Segmentação e Classificação Semântica de Trechos de Diários Oficiais Usando Aprendizado Ativo

Kattiana Constantino[1], Victor Augusto L. Cruz[1], Otávio M. M. Zucheratto[1], Celso França[1], Marcos Carvalho[1], Thiago H. P. Silva[2], Alberto H. F. Laender[1], Marcos André Gonçalves[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
31270-901 – Belo Horizonte – MG

{kattiana,victoraugusto,otaviozucheratto,celsofranca}@dcc.ufmg.br

{marcoscarvalho,laender,mgoncalv}@dcc.ufmg.br

[2]Departamento de Computação
Universidade Tecnológica Federal do Paraná
85905-490 – Toledo – PR

thiagosilva@utfpr.edu.br

Abstract. *Unrestricted and monitorable access to laws and regulations is an essential presupposition of democracy. This allows, for example, the detection of illicit acts and the monitoring of fraud in public actions (e.g., bids). However, each federated entity follows its own criteria for standardizing models and format in making this information available, for example, in municipal, state and Union official journals. In this context, our objective is to minimize the effort to deal with the textual extraction of these essential data by proposing a structure-oriented heuristic to segment excerpts from public documents, notably official journals. Subsequently, we semantically classify the extracted snippets with an active learning strategy that minimizes manual labeling effort. As a result of these efforts, we developed an annotation prototype integrated into the classification process, achieving 100% accuracy in extraction and 85% in classification with very little labeling effort.*

Resumo. *Acesso irrestrito e monitorável a leis e regulamentações é pressuposto essencial da democracia. Isso permite, por exemplo, a detecção de ilícitos e o monitoramento de fraudes em ações públicas (e.g., licitações). Contudo, cada ente federado segue seus próprios critérios de padronização de modelos e formato na disponibilização dessas informações, por exemplo, nos diários oficiais municipais, estaduais e da União. Nesse contexto, nosso objetivo é minimizar o esforço para lidar com a extração textual desses dados ao propor uma heurística orientada à estrutura para segmentar os trechos de documentos públicos. Posteriormente, classificamos semanticamente os trechos extraídos com uma estratégia de aprendizado ativo que minimiza o esforço manual de rotulação. Como resultado desses esforços, desenvolvemos um protótipo de anotação integrado ao processo de classificação, obtendo uma acurácia de 100% na extração e de 85% na classificação com muito pouco esforço de rotulação.*

1. Introdução

O acesso a dados públicos é relevante não só para observarmos as decisões dos entes federados (União, Estados, Distrito Federal e Municípios), mas também para acompanharmos como são definidas e executadas as políticas públicas destinadas à população, possibilitando assim democratizar as licitações e os pregões públicos, bem como fiscalizar as receitas e os gastos de cada órgão governamental [Pinto et al. 2021, Rangel et al. 2020]. As garantias de acesso aos dados públicos estão previstas na Constituição Federal Brasileira de 1988, sendo regulamentadas pela Lei de Acesso à Informação (Lei Nº 12.527) de 18 de Novembro de 2011[1]. Com a regulamentação dessa lei, tornou-se obrigatório a disponibilização dos dados públicos em sítios oficiais de cada um dos entes federados. Desse modo, é possível propor diversas análises com auxílio de técnicas estatísticas e computacionais para fins de monitoramento das atividades governamentais como, por exemplo, a detecção de fraudes em licitações ou editais de compras públicas.

Entretanto, não há uma padronização da forma como os atos governamentais são documentados (e.g., por meio de textos ou imagens), bem como se serão dispostos nos seus respectivos sítios ou se há rótulos associados a cada um dos atos. Além disso, nem todos os entes federados seguem o mesmo modelo de formatação para edição e organização dos documentos oficiais, havendo diversas formas de gerenciamento de documentos, bem como distintos meios de elaboração sem que haja respaldo de um modelo preciso de gestão oficial.

Este artigo aborda um esforço que visa lidar especificamente com dois aspectos críticos que envolvem documentos governamentais oficiais: *segmentação* e *classificação* semântica. O primeiro consiste em separar o texto de um Diário Oficial (DO) em trechos que nos permitam identificar o ente federado associado, o título e o conteúdo dos atos governamentais ali publicados, bem como o(s) responsável(eis) por sua criação. Entretanto, esse não é um problema trivial, visto que cada documento possui uma estrutura de apresentação específica que inclui diferentes componentes gráficos (e.g., separadores visuais). Em relação à classificação, o intuito é prever corretamente a qual classe uma nova observação pertence dado um conjunto pré-determinado de categorias como, por exemplo, se um trecho extraído pertence à classe *Leis* ou *Licitações*.

Para tratar tais questões, propomos segmentar documentos no formato PDF utilizando uma heurística orientada a estrutura que visa identificar elementos textuais que compõem um mesmo trecho a partir do distanciamento entre eles. Para a classificação semântica dos trechos, empregamos uma estratégia de aprendizado ativo que permite repetidas interações entre o classificador e o anotador, selecionando um número reduzido das instâncias mais úteis para o treino. Assim, as nossas principais contribuições neste artigo são:

1. Proposta de uma heurística baseada em aspectos estruturais para extração de trechos de DO's;
2. Uma estratégia para classificação semântica dos trechos extraídos usando aprendizado de máquina ativo e classificadores *transformers* do estado-da-arte;
3. Protótipo de uma ferramenta de anotação desenvolvida como um serviço Web integrado ao processo de classificação proposto.

[1] `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm`.

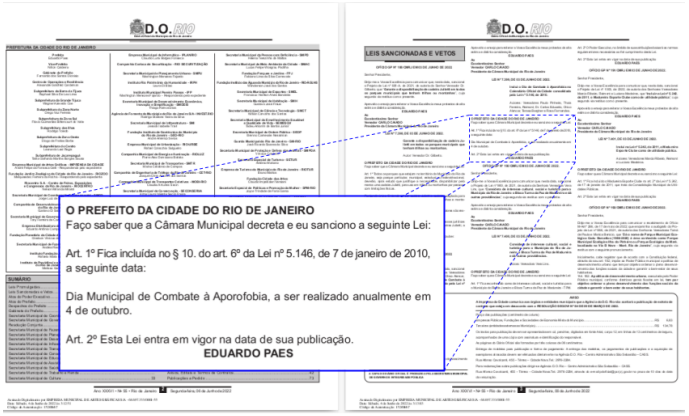

Figura 1. Trecho contido em uma edição do Diário Oficial do Município do Rio de Janeiro, páginas 2 e 3.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta uma visão geral do problema. As Seções 3 e 4 descrevem, respectivamente, as técnicas de segmentação e classificação adotadas. A seguir, a Seção 5 apresenta os resultados experimentais obtidos. Por fim, a Seção 6 apresenta algumas conclusões sumarizando os principais resultados obtidos e sugere alguns trabalhos futuros.

2. Visão Geral do Problema

Este artigo aborda os problemas de segmentação e classificação de informações contidas em diários oficiais. A tarefa de segmentação é o processo de separar trechos em blocos úteis, como sentenças, parágrafos ou seções [Pak and Teh 2018]. Conforme ilustrado na Figura 1, no contexto de diários oficiais, nosso interesse é extrair trechos contendo o ente federado, título/subtítulo, corpo e assinaturas. Já a tarefa de classificação busca prever a qual categoria uma observação pertence, dado um conjunto de categorias pré-existentes [Cunha et al. 2021]. Por exemplo, se parte do conteúdo de um trecho extraído contiver *"...RESOLVE: ART.1º - EXONERAR a servidora..."*, pode ser atribuída a ele a classe correspondente a *Movimentação de Pessoal*, enquanto um subtítulo de trecho iniciando com *"...Regula o acesso a informações previsto no inciso..."* pode ser considerado pertencente à classe de *Leis*.

No contexto do problema de segmentação e classificação de informação envolvendo grandes volumes de dados advindos da administração pública, alguns trabalhos têm procurado enfatizar que as etapas de tratamento e disponibilização sejam consideradas como pré-requisitos para a análise de dados mais complexos. Por exemplo, Pinto et al. [2021] realizaram a extração textual em diários oficiais por meio de expressões regulares e, como resultado, construíram uma base de conhecimento de acordo com uma gramática definida especificamente para atos de movimentação de pessoal da Prefeitura do Rio de Janeiro. Já Rangel et al. [2020] aplicaram técnicas de aprendizado de máquina supervisionado para inferir as categorias (e.g., *saúde* ou *finanças*) de documentos disponíveis em portais de dados governamentais, enquanto Pereira et al. [2021] discutem o problema da falta de padronização para designar categorias das ofertas de serviços públicos e propõem uma taxonomia para melhor categorizar os dados envolvidos. Por fim, vale ressaltar uma ferramenta para anotação e classificação de documentos proposta por Inuzuka et al. [2020] em parceria com uma empresa privada, na qual uma técnica de

aprendizado ativo é empregada para classificar se a informação contida em um trecho do Diário Oficial é ou não de teor jurídico.

Conforme já mencionado, o trabalho apresentado neste artigo contribui para a solução do problema exposto acima, destacando-se por ser uma parceria entre o Departamento de Ciência da Computação da Universidade Federal de Minas Gerais e o Ministério Público do Estado de Minas Gerais (MPMG) para realizar análises em grandes repositórios de dados públicos com a finalidade de caracterizar despesas públicas para apoiar investigações complexas. O artigo trata da gestão e classificação de dados e, em particular, lida com a extração de informação textual e sua classificação semântica. Sob uma perspectiva tecnológica, o resultado deste trabalho visa construir e disponibilizar ferramentas *open source*, democratizando o acesso eficiente aos dados públicos[2].

O MPMG definiu o escopo do trabalho sobre quais dados públicos seriam segmentados e classificados, selecionando os casos prioritários de acordo com seus trâmites internos. Com o intuito de omitir quais instâncias do poder são prioritárias, aqui limitamos ao considerar um repositório com uma coleção genérica contendo inúmeros entes federados. Precisamente, nos foi fornecido o repositório oficial do ano de 2020 com todos os Diários Oficiais da Associação Mineira dos Municípios[3], totalizando 1.640 documentos em formato PDF. Esse repositório possui a filiação de centenas de entes, resultando em trechos bem diversificados como um cenário interessante para avaliarmos nosso processo de classificação semântica.

3. Processo de Segmentação

Segmentação é o processo de dividir um documento em unidades que correspondam a uma mesma ação ou tópico [Pak and Teh 2018]. O problema não é trivial, visto que a maioria dos documentos trata de diversos assuntos elaborados por diversas instâncias dentro de uma mesma instituição. Neste sentido, o primeiro passo para realizarmos a segmentação é pré-processar os documentos com a finalidade de criar uma pseudo-estrutura que permita, posteriormente, identificar e extrair informações relevantes.

O formato mais comum para armazenar e publicar atos públicos em diários oficiais é o PDF. De modo geral, os trechos a serem extraídos estão bem posicionados no *layout* de apresentação. Além disso, a estrutura interna de um documento PDF informa as coordenadas em termos dos eixos X e Y para cada um de seus elementos, sejam eles caracteres ou figuras embutidas. Mais precisamente, devido à natureza não estruturada do formato PDF, precisamos identificar explicitamente quais conjuntos de caracteres formam uma palavra, quais conjuntos de palavras formam uma frase, quais conjuntos de frases formam um parágrafo e assim por diante, até definirmos todos os elementos que compõem um único ato.

Diante desse cenário, uma estratégia para extrair informações de um determinado trecho é explorar as distâncias entre os elementos textuais a fim de identificar quais são as margens que delimitam todos os conteúdos pertencentes a ele. A estratégia consiste em inicialmente converter[4] o documento PDF em um conjunto de elementos textuais e

[2]O estudo completo, com os códigos-fontes de segmentação e classificação, está disponível em: **https://github.com/MPMG-DCC-UFMG/M02**.

[3]https://www.diariomunicipal.com.br/amm-mg/

[4]Dentre as ferramentas testadas, *PDFMINER.SIX* (https://github.com/pdfminer/

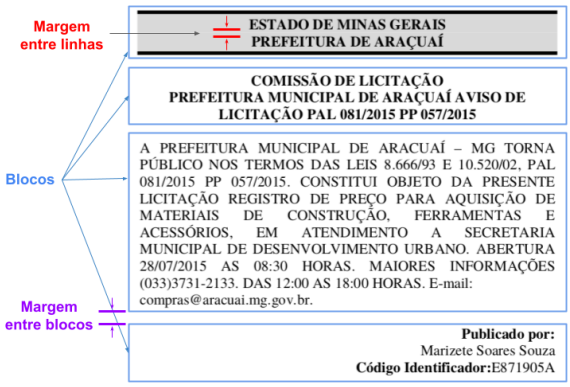

Figura 2. Trecho de um ato composto por blocos definidos de acordo com as margens entre subelementos que os compõem.

suas respectivas coordenadas X e Y. Nossos testes iniciais indicaram que há documentos que não respeitam a ordem de aparição de seus elementos na apresentação, sendo assim necessário considerar o pior caso em que todos os elementos precisam ser ordenados de acordo com as suas coordenadas. Uma vez que tais elementos estejam devidamente ordenados, definimos os tipos de elemento como sendo *caractere*, *palavra*, *linha* e *bloco*, onde um bloco é formado por linhas, uma linha é formada por palavras e uma palavra é formada por caracteres. Para cada atribuição de tipo a cada elemento, há limiares aceitáveis de distanciamento horizontal e vertical para decidir quais tipos de elemento serão agrupados ou mantidos disjuntos.

A Figura 2 exemplifica esse processo. No caso, o objetivo é determinar se as duas linhas *ESTADO DE MINAS GERAIS* e *PREFEITURA DE ARAÇUAÍ* fazem ou não parte de um mesmo bloco. Para isso, determina-se um limiar aceitável para a margem entre elas (destacado em vermelho). Similarmente, o corpo e a assinatura do trecho são duas partes distintas devido à distância definida para a margem entre os blocos (destacado em roxo). Agora que os blocos estão estruturados com suas partes bem definidas, cada segmento contido em um bloco corresponde às informações que desejamos extrair.

Além disso, há de se tratar a questão de que um trecho pode ser disposto em mais de uma página, bem como considerar que possa haver separação dos textos em colunas em uma mesma página. Tais considerações são resolvidas durante o processo inicial de ordenação, com uma adaptação para inspecionar a quantidade de colunas para ajustar as posições relativas de cada trecho como um fluxo único. Por fim, há ainda uma fase que busca descartar possíveis figuras embutidas (e.g., propagandas ou logotipos) que por ventura possam aparecer de forma inesperada e, assim, reposicionar as coordenadas relativas dos elementos textuais.

4. Classificação Semântica

Após a segmentação dos diários oficiais em trechos devidamente extraídos, a próxima etapa é classificá-los de acordo com a semântica contida em cada um dos atos. Essa classificação é importante para se realizar buscas mais bem elaboradas para aplicações finalísticas de análise de dados tais como detecção de fraudes em licitações e em editais.

`pdfminer.six`) e *pdftotext* (GNU/Linux) foram as ferramentas de melhor desempenho para a obtenção dos elementos textuais e suas respectivas posições relativas.

A maior dificuldade da etapa de classificação corresponde ao gigantesco volume de trechos sem quaisquer categorização ou rotulações, requerendo que haja intervenção manual para realizar anotações. Especificamente, há dois subproblemas inerentes ao processo de classificação a partir do repositório com trechos puros (i.e., sem metadados): (i) lidar com o problema de modelagem de tópico, i.e., um modelo estatístico para descobrir quais são as categorias/tópicos que ocorrem em nosso repositório de trechos textuais; (ii) definir qual é a técnica de aprendizado de máquina ideal para aprender como classificar os trechos. Essas duas subetapas são discutidas a seguir e, em seguida, um protótipo de classificação semântica como serviço Web é apresentado.

4.1. Modelagem de Tópicos

A identificação de tópicos semânticos é um tema relacionado ao Processamento de Linguagem Natural que consiste na criação de um modelo estatístico para descobrir tópicos abstratos em uma coleção de documentos [Blei 2012]. Mais especificamente, a modelagem de tópicos é uma técnica de aprendizado de máquina capaz de detectar padrões de palavras e frases em cada um dos trechos para agrupá-los automaticamente de acordo com a similaridade de suas características.

Em nosso contexto da administração pública, a modelagem consiste em definir os tópicos semânticos a partir da distribuição discriminativa dos termos que compõem cada grupo de trechos, com o objetivo de encontrar os tópicos que sejam evidentes, disjuntos e significativos. Para esse fim, aplicou-se a técnica[5] *Latent Dirichlet Allocation* [Blei et al. 2003], que assume uma distribuição de probabilidade de *Dirichlet* sobre dados textuais para estimar as probabilidades de palavras para cada grupo. Em seguida, realizou-se uma análise exploratória, variando o número de tópicos entre 4 e 16, para definir o número mais apropriado de grupos e definir a cada um deles um tópico semântico. Por fim, foram selecionados manualmente quatro grandes tópicos semânticos com o aval de especialistas em documentos oficiais: *Pessoal*, *Lei*, *Licitação* e *Orçamento*.

4.2. Processo de Classificação

Resumidamente, o problema de classificação visa predizer corretamente a qual classe uma nova observação pertence, dentro das possibilidades existentes. No contexto de aprendizado de máquina, a técnica de classificação é um método de aprendizado supervisionado, visto que, para a atribuição correta de novas observações à sua respectiva classe, é necessário um conjunto inicial de observações já categorizadas chamado de *conjunto de treino* [Cunha et al. 2021]. Após o treinamento do classificador, espera-se que, para qualquer nova observação, a classificação seja realizada automaticamente pelo modelo. O problema admite diversas estratégias de modelagem com graus diferentes de complexidade como classificadores lineares, *Support Vector Machines* (SVM), árvores de decisão, redes neurais e outros.

Considerando o cenário da administração pública em que se deseja classificar um grande volume de dados que, inicialmente, não possui rótulos associados às suas respectivas classes, a estratégia de aprendizado ativo se mostra adequada. Essa técnica consiste

[5]Ressalta-se que também foram testadas outras abordagens de modelagem de tópicos específica para textos curtos como BTM (*Biterm Topic Models for Short Text*) e *clustering* (k-means) [Cunha et al. 2021]. Porém, os grupos foram melhor definidos pelo LDA.

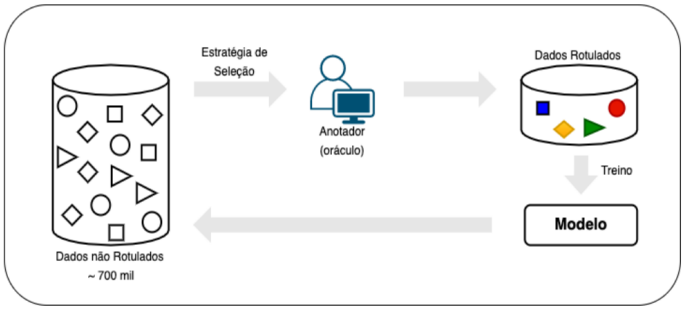

Figura 3. Processo de classificação dos trechos usando aprendizado ativo.

em repetidas interações entre o classificador e um oráculo (usuário que será consultado para classificar observações específicas selecionadas pelo modelo), possibilitando que o treinamento do modelo utilize um número reduzido de observações que, ao considerar um processo de seleção bem feito, permite um aprendizado significativo por parte do classificador com um baixo custo de rotulação que nesse domínio é caro, considerando a necessidade do envolvimento de espacialistas de domínio [Lewis and Catlett 1994].

A seleção das observações que serão consultadas ao oráculo pode ser realizada de maneiras distintas. Uma opção seria consultar o oráculo sobre as observações que se encontram próximas a fronteiras entre classes, pois o modelo encontraria mais dificuldades em classificar essas observações. No entanto, pode ser interessante combinar esses casos com instâncias que o modelo tem certa confiança para classificar, pois, dessa maneira, será possível consolidar as classes já determinadas pelo classificador.

A escolha de uma quantidade reduzida de amostras mais significativas para o aprendizado do classificador (amostra mais útil) depende das estratégias adotadas. Essas estratégias são baseadas na incerteza de classificação [Lewis and Catlett 1994] (*classification uncertainty*), portanto, são chamadas de medidas de incerteza. Na abordagem adotada, foram usadas três medidas integradas: incerteza de classificação (*classification uncertainty*), fronteira de classificação (*classification margin*) e entropia de classificação (*classification entropy*). Incerteza de classificação é a medida mais simples. Ao consultar os rótulos com base nessa medida de classificação, a estratégia seleciona a amostra com a maior incerteza. No caso da medida referente à fronteira de classificação, ao consultar os rótulos, a estratégia seleciona a amostra com a menor fronteira, pois, quanto menor a fronteira (margem) de decisão, mais incerta é a decisão. Por fim, em relação à medida de entropia de classificação, em termos de heurística, ela é proporcional ao número médio de suposições que alguém deve fazer para encontrar a classe verdadeira. Quanto mais uniforme for a distribuição, maior será a entropia. Portanto, a amostra mais incerta é aquela com elevado valor de entropia.

A Figura 3 ilustra o processo de rotulação e classificação dos segmentos dos diários oficiais. Por meio de uma estratégia de seleção previamente definida (baseada nas amostras de incerteza, por margem e por entropia), a ferramenta seleciona um conjunto de dados não-rotulados do repositório que contém aproximadamente 700 mil segmentos. Esse conjunto selecionado é apresentado ao anotador (o oráculo), o qual é responsável por identificar e rotular as informações do segmento, formando, assim, o conjunto de treinamento. O módulo de aprendizado supervisionado recebe o conjunto de treinamento, onde

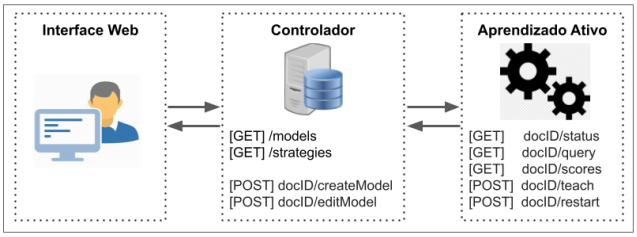

Figura 4. Arquitetura do serviço Web integrando a anotação de instâncias e aprendizado ativo.

cada exemplo, representado por suas características ou atributos, foi rotulado de acordo com a classe à qual pertence. Como resultado, o sistema de aprendizado deve construir e atualizar o modelo, que permite predizer classes para a novas entradas, diferentes dos exemplos rotulados previamente, reiniciando o ciclo de aprendizado de máquina.

Os resultados iniciais demonstraram que as classes são naturalmente desbalanceadas, motivando o emprego de duas estratégias para tratar essa questão: (i) priorizar as classes minoritárias na estratégia de seleção de segmentos a serem rotulados e (ii) lidar com o desbalanceamento por redução do número de observações da classe majoritária (i.e., *undersampling*) diretamente na fase de treino do modelo.

4.3. Aprendizado Ativo como Serviço Web

Embora haja boas ferramentas *open source* para anotação de dados, como Brat[6] e Doccano[7], elas não suportam técnicas de aprendizado de máquina que requerem interatividade com o anotador. De fato, as melhores alternativas são comerciais e de código fechado, com destaque para a ferramenta Prodigy[8]. Diante disso, propomos um protótipo de anotação integrado com aprendizado ativo por meio de um serviço Web. Tecnicamente, a proposta da solução segue um padrão similar das alternativas de código fechado, o que tem como principal vantagem uma integração fácil com os modelos preditivos, seja para aplicações locais, em rede e até para dispositivos móveis.

Nosso protótipo contém três módulos, conforme ilustrado na Figura 4, separados em Interface Web (*front-end*), Controlador (*middleware*) e Aprendizado Ativo (*back-end*). De forma geral, a Interface Web permite que os usuários realizem anotações interativas, enquanto o Controlador executa um conjunto de requisições GET/POST (pré-definidas) advindas da Interface Web ao módulo Aprendizado Ativo. Além disso, cada módulo é independente, ou seja, novas implementações de modelos preditivos e estratégias de seleção podem ser acopladas de forma transparente ao anotador.

A Figura 5 exemplifica a interface de usuário para o processo de anotação interativo integrado ao modelo de aprendizado ativo. À medida que o anotador decide qual é o rótulo mais apropriado para uma determinada instância, o gráfico mostra o progresso do modelo. Além disso, destaca-se que a implementação permite múltiplos usuários anotando um mesmo conjunto de instâncias ao mesmo tempo. Portanto, o processamento preditivo é todo realizado em tempo real e pode ser finalizado assim que haja satisfação com a acurácia alcançada.

[6]Disponível em: `https://brat.nlplab.org/`.
[7]Disponível em: `https://doccano.github.io/doccano/`.
[8]Disponível em: `https://prodi.gy/`.

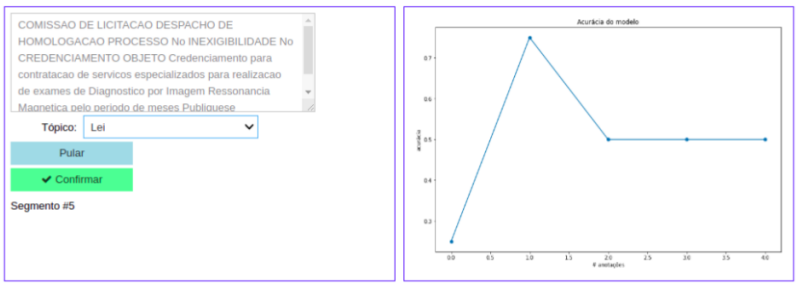

Figura 5. Interface com a interação do anotador com o serviço Web.

Tabela 1. Estatísticas dos dados presentes nos 1.640 documentos.

	# Entidades por documento	# Segmentos por documento
Média	96,9	426,7
Mediana	92	416,5
Desvio Padrão	39,5	190

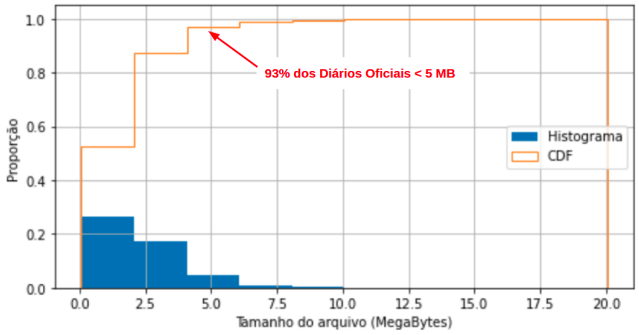

Figura 6. Distribuição dos tamanhos dos documentos.

5. Avaliação Experimental

5.1. Avaliação Experimental da Segmentação de Trechos

Como entrada para o processo de segmentação, foram considerados 1.640 documentos em formato PDF (ver Seção 2). Como saída, o processo de segmentação extraiu 645 Entidades Federadas e um total de 699.759 trechos. Também verificou-se que um diário oficial pode ter várias Entidades Federadas e, por sua vez, cada uma delas pode ter vários trechos (i.e., atos sob sua responsabilidade). A Tabela 1 mostra as estatística básicas do número de entidades e o número de entes federados extraídos por documento.

Dado que parte do processo de segmentação consiste na ordenação dos elementos textuais de acordo com suas coordenadas em cada página (com complexidade igual a $O(n \log n)$), a avaliação do tempo total de processamento de casos reais é de suma importância. A Figura 6 apresenta a distribuição dos tamanhos dos documentos, onde o eixo Y do gráfico se refere à proporção de arquivos em formato PDF e o eixo X aos tamanhos dos arquivos. Observa-se que 93% de todos os documentos têm tamanho inferior a 5 MB. O processamento do maior documento (20,1 MB com 667 páginas) durou 12 minutos, em uma máquina Intel Duo CPU 2.6 GHz com 4 G de memória RAM, o que é bem aceitável dado a frequência diária de publicação dos documentos por cada ente federado.

A qualidade das extrações foi manualmente avaliada por meio de 5.219 amostras aleatoriamente selecionadas através do protótipo de anotação (i.e., informa-se se o seg-

Tabela 2. Proporção das classes semânticas das amostras de Diário Oficiais.

Classe	Proporção
Licitação	62,45%
Pessoal	24,90%
Orçamento	6,88%
Lei	5,78%

mento foi ou não devidamente coletado), sendo verificado que os segmentos foram corretamente extraídos (i.e., acurácia de 100%). Por outro lado, há pequenas imperfeições, como elementos advindos de tabelas ligeiramente fora de suas posições. Contudo, o conteúdo textual dos segmentos foi totalmente preservado, de modo a ser utilizado em etapas futuras de classificação e busca textual. Portanto, a estratégia baseada em estrutura gera um padrão bem determinado de acordo com as distâncias entre os elementos textuais. Em especial, a estratégia se mostrou uma escolha técnica adequada para lidar com possíveis formatações fora de ordem, que foi um problema recorrente durante nossos testes.

5.2. Avaliação Experimental da Classificação Semântica

Nesta subseção, apresentamos a avaliação experimental da classificação semântica com o uso de aprendizado ativo, cujo processo de anotação foi realizado seguindo a proposta definida na Seção 4. Especificamente, selecionamos como modelos preditivos quatro classificadores do estado-da-arte: (i) **SVM**, amplamente usado para a classificação de texto, tendo produzido os melhores resultados num *benchmark* recente [Cunha et al. 2021] em cenários similares aos aqui descritos, ou seja, coleções pequenas ou medianas, com poucos dados rotulados e alto desbalanceamento de classes; (ii) **BERT** [Devlin et al. 2019] que tem sido o estado-da-arte em inúmeras tarefas de PLN [Cunha et al. 2021, Garg et al. 2020]; (iii) **BERTimbau** [Souza et al. 2020] que estende o BERT para reconhecer sentenças em português brasileiro [Santos et al. 2019, Rodrigues et al. 2019]; e (iv) **LaBSE** [Feng et al. 2022] que explora o uso de modelos de linguagem capazes de representar sentenças em vários idiomas. Já para o algoritmo de seleção de instâncias úteis, a estratégia *uncertainty* [Lewis and Catlett 1994] obteve os melhores resultados. Tais algoritmos estão de acordo com outras iniciativas de classificação de textos extraídos de diários oficiais [Inuzuka et al. 2020, Rangel et al. 2020].

Conforme já mencionado, há aproximadamente 700 mil trechos textuais não rotulados que precisam ser devidamente classificados. Para isso, o processo de aprendizado ativo requer um conjunto inicial para tomar as suas primeiras decisões. Nesse sentido, foram selecionados aleatoriamente 29 exemplos iniciais, separando um conjunto inicial com 10 instâncias e o conjunto de teste com os 19 restantes. Em seguida, foram realizadas as anotações de acordo com a estratégia que seleciona o trecho textual em que há uma maior incerteza sobre ele, ou seja, o exemplo mais útil para ser rotulado. Dessa forma, foi rotulado um total de 727 trechos textuais (equivalente a 0.1% do repositório). Como resultado, observamos que a natureza do problema é altamente desbalanceada, com o assunto relacionado a *Licitação* sendo bem mais representativo (62,45%) do que o assunto referente a *Orçamento* (5,78%), conforme a Tabela 2.

Para fins de avaliação da tarefa de classificação semântica, consideramos a *acurácia* para medir a efetividade global em relação a todas as decisões tomadas pelo classificador, a *precisão* que informa quantas classes estão corretas dentre as que foram

associadas a uma determinada classe, a *revocação* que indica quais classes foram corretamente associadas em relação ao total de instâncias de uma determinada classe e a medida *F1* que corresponde à média harmônica entre a precisão e a revocação [Cunha et al. 2021]. No caso da medida F1, apresentamos os resultados da variante Macro(F1) que indica as suas médias por classe, sendo assim mais adequada para problemas altamente desbalanceados.

A avaliação da robustez da classificação semântica foi realizada por validação cruzada (*cross-validation*), que é uma técnica experimental que explora o treinamento de vários modelos em subconjuntos distintos de dados de entrada e a avaliação deles no subconjunto complementar dos dados [Cunha et al. 2021]. Um dos principais objetivos da validação cruzada é verificar a generalização dos modelos para diferentes conjuntos de treino e teste. Para a escolha dos melhores parâmetros dos classificadores utilizamos uma validação cruzada dentro do treino (*nested cross-validation*). Por fim, dado que as classes rotuladas são desbalanceadas, o processo de aprendizagem empregou a técnica de *undersampling* (apenas no treinamento), que equilibra o número de classes desiguais ao manter todas as instâncias da classe minoritária e diminuir o tamanho da classe majoritária.

A Tabela 3 mostra a média dos resultados do *fold* de teste do processo de validação cruzada com *5-fold*. Como pode ser visto, a acurácia de todos os classificadores é similar (estatisticamente) e bastante efetiva (cerca de 85%). Por outro lado, os resultados de MacroF1, que é a medida mais adequada em problemas com desbalanceamento, do Bertimbau e do LaBSE são bastante superiores aos do SVM (12% de ganho) e aos do BERT (5.6% superiores). Particularmente, a precisão do LabSE é bastante alta, cerca de 84%.

A Tabela 4 mostra a Precisão, a Revocação e a Medida F1 por classe. Como esperado, a classe majoritária possui os valores mais altos para todas as métricas (entre 91% e 94%) para todos os classificadores, com os maiores valores obtidos pelo SVM para Licitações. Apesar de um valor de MacroF1 médio um pouco inferior, o SVM é o classificador que consegue a performance mais "equilibrada" para as quatro classes, principalmente para as minoritárias. Uma explicação para este caso é o SVM se beneficiar da representação dos dados (i.e., baseada em TF-IDF), uma vez que ele lida muito bem com representações esparsas. Por outro lado, a efetividade do LaBSe, em três de quatro classes fica ao redor de 90%. Assim, dependendo da aplicação e da importância da classe *Lei*, pode-se escolher entre um desses classificadores.

Sobre os erros de classificação, observamos dois tipos principais de erro: *erro de anotação* e *erro de predição* do modelo. Para o primeiro caso, podemos ilustrá-lo com o trecho *"Termo de aditamento ao Contrato [anônimo], Processo [anônimo], para aquisição de ..."*, com o título *"PORTARIAS/LEIS"*, que foi rotulado manualmente como *Lei*, mas é um caso específico de *Licitação*. Nesse caso, há informação no título do trecho que é relacionada a categorias às quais ele se enquadraria (i.e., portaria ou lei), induzindo o erro do anotador. Porém, observando atentamente o trecho, trata-se de uma portaria para designar aditamento em um contrato de licitação.

Para o segundo caso, o modelo SVM predisse erroneamente como sendo da classe *Pessoal* o fragmento de trecho *"PORTARIA [anônimo] REGULAMENTA A LEI COMPLEMENTAR [anônimo] QUE INSTITUI O QUADRO PESSOAL DA ..."*, o qual corresponde realmente a uma regulamentação de *Lei*. De fato, este exemplo de erro ilustra a

Tabela 3. Efetividade dos modelos SVM, BERT, BERTimbau e LaBSE utilizando validação cruzada com 5-fold

	SVM	BERT	BERTimbau	LaBSE
Acurácia	0,85 ± 0,04	0,84 ± 0,02	0,86 ± 0,01	0,86 ± 0,02
Macro F1	0,67 ± 0,06	0,71 ± 0,05	0,75 ± 0,05	0,75 ± 0,03
Precisão	0,74 ± 0,06	0,74 ± 0,07	0,78 ± 0,08	0,84 ± 0,06
Revocação	0,67 ± 0,07	0,70 ± 0,05	0,73 ± 0,03	0,75 ± 0,03

Tabela 4. Precisão (P), revocação (R) e F1-score (F1) dos classificadores por classe.

	SVM			BERT			BERTimbau			LaBSE		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Lei	0,60	0,75	0,67	0,36	0,18	0,24	0,52	0,23	0,32	0,50	0,27	0,35
Licitação	0,94	0,94	0,94	0,88	0,96	0,91	0,88	0,95	0,91	0,88	0,94	0,91
Orçamento	0,71	0,83	0,77	0,92	0,83	0,87	0,92	0,83	0,87	0,92	0,88	0,90
Pessoal	0,83	0,68	0,75	0,81	0,82	0,82	0,88	0,92	0,90	0,87	0,90	0,88

dificuldade de se classificar trechos que possuem ambiguidade até para uma avaliação humana. No caso, o termo "quadro pessoal" foi considerado pelo algoritmo como fortemente associado à classe *Pessoal*.

6. Conclusões e Trabalhos Futuros

Neste artigo foi abordado um problema real em termos de pré-processamento e organização de documentos públicos a partir de uma parceria estabelecida entre o Departamento de Ciência da Computação da Universidade Federal de Minas Gerais e o Ministério Público do Estado de Minas Gerais. Mais especificamente, o artigo abordou dois aspectos críticos que envolvem o processamento de diários oficiais: o processo de segmentação, ou seja, a extração de trechos textuais, e o processo de classificação semântica dos trechos extraídos. Como solução, propusemos uma heurística orientada à estrutura para segmentar os trechos de diários oficiais. Posteriormente, propusemos classificar semanticamente os trechos extraídos com uma estratégia de aprendizado ativo que minimiza o esforço manual de rotulação. Por fim, implementamos como um serviço Web a rotulação de dados integrado ao processo de aprendizado ativo. Nossa avaliação experimental consistiu em processar 1.640 documentos, assim extraindo 645 entidades federadas distintas com um total de 699.759 trechos. Em relação à classificação semântica, com muito pouco esforço de rotulação e adotando técnicas para lidar com o desbalanceamento natural das classes semânticas, nossas avaliações por validação cruzada indicaram um valor para *MacroF1* de 75% e uma acurácia global de 85%. Como trabalho futuro, pretendemos expandir a fonte de documentos governamentais para além do escopo determinado pela parceria com o órgão investigativo e tratar casos particularmente difíceis de segmentação que possuem imagens embutidas (e.g., documentos digitalizados). Também pretendemos aplicar o aprendizado ativo para aplicações finalísticas como, por exemplo, para decidir se há ou não indícios de fraude em trechos de documentos oficiais.

Agradecimentos

Gostaríamos de agradecer o apoio do Ministério Público de Minas Gerais, por meio do projeto Capacidades Analíticas, bem como à CAPES, CNPq, FAPEMIG e Fundação Araucária pelo apoio individual aos autores.

Referências

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S. D., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., Rosa, T., Rocha, L., and Gonçalves, M. A. (2021). On the Cost-Effectiveness of Neural and Non-Neural Approaches and Representations for Text Classification: A Comprehensive Comparative Study. *Inf. Process. Manag.*, 58(3):102481.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 878–891. Association for Computational Linguistics.

Garg, S., Vu, T., and Moschitti, A. (2020). TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 7780–7788. AAAI Press.

Inuzuka, M., do Nascimento, H., Almeida, F., Barros, B., and Jradi, W. (2020). Doclass: open-source software to support document labeling and classification. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 105–112. SBC.

Lewis, D. D. and Catlett, J. (1994). Heterogeneous Uncertainty Sampling for Supervised Learning. In *Machine Learning Proceedings 1994*, pages 148–156. Elsevier.

Pak, I. and Teh, P. L. (2018). Text Segmentation Techniques: A Critical Review. *Innovative Computing, Optimization and Its Applications*, pages 167–181.

Pereira, G. C., Monteiro, I. T., Vasconcelos, D. R., Braz, L., and Silva, C. H. (2021). Classificação taxonômica de categorias de serviços públicos para aplicações digitais. In *Anais do IX Workshop de Computação Aplicada em Governo Eletrônico*, pages 119–130. SBC.

Pinto, F. A. D., Haeusler, E. H., and Lifschitz, S. (2021). Transparência pública automatizada a partir da gramática do diário oficial. In *Anais do IX Workshop de Computação Aplicada em Governo Eletrônico*, pages 59–70. SBC.

Rangel, M., Bernardini, F., Viterbo, J., Monteiro, R., Seixas, E., and dos Santos Pinto, H. (2020). Uso de Aprendizado de Máquina para Categorização Automática de Conjuntos de Dados de Portais de Dados Abertos. In *Anais do VIII Workshop de Computação Aplicada em Governo Eletrônico*, pages 120–131. SBC.

Rodrigues, R., da Silva, J., Castro, P., Félix, N., and Soares, A. (2019). Multilingual Transformer Ensembles for Portuguese Natural Language Tasks. In *Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symposium in Information and Human Language Technology (STIL 2019)*, pages 27–38. CEUR-WS.org.

Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., and Vieira, R. (2019). Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 437–442. IEEE.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems, (BRACIS)*, pages 403–417. Springer.