LiPSet: Um conjunto de Dados com Documentos Rotulados de Licitações Públicas

Mariana O. Silva[1], Amanda F. Paula[1], Gabriel P. Oliveira[1], Iago A. D. Vaz[1], Henrique Hott[1], Larissa D. Gomide[1], Arthur P. G. Reis[1], Bárbara M. A. Mendes[1], Clara A. Bacha[1], Lucas L. Costa[1], Michele A. Brandão[1,2], Anisio Lacerda[1], Gisele L. Pappa[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG, Brasil

[2]Instituto Federal de Minas Gerais (IFMG) – Ribeirão das Neves, MG, Brasil

{mariana.santos,amanda.fagundes,gabrielpoliveira,iagoadvaz}@dcc.ufmg.br
{henriquehott,larissa.gomide}@dcc.ufmg.br, {arthurpetrocchi,barbaramit}@ufmg.br
{clarabacha,lucas-lage}@ufmg.br, {michele.brandao,anisio,glpappa}@dcc.ufmg.br

Abstract. *In this work, we present LiPSet, a dataset with labeled documents from public bids from Minas Gerais. After an overview of the manual collection and labeling process, we present a brief exploratory data analysis to summarize the main features and contributions of the proposed dataset. In addition, we discuss potential applications and main challenges involving the use of LiPSet.*

Resumo. *Neste trabalho, é apresentado o LiPSet, um conjunto de dados com documentos rotulados de licitações públicas de Minas Gerais. Após uma visão geral do processo de coleta e rotulação manual, uma breve análise exploratória de dados é apresentada para resumir as principais características e contribuições do conjunto de dados proposto. Além disso, são discutidas potenciais aplicações e principais desafios que envolvem o uso do LiPSet.*

1. Introdução

Com a Lei de Acesso às Informação (Lei n.º 12.527, sancionada em 18 de novembro de 2011),[1] os cidadãos passaram a ter acesso a informações públicas dos três poderes da União, dos estados, do Distrito Federal e dos municípios. Apesar dessas informações serem disponibilizadas em formatos diferentes de arquivos, em geral, sem ou com pouca padronização, elas são importantes para muitas aplicações [Muniz and Lóscio 2018, Mata et al. 2019, Shimron et al. 2022]. Por exemplo, Pereira [2022] investiga como dados abertos governamentais sobre a educação brasileira são utilizados por uma determinada comunidade, até mesmo para a definição de projetos. Por outro lado, Costa *et al.* [2022] utilizam dados de licitações públicas para identificar a ocorrência de possíveis fraudes nessas licitações.

Especificamente, dados abertos sobre licitações públicas envolvem diferentes tipos de documentos, incluindo o texto de editais, erratas, atas, contratos, adjudicação e homologação, dentre outros. Cada um desses documentos possui um formato específico, trazendo informações a respeito das diferentes etapas de um processo licitatório, desde sua

[1]Sobre a Lei de Acesso à Informação: https://www.gov.br/capes/pt-br/acesso-a-informacao/servico-de-informacao-ao-cidadao/sobre-a-lei-de-acesso-a-informacao

publicação até sua homologação. A análise e manipulação desses documentos depende do processamento de dados não-estruturados, especificamente, em formato de texto. Isso requer o uso de técnicas específicas para lidar com esse tipo de dado, por exemplo, da área de Processamento Natural de Linguagem (PLN). PLN é uma área de pesquisa ampla que inclui o desenvolvimento de modelos computacionais para solucionar problemas que dependem de informações expressas em linguagem natural [Meera and Geerthik 2022].

A coleta, análise e processamento desses documentos é um desafio para os humanos (por depender de ferramentas capazes de gerenciar grandes volumes de documentos, além de precisar analisá-los para ampliar a compreensão do tipo de documento sendo consultado) e para as máquinas (por ter que ser capaz de automatizar as tarefas de coleta, análise e processamento). Este trabalho representa um passo em direção à superação desses desafios ao apresentar LiPSet, um conjunto de dados formado a partir da coleta, processamento e rotulação de documentos de licitações públicas de Minas Gerais.

Assim, este artigo está organizado conforme segue. A Seção 2 descreve os trabalhos relacionados. A Seção 3 descreve as características do conjunto de dados LiPSet. A Seção 4 caracteriza o LiPSet e a Seção 5 apresenta uma aplicação desse conjunto de dados. Já a Seção 6 descreve as principais limitações e desafios do LiPSet. Finalmente, a Seção 7 detalha as considerações finais.

2. Trabalhos Relacionados

Desde a sanção da Lei de Acesso à Informação, vários trabalhos foram publicados visando o uso de dados abertos do governo [Lima et al. 2020, Lyra et al. 2021]. Em geral, dados públicos foram coletados e organizados por meio de diferentes estratégias, por exemplo, armazenando em bases de dados orientadas a grafos [van Erven et al. 2017], ou realizando a rotulação dos dados [Lima et al. 2020]. Em particular, para dados de licitações públicas, diversas pesquisas apresentam resultados promissores para a aplicação desses dados. Gabardo e Lopes [2014], por exemplo, utilizam técnicas de análise de redes sociais para verificar a formação de cartel em empresas de construção civil no estado do Paraná. Similarmente, Silva *et al.* [2020] abordam a aplicação de técnicas de mineração de dados para apoio a processos de auditoria interna do exército.

Apesar do avanço em pesquisas sobre dados abertos, alguns desafios ainda persistem ao lidar com dados das diferentes esferas governamentais no Brasil [Oliveira and Silveira 2018], são eles: falta de padronização, fontes heterogêneas e a necessidade de uso de técnicas de PLN. Nesse sentido, muitos trabalhos visam suprir a demanda por dados com mais qualidade para aumentar a acessibilidade e reusabilidade dos mesmos. Um exemplo é o QualiSuS, um banco de dados relacional criado a partir da extração de dados do Portal DataSUS [Clarindo et al. 2020]. Nele, padrões são adotados, como identificadores de doenças e os dados disponibilizados em formato CSV e JSON. Já na área jurídica, o JusBD apresenta um conjunto de dados não rotulado com a finalidade de auditorias forenses em dados do senso do poder judiciário[Mata et al. 2019]. Finalmente, Araújo e Souza [2011] utilizam um coletor Web para obtenção de dados de políticos brasileiros.

Em geral, lidar com a extração de informações em documentos no formato PDF não é trivial, principalmente, pela falta de padronização [Mata et al. 2019]. Por isso, o LiPSet contribui para a análise de documentos públicos governamentais, ao ser formado

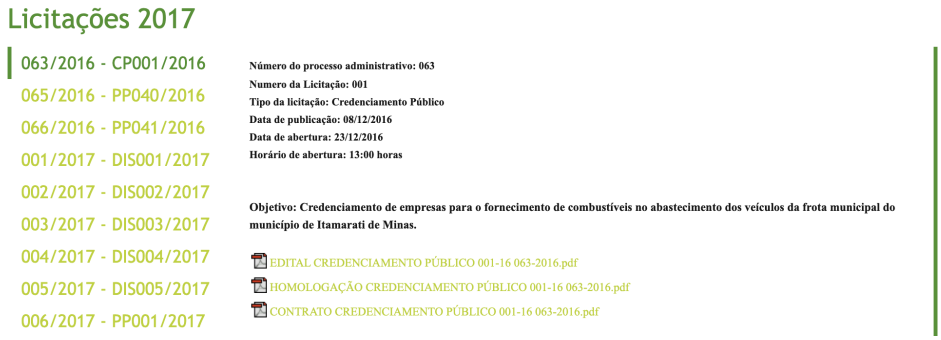

Figura 1. Página do Portal da Transparência do município Itamarati de Minas.

por dados extraídos de arquivos em formato PDF. Além disso, foram utilizadas técnicas de PLN para a extração e pré-processamento do conteúdo textual de tais documentos. É importante destacar que o LiPSet também possui dados rotulados, facilitando o seu como entrada de algoritmos de aprendizado de máquina supervisionados.

3. LiPSet

Esta seção apresenta o LiPSet, um conjunto de dados com documentos rotulados de licitações públicas de Minas Gerais. As Seções 3.1 e 3.2 descrevem os processos de coleta e construção. Em seguida, a Seção 3.3 apresenta como os dados são armazenados e organizados. Por fim, a Seção 3.4 detalha a localização pública para download do LiPSet.

3.1. Coleta dos Dados

Para a coleta de documentos de licitações públicas, as fontes de dados utilizadas foram portais da transparência e/ou de licitações de 18 municípios de Minas Gerais. A primeira etapa do processo de coleta dos dados é o levantamento dos links dos portais que devem ter suas licitações coletadas. Ou seja, são definidos especificamente quais municípios devem ter os documentos de licitações coletados. Em seguida, o portal de cada município é analisado, visto que não há um padrão de disponibilização dos documentos entre a maioria deles. Após essas análises, baseado na estrutura e formatação dos portais, um *Web Crawler* é desenvolvido para acessar cada link disponível na página do portal para baixar os documentos de licitação automaticamente.

Como exemplo, a Figura 1 mostra uma página do Portal da Transparência do município Itamarati de Minas.[2] Na página do portal, observa-se que existem links (em verde) no canto esquerdo nos quais é possível acessar cada documento de licitação. Dessa forma, para cada link existente na página, o *Web Crawler* simula um *click* nesses links para baixar os arquivos desejados. Especificamente, o processo de coleta dos dados ocorreu em dois períodos distintos: julho e dezembro de 2021. Por questões de privacidade do Programa de Capacidades Analíticas em parceria com o Ministério Público de Minas Gerais, não é possível disponibilizar o código do coletor dos portais da transparência.

A Tabela 1 apresenta a quantidade total de arquivos coletados para cada um dos municípios e o mês de coleta. No total, foram coletados 9.761 documentos no formato

[2]Portal da Transparência de Itamarati de Minas: `http://itamaratideminas.mg.gov.br/licitacoes.html`

Tabela 1. Distribuição dos arquivos PDF coletados em julho e dezembro de 2021.

Mês de coleta	#	Município	# Documentos
Julho	1	Arantina	937
	2	Coqueiral	1.528
	3	Cristais	1.737
	4	Ijaci	455
	5	Itamarati de Minas	1.111
	6	Olaria	42
	7	Passa-Vinte	412
	8	Pirapetinga	1.108
	9	Ribeirão Vermelho	686
	10	São Bento Abade	275
Dezembro	1	Bias Fortes	159
	2	Cana Verde	402
	3	Contagem	136
	4	Governador Valadares	93
	5	Palma	93
	6	Pedro Teixeira	179
	7	Rio Preto	279
	8	São Tomé	129
Total	18		**9.761**

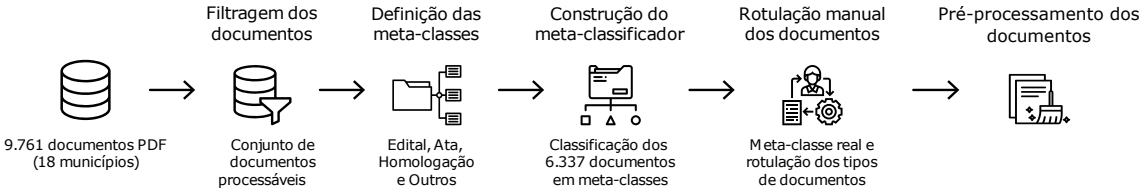

Figura 2. Metodologia utilizada na rotulação de documentos de licitação.

PDF (*Portable Document Format*). Em particular, no portal de cada município, também são disponibilizados arquivos em formatos HTML, DOC e CSV (*Comma-separated values file*). Entretanto, por questões de padronização do processamento, neste trabalho, documentos em formatos diferentes do PDF foram desconsiderados. Além disso, os arquivos de formato HTML e CSV contém informações extraídas diretamente das páginas web que foram visitadas pelo coletor e sua coleta está mais relacionada à estrutura da página em questão do que a seu conteúdo.

3.2. Metodologia para rotulação dos documentos

Nesta seção, apresentamos a metodologia elaborada para a construção do LiPSet, onde são rotulados os documentos de licitações públicas. A Figura 2 ilustra a visão geral da metodologia aplicada e cada etapa é brevemente descrita a seguir.

Filtragem dos documentos. Após coletados, os documentos passam por um processo de filtragem, onde são separados em não-processáveis (documentos escaneados/corrompidos) e processáveis (documentos de onde é possível extrair texto de forma direta). Essa separação é feita utilizando a biblioteca PDFPlumber[3] do Python, que não consegue extrair texto de documentos escaneados, corrompidos ou imagens. Quando nenhum texto é extraído do documento, ele é considerado não-processável. Documentos considerados processáveis seguem para as próximas etapas do fluxo proposto.

[3]PDFPlumber: `https://github.com/jsvine/pdfplumber`

Definição das meta-classes. A definição das meta-classes levou em consideração tipos de documentos considerados essenciais em um processo licitatório. A partir do conhecimento empírico adquirido pela análise dos documentos, chegou-se à proposta presente neste trabalho, resultando em quatro meta-classes, são elas: Ata (abrange todos os documentos de ata disponíveis), Edital (abrange documentos de edital e convites enviados em licitações da modalidade Convite), Adjudicação/Homologação (abrange documentos de adjudicação, homologação ou que apresentam ambas as informações em um mesmo documento) e Outros (contempla os arquivos pertencentes aos demais tipos de documentos, por exemplo: erratas, anexos, contratos e memoriais descritivos). A Seção 4 apresenta uma descrição mais aprofundada das meta-classes aqui citadas.

Construção do meta-classificador. A partir das constantes interações com os documentos de licitação, perceberam-se padrões estruturais dos documentos. Também observou-se a presença de palavras-chave que se mostraram atributos promissores para a separação dos documentos entre as meta-classes propostas. A partir desse arcabouço, optou-se por desenvolver um método de classificação heurístico capaz de realizar a separação proposta, que facilitaria no processo manual de rotulação. O Algoritmo 1 apresenta os principais passos do classificador heurístico de meta-classes. Esse classificador é baseado em palavras-chave e consiste na análise da ocorrência dessas palavras no título e no conteúdo de cada documento de licitação. A Tabela 2 apresenta as palavras-chave definidas para identificar cada meta-classe nos documentos. Após o processo de concepção e refinamento das regras utilizadas em sua construção, o classificador de meta-classes é aplicado a todos os 6.337 documentos, resultantes após a filtragem.

Algoritmo 1: Classificador heurístico de meta-classes

```
   Entrada: Documentos de processos licitatórios em PDF
   Saída: Meta-classe atribuída a cada documento
1  início
2      para cada documento dPDF de cada cidade c faça
3          Extração do título e conteúdo da primeira página do documento;
4          Declara a variável countPalavrasTitulo; // Ocorrência da palavra-chave no título, por meta-classe
5          Declara a variável countPalavrasConteudo; // Ocorrência da palavra-chave no conteúdo, por
             meta-classe
6          para cada meta-classe faça
7              Atualiza countPalavrasTitulo com quantidade de palavras-chave que ocorreram no título;
8              Atualiza countPalavrasConteudo com quantidade de palavras-chave que ocorreram no
                 conteúdo;
9          fim
10         se existirem palavras-chave da meta-classe "Outros" então
11             meta_classe ← "Outros"
12         fim
13         se existirem palavras-chave da meta-classe "Homologação/Adjudicação" então
14             meta_classe ← "Homologação/Adjudicação"
15         fim
16         Ordena countPalavrasTitulo em ordem decrescente;
17         Ordena countPalavrasConteudo em ordem decrescente;
18         se houver ocorrência de palavra-chave no conteúdo então
19             meta_classe ← meta-classe associada
20         fim
21         meta_classe ← "Outros"
22     fim
23 fim
24 retorna Lista de documentos rotulados por meta-classe
```

Rotulação manual dos documentos. A partir da mobilização de sete membros do Programa de Capacidades Analíticas - MPMG/UFMG, todos os 6.337 documentos foram

Tabela 2. Meta-classes e palavras-chave associadas.

Meta-classe	Palavras-chave	Meta-classe	Palavras-chave
Ata	`ata`, `sessão pública`	Edital	`convite`, `edital`
Homologação/ Adjudicação	`homologação`, `adjudicação`	Outros	`cronograma`, `aditamento`, `ordem de serviço`, `resposta`, `extrato`, `diário oficial`, `aviso de`, `retificação`, `contrato administrativo`

Tabela 3. Dicionário de dados, contendo um exemplo de entrada.

Campo	Tipo de Dado	Exemplo
`file_id`	*string*	"d2a0a04e5954c3095c1c1bbabcb5a107"
`original_name`	*string*	"d2a0a04e5954c3095c1c1bbabcb5a107.pdf"
`n_pages`	*int*	1
`text_content`	*array*	["\n \n \n \n \nPREFEITURA MUNICIPAL DE OLARIA - TERMO DE RETIFICAÇÃO - Processo Licitatório nº \n055/2019 Pregão Presencial nº 014/2019, SOFREU ALTERAÇÕES na data de entrega de \ndocumentos de habilitação e proposta, devido o objeto da licitação estar escrito incorretamente, \ndessa forma, ONDE SE LÊ dia 22/05/2019, LEIA – SE dia 30/05/2019 as 09:00 (nove) horas ..."]
`table_content`	*array*	[ ]
`status`	*string*	"SUCCESS"
`city`	*string*	"olaria"
`text_preprocessed`	*string*	" termo retificacao processo licitatorio pregao presencial sofreu alteracoes data entrega documentos habilitacao proposta devido objeto licitacao estar escrito incorretamente forma le dia leia ... "
`meta_class`	*string*	"OUTROS"
`type_document`	*string*	"errata"

analisados e rotulados manualmente quanto à meta-classe real e o tipo de documento. Após o processo de rotulação, os resultados foram processados, culminando assim em um total de 56 tipos de documentos. Aqui, tipos de documentos referem-se a classes mais específicas, incluindo errata, aviso, ratificação, publicação diário oficial, contrato, aditamento, entre outras. Como a classe real da maioria dos documentos está presente no título do arquivo, o processo de rotulação manual foi realizado através da verificação dos títulos de cada documento. Portanto, não foi necessária a verificação de concordância entre os rotuladores, devido à confiabilidade do processo de rotulação.

Pré-processamento dos documentos. Após a rotulação manual dos documentos, o texto de cada documento foi pré-processado utilizando um conjunto de funções, que inclui transformação para letra minúscula e remoção de: nomes próprios, e-mails, URLs, pronomes, advérbios, caracteres especiais, acentos, *stop-words*, horas, símbolos de números, números, palavras contraídas e reduzidas, letras sozinhas no texto e espaços extras.

3.3. Armazenamento e Organização dos Dados

Para armazenar os dados relativos a um documento, foram utilizados arquivos no formato JSON (*JavaScript Object Notation*), que pode ser facilmente convertido em dicionários. Dentre as vantagens da escolha deste formato, está o fato de que um JSON consegue armazenar diferentes tipos de dados, possibilitando uma grande flexibilidade em relação aos dados armazenados. Na Tabela 3, são apresentados os campos presentes em cada arquivo JSON, o tipo de dado correspondente e um exemplo de entrada.

Em relação às informações armazenadas em cada campo, no campo "file_id", é armazenado um código padronizado de identificação hexadecimal único para cada docu-

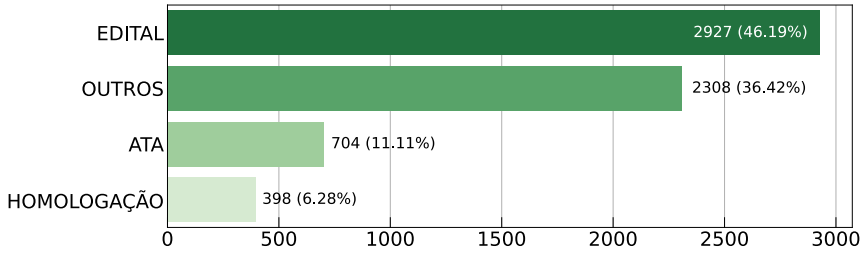

Figura 3. Quantidade de documentos por meta-classe.

mento, sendo o nome original do documento na base de dados de origem armazenado no campo "original_name" e o número de páginas inserido no campo "n_pages". Já os campos "text_content" e "table_content" possuem cada um, um array com os textos e tabelas presentes no arquivo original do documento. Ambas informações foram extraídas utilizando a biblioteca PDFPlumber. Por fim, os campos "status", "city" e "text_preprocessed" armazenam, respectivamente, o status do documento (i.e., se é processável ou não), a cidade de origem do documento e o texto pré-processado.

Visando a usabilidade do conjunto de dados em aplicações futuras, cada arquivo JSON apresenta os campos "meta_class" e "type_document". Tais campos armazenam, respectivamente, a meta-classe e o tipo de documento obtidos a partir da rotulação manual descrita na Seção 3.2. Ambas informações podem ser úteis, por exemplo, em tarefas de classificação de texto e documentos. Na Seção 5, descrevemos melhor como o LipSet pode ser utilizado em tal contexto.

3.4. Usabilidade

Seguindo os princípios da ciência aberta, o LiPSet está publicamente disponível em um repositório no Zenodo.[4] Para cada município do conjunto de dados, é disponibilizado um arquivo contendo as informações de todos os documentos de licitações públicas, incluindo o texto processado, a meta-classe e o tipo de documento. Tal arquivo está disponível em duas versões (CSV e JSON), que podem ser usadas a depender da aplicação considerada. Por exemplo, arquivos CSV podem ser mais úteis em análises complexas em linguagem Python ou R, enquanto arquivos JSON têm maior utilidade em aplicações Web.

4. Caracterização

Esta seção apresenta uma caracterização dos documentos presentes no LiPSet por meta-classe e município (Seção 4.1) e por quantidade de páginas (Seção 4.2).

4.1. Distribuição dos Documentos por Meta-classe e Município

Dos 9.761 documentos de 18 municípios que compõem o LiPSet, 2.223 não foram classificados por serem não-processáveis, identificados quando o campo "status" é *"FAILED"*. Além disso, 1.201 documentos são do tipo PDF, mas não fazem parte de nenhuma meta-classe por serem imagens, plantas e fotografias anexadas. Por conseguinte, o LiPSet possui 6.337 documentos de licitações públicas classificados em uma das quatro meta-classes definidas como: Adjudicação/Homologação, Atas, Edital e Outros. A Figura 3 apresenta a distribuição da quantidade de documentos por meta-classe e mostra que cerca de 83% dos documentos pertencem às meta-classes Edital e Outros.

[4]LiPSet: `https://doi.org/10.5281/zenodo.6974237`

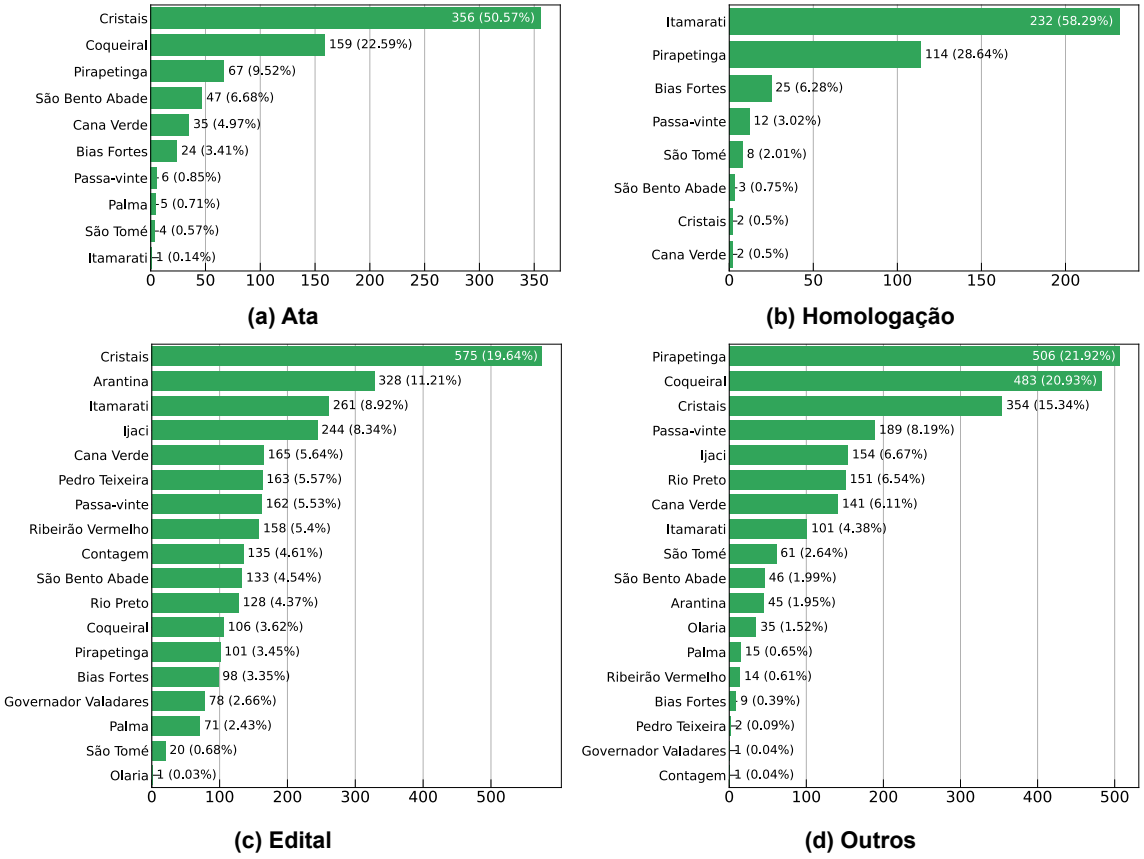

Figura 4. Quantidades de documentos de cada meta-classe, por município.

Nas Figuras 4a, 4b, 4c e 4d, também foi analisada a distribuição de documentos de cada meta-classe, por município. Os 704 documentos rotulados com a meta-classe Ata estão distribuídos em dez cidades, sendo que aproximadamente 51% são do município de Cristais (Figura 4a). Um comportamento semelhante é observado para a meta-classe Homologação, na Figura 4b. Dos 398 documentos rotulados, aproximadamente 87% deles pertencem a Itamarati e Pirapetinga. Por outro lado, todos os 18 municípios apresentam documentos das meta-classes Edital e Outros, conforme ilustrado nas Figuras 4c e 4d. Os municípios de Cristais, Arantina e Itamarati são responsáveis por quase 40% dos editais, enquanto Pirapetinga, Coqueiral e Cristais apresentam mais de 50% dos documentos rotulados como Outros.

4.2. Característica dos Documentos

Em relação ao conteúdo dos documentos, a Figura 5 exibe a distribuição do número de páginas dos documentos, por meta-classe. Para a meta-classe Edital, verifica-se uma grande dispersão entre os valores do número de páginas dos documentos, representada pelos diversos *outliers*. Nesse contexto, o menor número de páginas é igual a 1 e o maior é igual a 943. Já o valor mais frequente para o números de páginas para editais é igual a 27. Na meta-classe Outros, os valores dos números de páginas variam entre 1 página, valor mais predominante, e 262 páginas, o *outlier* mais discrepante. Já na meta-classe Homologação, não há *outliers*, e o valor do número de páginas dos documentos varia entre 1 e 28, sendo o valor de maior predominância igual a 5. Por fim, na meta-classe Ata, observa-se uma grande dispersão entre os valores do número de páginas dos

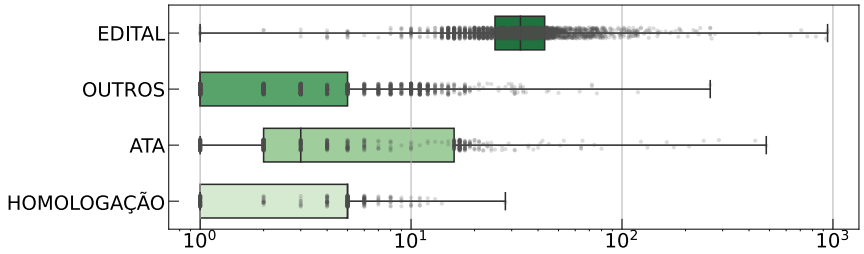

Figura 5. Distribuição do nº de páginas em escala logarítmica, por meta-classe.

documentos, confirmada pelos *outliers*. Os valores dos números de páginas variam entre 1 e 483 páginas, sendo 1 o valor mais predominante.

Analisar o número de páginas pode auxiliar na decisão de quantas páginas são consideradas em aplicações utilizando o LiPSet. Por exemplo, para a tarefa de classificação de documentos de licitação, é possível considerar apenas o vocabulário do texto das primeiras páginas, visto que a maioria dos documentos não possui mais de 10 páginas.

5. Aplicações

Conforme já mencionado, o LipSet é formado por documentos de licitações públicas de 18 municípios de Minas Gerais. Esses documentos foram rotulados, o que aumenta e potencializa sua aplicação. Algumas delas são descritas a seguir.

Treinamento de modelos de classificação. O LipSet pode ser utilizado para treinar um modelo de classificação e então obter os tipos dos documentos de outros municípios. Por meio desse modelo, é possível obter a classificação de novos documentos sem o esforço de novas rotulações manuais. Além disso, uma versão adaptada do modelo de classificação poderia ser utilizado para classificar documentos do processo licitatório de órgãos federais ou de outros estados. Isso é possível pelo fato dos processos licitatórios serem similares nas diferentes esferas do governo, resultando em documentos parecidos.

Análises aprofundadas de tipos específicos de documentos. Uma vez que se conheça os tipos de documentos dos municípios, é possível o desenvolvimento de fluxos específicos para extração de informações relevantes para cada tipo. Isso é importante por cada tipo de documento possuir um conjunto diverso de dados e ser estruturado de forma diferente. Resultando em fluxos mais precisos e significativamente mais fáceis de se implementar.

Análise de despesas públicas. As informações extraídas por esses fluxos específicos podem ser utilizadas para a organização de bases de dados históricas de órgãos públicos de forma automatizada e eficiente. Assim, é possível a criação de métricas mais extensivas para a análise de despesas públicas e mudanças de preços em produtos e serviços.

Detecção de fraudes. Os fluxos específicos para extração de informações do LiPSet podem ser utilizados na implementação de trilhas de auditagem para geração de alertas de fraude em licitações públicas, conforme o arcabouço proposto em [Costa et al. 2022]. Além disso, Velasco *et al.* [2021] propõem uma metodologia que utiliza algoritmos de mineração de dados para detecção de padrões de corrupção, auxiliando na identificação de fraudes. Já Anowar e Sadaoui [2019] desenvolveram um classificador de fraude que distingue entre licitantes legítimos e não legítimos, utilizando algoritmos de aprendizado de máquina supervisionado. Alguns estudos também exploraram metodologias baseadas

em redes neurais para a detecção de fraudes [Pereira and Murai 2021, Abidi et al. 2021]. Esses trabalhos utilizam técnicas de mineração de dados ou aprendizado de máquina, e possuem metodologias fortemente dependente dos dados. Portanto, um classificador de documentos treinado sobre o LipSet poderia auxiliar na extração de novos dados de licitações, e poderia aprimorar as metodologias citadas.

6. Desafios e Limitações

O LiPSet possui algumas limitações que podem ser o foco de melhorias em trabalhos futuros. Essas limitações estão relacionadas ao desafio de termos que lidar com muitos documentos diferentes, muitas vezes disponibilizados sem padronização em portais da transparência dos municípios. A seguir, os principais desafios e limitações são listados.

Processamento de documentos apenas em formato PDF. Conforme já mencionado, o LiPSet não inclui dados de documentos em formatos diferente do PDF. Também são descartados documentos escaneados ou contendo apenas imagens. Como alguns municípios possuem muitos documentos escaneados, a depender do foco da pesquisa, pode ser necessário que alguns trabalhos os considerem.

Falta de padronização. O LiPSet é composto por documentos muito distintos com pouca ou nenhuma padronização. Isso pode atrapalhar no resultado de aplicações que o utilizem como entrada. Uma sugestão de melhoria para esse problema é agrupar os documentos não apenas pela meta-classe, mas também por sua similaridade (i.e., por meio do uso de alguma abordagem que faça esse cálculo).

Desbalanceamento. Como os documentos são bastante distintos, para muitos deles, não foi possível categorizá-los em detalhes. Por isso, grande parte dos documentos foi agrupada na meta-classe Outros. Assim, tal desbalanceamento pode trazer problemas de representatividade dos dados. Portanto, trabalhos que necessitem considerar todos os documentos de licitação ou relacionados a licitações precisam explorar arquivos nessa meta-classe para melhor entendê-los.

Quantidade limitada de municípios. O LiPSet considera documentos de apenas 18 municípios mineiros, pois esse foi o foco de estudo da pesquisa que utilizou esse conjunto de dados. Apesar de uma abrangência relativamente pequena quando comparada à quantidade de municípios brasileiros, a metodologia de construção e rotulação do conjunto de dados apresentada neste trabalho pode ser aplicada para a ampliação do LiPSet e/ou construção de novos conjuntos de dados com documentos de outros municípios.

7. Conclusão

Este trabalho apresentou LiPSet, um conjunto de dados formado por documentos de licitações públicas. Para construção desse conjunto de dados, foram coletados documentos do Portal da Transparência de 18 municípios mineiros. Esses documentos foram então processados e rotulados para poderem ser utilizados facilmente em diferentes aplicações. A caracterização mostrou que há um grande desbalanceamento entre a quantidade de documentos por meta-classe e município. A descrição da aplicação, desafios e limitações revelaram o potencial de uso do LiPSet, bem como as oportunidades de pesquisa para melhorá-lo. Como trabalhos futuros, planeja-se considerar documentos de mais municípios mineiros para ampliar o escopo do LiPSet. Também espera-se aplicar algoritmos de similaridade para agrupar os documentos e, assim, facilitar o uso do conjunto de dados.

Agradecimentos. Ao Ministério Público de Minas Gerais (MPMG) pelo apoio através do Projeto Capacidades Analíticas. Ao CNPq, CAPES e FAPEMIG pelo apoio aos pesquisadores envolvidos.

Referências

Abidi, W. U. H. et al. (2021). Real-time shill bidding fraud detection empowered with fussed machine learning. *IEEE Access*, 9:113612–113621.

Anowar, F. and Sadaoui, S. (2019). Multi-class ensemble learning of imbalanced bidding fraud data. In *Canadian AI*, volume 11489 of *Lecture Notes in Computer Science*, pages 352–358. Springer.

Araújo, L. R. and Souza, J. F. (2011). Aumentando a transparência do governo por meio da transformação de dados governamentais abertos em dados ligados. *RESI*, 10(1).

Clarindo, J. P. et al. (2020). Qualisus: um dataset sobre dados da saúde pública no brasil. *SBBD DSW*, pages 418–428.

Costa, L., Reis, A., Bacha, C. A., Oliveira, G. P., Silva, M. O., Teixeira, M. C., Brandão, M. A., Lacerda, A., and Pappa, G. (2022). Alertas de fraude em licitações: Uma abordagem baseada em redes sociais. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 37–48, Porto Alegre, RS, Brasil. SBC.

Gabardo, A. C. and Lopes, H. S. (2014). Using social network analysis to unveil cartels in public bids. In *ENIC*, pages 17–21. IEEE.

Lima, M. et al. (2020). Inferring about fraudulent collusion risk on brazilian public works contracts in official texts using a bi-lstm approach. In *EMNLP*, pages 1580–1588.

Lyra, M. S. et al. (2021). Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil) municipalities. *Appl. Network Sci.*, 6(1):1–10.

Mata, W. R. R. et al. (2019). JusBD: Um Banco de Dados para Obtenção de Informações do Poder Judiciário. pages 398–407.

Meera, S. and Geerthik, S. (2022). Natural language processing. *Artificial Intelligent Techniques for Wireless Communication and Networking*, pages 139–153.

Muniz, R. I. V. C. S. and Lóscio, B. F. (2018). Publicação de Dados Abertos Conectados Sobre os Transplantes Realizados no IMIP. In *SBBD WTDBD*, Rio de Janeiro, Brasil.

Oliveira, E. F. and Silveira, M. S. (2018). Open government data in brazil a systematic review of its uses and issues. In *dg.o*, pages 1–9.

Pereira, L. S. (2022). Caracterização da comunidade que utiliza dados abertos governamentais sobre a educação brasileira. Master's thesis, Universidade Federal de Campina Grande, Campina Grande, Brasil.

Pereira, R. and Murai, F. (2021). Quão efetivas são redes neurais baseadas em grafos na detecção de fraude para dados em rede? In *BraSNAM*, pages 205–210. SBC.

Shimron, E. et al. (2022). Implicit data crimes: Machine learning bias arising from misuse of public data. *the National Academy of Sciences*, 119(13):e2117203119.

Silva, L. C. et al. (2020). Utilização de técnicas de mineração de dados para detectar possíveis relacionamentos entre empresas participantes de licitações nas forças armadas. *Acanto em Revista*, 7(7):85–85.

van Erven, G. C. G. et al. (2017). Detecting evidence of fraud in the brazilian government using graph databases. In *WorldCIST*, pages 464–473. Springer.

Velasco, R. B. et al. (2021). A decision support system for fraud detection in public procurement. *Int. Trans. Oper. Res.*, 28(1):27–47.