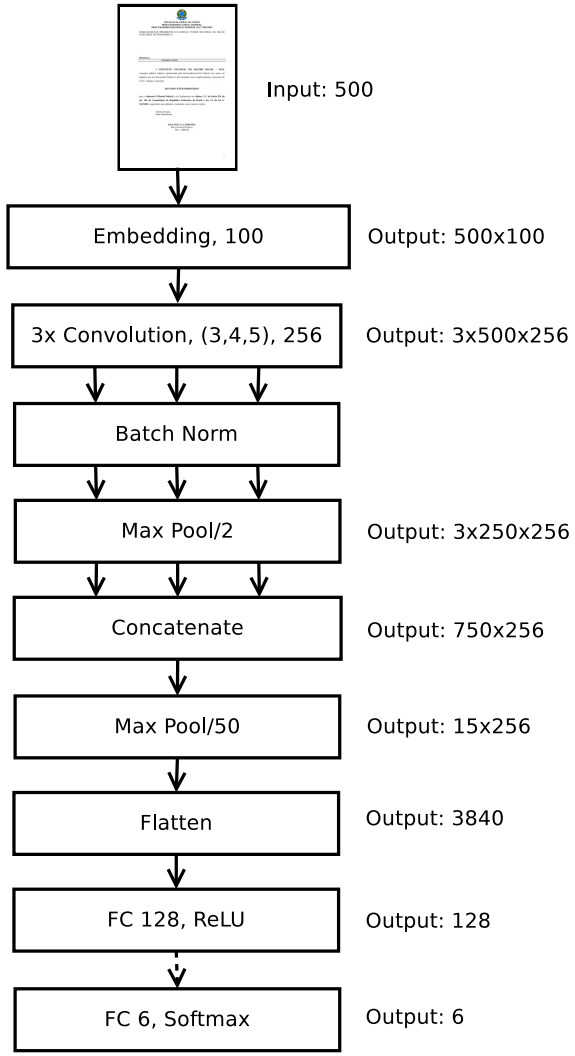
Sequence-aware multimodal page classification of Brazilian legal documents

- Extender SVic com imagens das páginas dos processos.
- Combinar texto + imagem + sequência para classificação de página de processo.

Dados utilizados

- 6.510 recursos extraordinários $\rightarrow$ 339.478 páginas.
- Texto armazenado em csv extraído por sistemas de OCR e pré-processado (letras minúsculas, remoção de stop words).
- Imagem armazenada em JPEG extraída de PDF.
- Seis tipos de documento: (ARE, Petição de RE, Despacho, Sentença, Acórdão e outros).
- 33.840 imagens sem texto correspondente e 4 textos sem imagem.

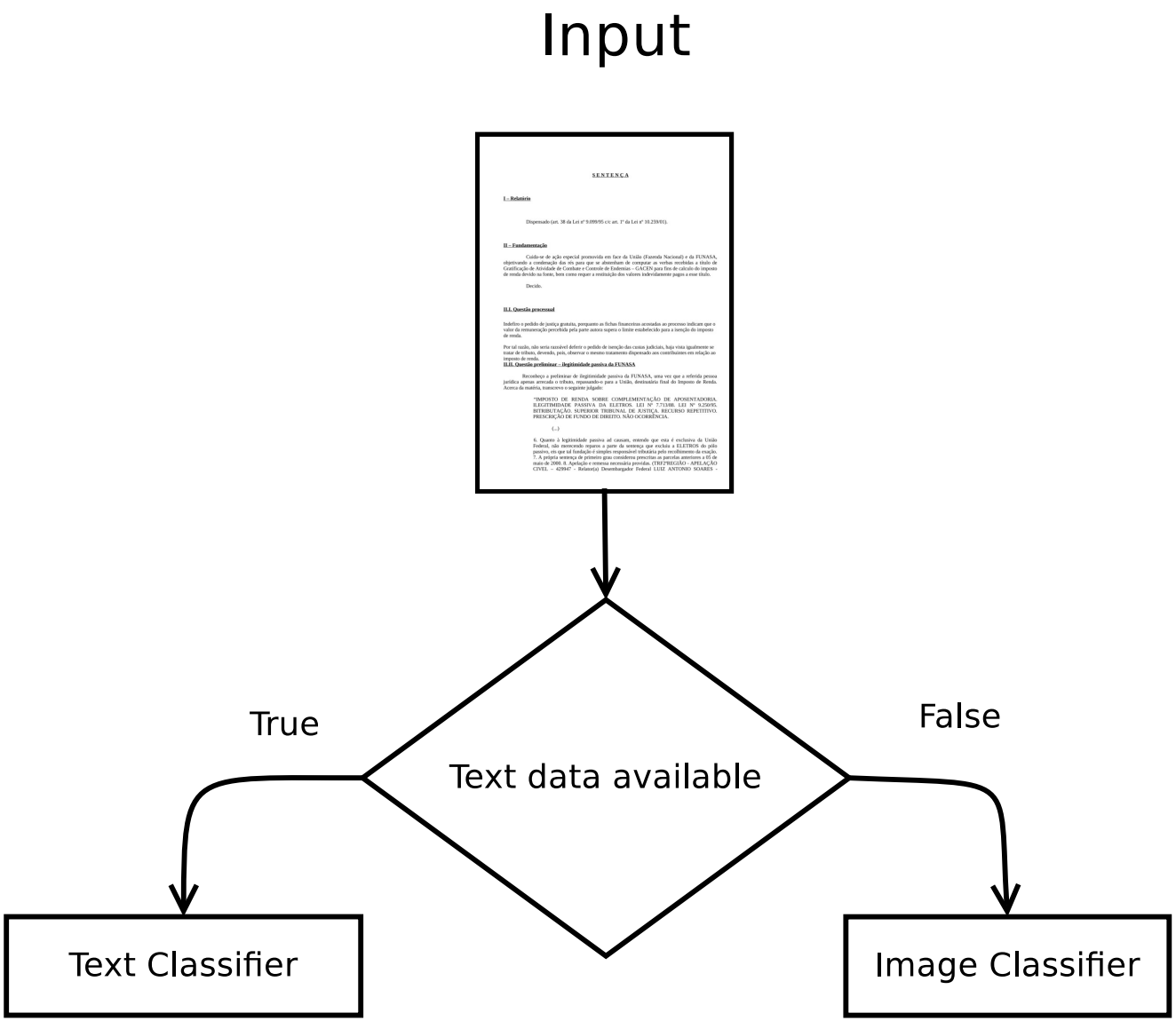
Classificação de texto



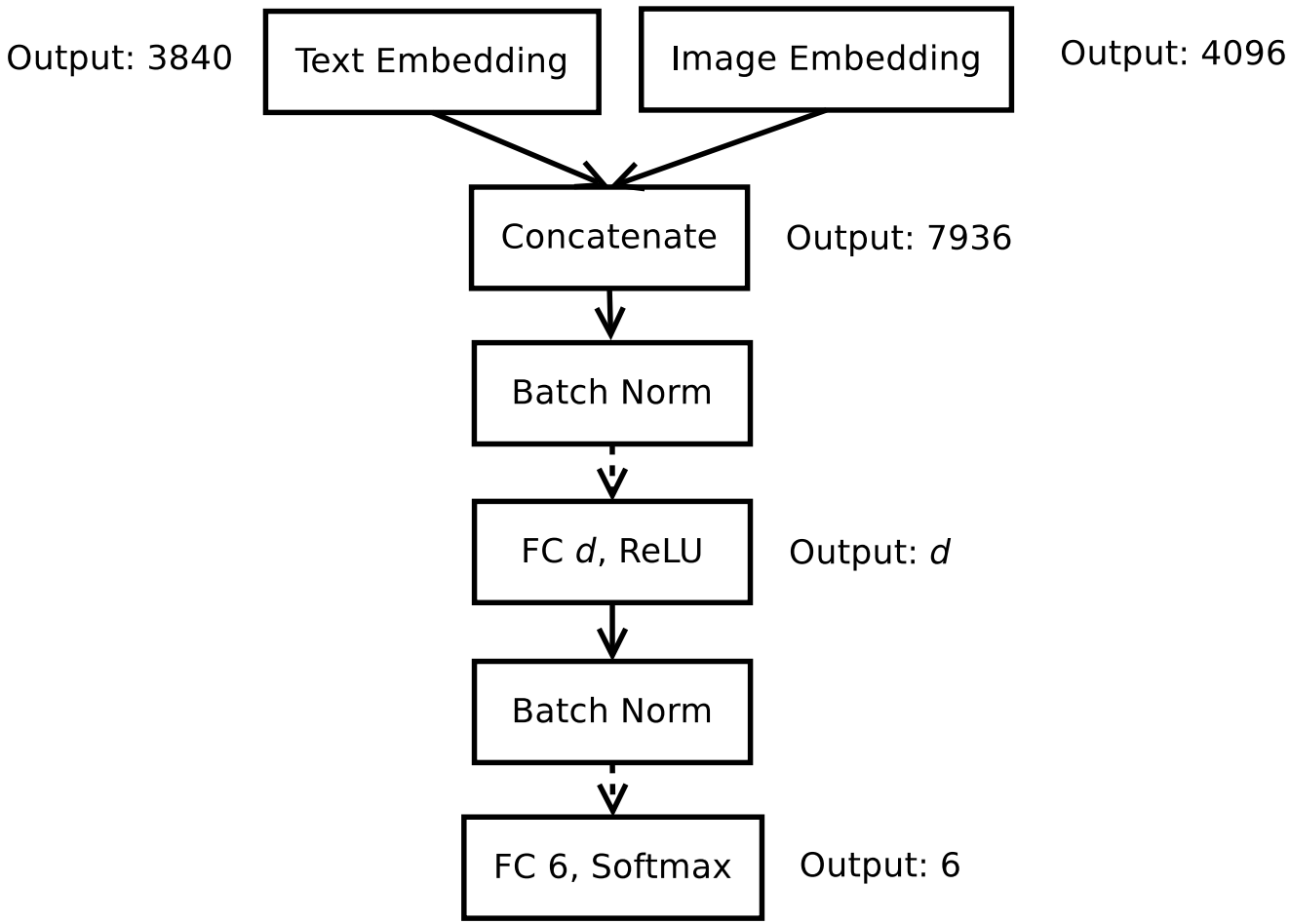
Classificação de imagem

- Fine-tune de Resnet50 [He et al., 2016] pré-treinada na base de dados ImageNet [Russakovsky et al., 2015].
  1. Treina a cabeça por um epoch.
  2. Treina todas as camadas por seis epochs.

Baseline para Fusão

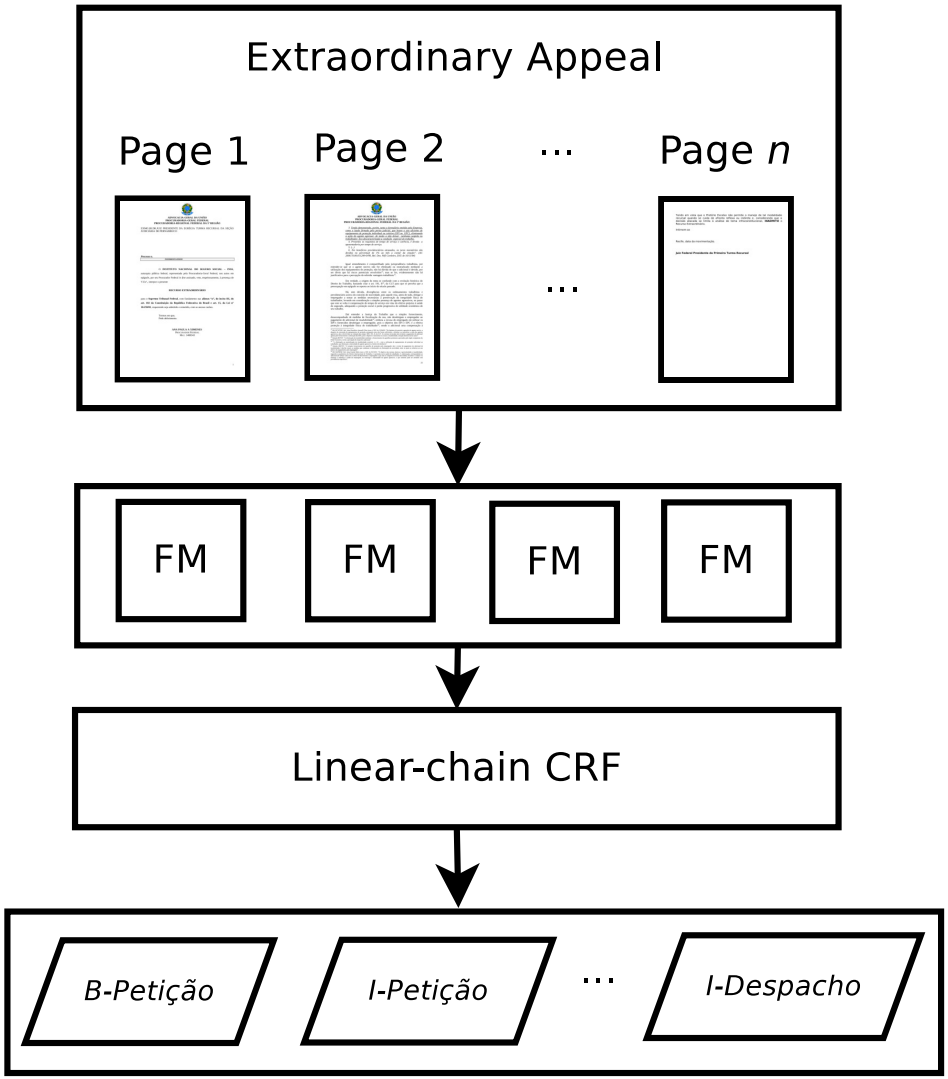


Módulo de fusão



- Embeddings aprendíveis para entradas ausentes.

Classificação de sequência (pós-processamento com CRF)



Classificação de sequência (biLSTM)

- 4 variantes de redes recorrentes:
  1. BiLSTM
  2. BiLSTM-CRF
  3. BiLSTM-F
  4. BiLSTM-F-CRF

Resultados I

Tabela: Escores F_1 (em %, conjunto de teste) dos principais métodos para classificação de imagem, texto, fusão e sequência. Resultados de imagem usando o conjunto de teste de imagens; todos os outros, o conjunto de teste de texto. Um classificador majoritário obtém um escore F_1 com média ponderada de 84.41/84.07 e aritmética de 15.73/15.71 nos conjuntos de teste de texto e imagem, respectivamente.

Classe	Texto	Imagem		Fusão	Sequência	
	CNN	Resnet50-w	Resnet50	FM	FM+CRF	BiLSTM-F
Acórdão	89.96	18.45	06.78	90.74	**91.56**	88.97
ARE	55.72	11.33	00.00	57.92	60.74	**61.16**
Despacho	62.94	08.44	00.00	63.98	62.69	**64.07**
Others	97.31	61.72	95.02	97.24	**97.67**	97.46
RE	75.59	32.59	34.96	75.47	78.43	**79.67**
Sentença	80.53	43.52	48.67	82.04	83.42	**85.26**
M.a.	77.01	29.34	30.91	77.90	79.09	**79.43**
M.p.	94.72	58.09	87.67	94.72	**95.38**	95.30

Resultados II

Tabela: Impacto do número de unidades ocultas e aprendizado de embeddings para entradas ausentes nos escores F_1 (em %, conjunto de validação). O sufixo *-zero* indica o uso de vetor de zeros para entradas ausentes.

Method	Average F_1
FM-512	74.49
FM-512-zero	68.02
FM-128	**75.70**
FM-128-zero	72.95

Resultados III

Tabela: Escores F_1 (em %, conjunto de teste) do classificador híbrido e de uma versão do módulo de fusão que ignora a entrada de imagem (sem ativ. img.) comparados aos do módulo de fusão. Para o classificador híbrido reportamos resultados usando ambos classificadores de imagem: com (HC-w) e sem (HC) penalização de classes frequentes. Entre parênteses, a diferença de performance quando se compara com o módulo de fusão original (FM).

Classe	Text test split		Text + image test split		
	FM	fusão sem ativ. img.	FM	HC-w	HC
Acórdão	90.74	88.27 (-2.47)	88.5	41.36 (-47.14)	87.68 (-0.82)
ARE	57.92	54.09 (-3.83)	56.6	49.02 (-7.58)	43.91 (-12.69)
Despacho	63.98	62.01 (-1.97)	63.79	42.71 (-21.08)	61.85 (-1.94)
Others	97.24	97.27 (+0.03)	97.03	95.80 (-1.23)	97.02 (-0.01)
RE	75.47	73.26 (-2.21)	75.05	72.11 (-2.94)	75.00 (-0.05)
Sentença	82.04	79.58 (-2.46)	81.21	74.07 (-7.14)	79.68 (-1.53)
M.a.	77.90	75.74 (-2.16)	77.03	62.51 (-14.52)	74.19 (-2.84)
M.p.	94.72	94.47 (-0.25)	94.32	92.58 (-1.74)	93.95 (-0.37)

Resultados IV

Tabela: Comparação entre os escores F_1 (em %, conjunto de validação) dos diferentes métodos para classiifcação de sequência usando LSTM.

Method	Average F_1	Weighted F_1
BiLSTM	77.16	94.25
BiLSTM-CRF	78.45	94.46
BiLSTM-F	**79.03**	**94.81**
BiLSTM-F-CRF	78.87	94.58

DODF dataset

Inferring the source of official texts: can SVM beat ULMFiT?

- Conjunto de dados com textos do DODF rotulados (órgão de origem) e não rotulados.
- Samples rotulados e não rotulados $\rightarrow$ ULMFiT [Howard and Ruder, 2018].

Dados

- 717 documentos com anotação para 19 órgãos públicos originários.
- 1,926 documentos sem rótulos usados para adaptar um modelo de linguagem pré-treinado. Total de 984.580 tokens.

Métodos rasos

- Dois tipos de sacos-de-palavras:
  - valores tf-idf;
  - contagens de token.
- Dois classificadores:
  - Naïve Bayes (NB);
  - Support Vector Machine (SVM).

Transfer learning

Universal Language Model Fine-Tuning (ULMFiT) [Howard and Ruder, 2018] [1]

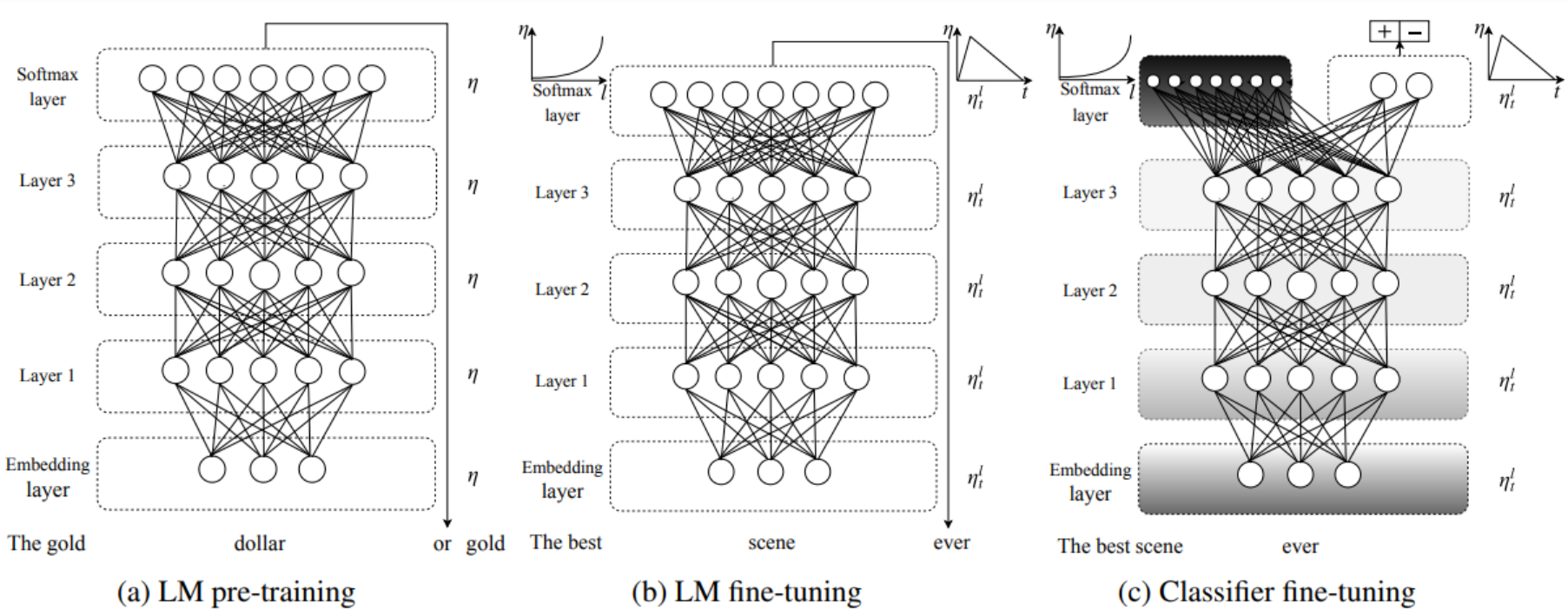


Figura: Imagem retirada de [Howard and Ruder, 2018].

[1] Modelo pré-treinado em 166.580 artigos da Wikipedia (100.255.322 tokens) disponível em `https://github.com/piegu/language-models/tree/master/models`.

Resultados I

Tabela: Escores F_1 (em %) das classes no conjunto de teste.

Classe	NB	SVM	F-ULMFiT	B-ULMFiT	F+B-ULMFiT	Quantidade
Casa Civil	66.67	78.95	80.00	82.35	**88.24**	18
Controladoria	80.00	80.00	**100.00**	**100.00**	**100.00**	2
Defensoria Pública	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	8
Poder Executivo	80.00	85.71	78.26	**90.91**	86.96	10
Poder Legislativo	66.67	**100.00**	66.67	66.67	**100.00**	1
Agricultura	50.00	**66.67**	57.14	50.00	57.14	4
Cultura	**91.67**	**91.67**	**91.67**	**91.67**	**91.67**	13
Desenv. Econômico	**66.67**	**66.67**	**66.67**	**66.67**	**66.67**	4
Desenv. Urbano	75.00	75.00	75.00	**85.71**	75.00	4
Economia	66.67	**100.00**	**100.00**	**100.00**	**100.00**	1
Educação	76.19	**91.67**	81.48	75.00	88.00	13
Fazenda	90.48	90.48	95.00	95.24	**97.56**	21
Justiça	**75.00**	66.67	60.00	66.67	66.67	5
Obras	88.24	**90.91**	88.24	76.92	85.71	18
Saúde	92.75	92.31	92.31	94.12	**95.52**	32
Segurança Pública	**98.99**	94.34	94.34	97.09	94.34	50
Transporte	94.74	**97.56**	92.31	92.31	**97.56**	20
Meio Ambiente	**100.00**	**100.00**	66.67	0.00	0.00	2
Tribunal de Contas	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	11
F_1 Médio	82.09	**87.82**	83.46	80.6	83.74	237
F_1 Ponderado	88.68	90.49	88.90	88.88	**90.88**	237
Acurácia	88.61	90.72	89.03	89.45	**91.56**	237

Conclusões

Conclusões I

- Um dos principais desafios de pesquisa em NLP é o treino de modelos que generalizam bem. Dados anotados de diferentes domínios são necessários para subsidiar esforços nesse sentido.
- Apresentamos três novos conjuntos de dados de domínios específicos em português, e treinamos e avaliamos diferentes modelos usando cada um deles.

Conclusões II

- Achados:
  - Um modelo BiLSTM-CRF treinado nos dados do LeNER-Br é capaz de reconhecer tanto entidades específicas do domínio quanto entidades gerais sem a necessidade de pré-processamento específico ou engenharia de características.
  - Modelos saco-de-palavras podem atingir resultados comparáveis aos de modelos de aprendizado de profundo.
    - ★ Ainda mais em conjuntos pequenos.
  - Tópicos extraídos usando LDA podem servir como ponto de partida para organização de casos do STF.
  - Classificação de página de RE melhora com cada modalidade adicional de entrada.

Limitações

- Modelos treinados apenas para apoiar e encorajar trabalhos futuros.
  - Sem buscas extensivas por hiper-parâmetros em modelos de aprendizado profundo.
  - Busca de hiper-parâmetros e modelos mais recentes (Ex: BERT [Devlin et al., 2018], ELMo [Peters et al., 2018]).
- Módulos do sistema multimodal treinados separadamente.
  - Treino ponta-a-ponta do sistema.
- Documentos não anotados por mais de uma pessoa: não foi possível calcular métricas de concordância entre anotadores.
  - Revisão cuidadosa (LeNER-Br e DODF).
  - Fluxo de trabalho diário do STF (VICTOR).

Referências I

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018).

BERT: pre-training of deep bidirectional transformers for language understanding.

CoRR, abs/1810.04805.

He, K., Zhang, X., Ren, S., and Sun, J. (2016).

Deep residual learning for image recognition.

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Howard, J. and Ruder, S. (2018).

Fine-tuned language models for text classification.

CoRR, abs/1801.06146.

Luz de Araujo, P. H. and de Campos, T. E. (2020).

Topic modelling brazilian supreme court lawsuits.

In *International Conference on Legal Knowledge and Information Systems (JURIX)*, Frontiers in Artificial Intelligence and Applications, pages 113–122, Prague, Czech Republic. IOS Press.

Luz de Araujo, P. H., de Campos, T. E., Ataides Braz, F., and Correia da Silva, N. (2020a).

VICTOR: a dataset for Brazilian legal documents classification.

In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France. European Language Resources Association.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018).

Lener-br: a dataset for named entity recognition in brazilian legal text.

In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Canela, RS, Brazil.

Referências II

Luz de Araujo, P. H., de Campos, T. E., and Magalhaes Silva de Sousa, M. (2020b).
Inferring the source official texts: can SVM beat ULMFiT?
In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), Evora, Portugal. Springer.
Code and data available from `https://cic.unb.br/~teodecampos/KnEDLe/`.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015).
ImageNet Large Scale Visual Recognition Challenge.
International Journal of Computer Vision (IJCV), 115(3):211–252.