UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Henrique Ribeiro Hott

Classificação e Modelagem de Tópicos para Documentos de Licitação via NLP e Deep Learning

Belo Horizonte
2024

Henrique Ribeiro Hott

Classificação e Modelagem de Tópicos para Documentos de Licitação via NLP e Deep Learning

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Fabrício Murai Ferreira
Coorientador: Jefersson Alex dos Santos

Belo Horizonte
2024

Hott, Henrique Ribeiro.

H834c Classificação e modelagem de tópicos para documentos de licitação via NLP e deep learning [recurso eletrônico] / Henrique Ribeiro Hott. – 2024.

1 recurso online (100 f. il., color.) : pdf.

Orientador: Fabrício Murai Ferreira
Coorientador: Jefersson Alex dos Santos.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 90-100.

1. Computação – Teses. 2. Engenharia de software – Teses. 3. Processamento da linguagem natural (Computação) – Teses 4. Aprendizado profundo – Teses. 3. Informações eletrônicas Governamentais - Dados conectados – Teses. 5. Licitação pública – Belo Horizonte - Teses. I. Ferreira, Fabrício Murai. II. Santos, Jefersson Alex dos III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6*32(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais – ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS

CLASSIFICAÇÃO E MODELAGEM DE TÓPICOS PARA DOCUMENTOS DE LICITAÇÃO VIA NLP E DEEP LEARNING

HENRIQUE RIBEIRO HOTT

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Fabrício Murai Ferreira - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Jefersson Alex dos Santos - Coorientador
Departamento de Ciência da Computação - UFMG

Prof. Renato Antônio Celso Ferreira
Departamento de Ciência da Computação - UFMG

Profa. Adriana Silvina Pagano
Faculdade de Letras - UFMG

Belo Horizonte, 17 de dezembro de 2024.

Documento assinado eletronicamente por **Fabricio Murai Ferreira**, **Professor do Magistério**

Superior, em 18/12/2024, às 15:32, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jefersson Alex dos Santos**, **Supervisor(a)**, em 20/12/2024, às 11:08, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Renato Antonio Celso Ferreira**, **Professor do Magistério Superior**, em 20/12/2024, às 14:03, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Adriana Silvina Pagano**, **Professora do Magistério Superior**, em 30/12/2024, às 12:50, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3837188** e o código CRC **51B04B1D**.

Referência: Processo nº 23072.274394/2024-78 SEI nº 3837188

Aos meus pais e familiares, aos amigos que são minha família estendida e aos meus sábios mestres, minha mais sincera gratidão.

Agradecimentos

Tenho grande apreço por esta seção de agradecimentos. Nela, nós, autores acadêmicos, podemos nos desprender brevemente das tecnicidades da ciência e expressar livremente palavras de carinho e gratidão. O fato de que esta seção muitas vezes é ignorada ou subestimada pela maioria dos leitores e autores só aumenta seu charme único, como um pequeno grafite em um centro urbano: algo que muitos passam sem notar, mas que guarda uma mensagem especial para aqueles que param para observá-lo. Minha jornada foi intensa, marcante e desafiadora; por isso, tenho muito a agradecer e fico feliz por poder escrever, neste pequeno e modesto "muro", minhas palavras mais sinceras.

Em primeiro lugar, agradeço à professora Gisele Lobo Pappa, minha primeira orientadora, que me proporcionou a oportunidade de vivenciar o que mais amo fazer: pesquisa. Foi graças a você que participei do meu primeiro congresso e tive a liberdade de mostrar meu potencial em nossos trabalhos. Aos meus colegas de pesquisa, Mariana Oliveira, Gabriel Oliveira e Michele Brandão, deixo minha gratidão. O talento de vocês marcou profundamente a minha jornada, trazendo conhecimento e inspiração. Suas palavras me motivaram a crescer e evoluir como pesquisador a cada dia. Agradeço também aos meus atuais orientadores, Fabrício Murai e Jefersson A. Santos, que me acolheram com aptidão e paciência nesta etapa final da jornada. Vocês me mostraram que ainda havia possibilidades a serem exploradas e me guiaram com excelência até este momento.

O caminho foi, muitas vezes, sombrio, mas tive a sorte de contar com pessoas que trouxeram luz quando tudo parecia perder o sentido. Aos meus amigos PP e Gustavo Pelli, obrigado por virem até mim, me darem suporte e carinho, e me lembrarem de que eu nunca estive sozinho. Aos outros amigos, João Gabriel, Pedro Egg e Amanda, sou grato por todas as jogatinas e momentos compartilhados que me ajudaram a recarregar as forças e seguir em frente.

Por fim, um agradecimento especial aos meus pais. Vocês dividiram comigo o peso de cada desafio e estiveram presentes em todos os momentos, bons e difíceis. Sem o amor e o suporte incondicional de vocês, eu jamais teria chegado até aqui. Em sua honra, dedico não apenas este trabalho, mas também todos aqueles que ainda estão por vir.

"*Raramente, [...] nós nos curamos em isolamento. A cura é um ato de comunhão.*"
(Bell Hooks - Tudo sobre o amor: Novas perspectivas)

Resumo

Dados abertos governamentais (OGD, do inglês *Open Government Data*) englobam dados sobre ações, gastos e investimentos governamentais, disponibilizados de forma acessível e transparente ao público. No Brasil, a Lei de Acesso à Informação (Lei n.º 12.527 de 2011) assegura aos cidadãos o direito de acesso a informações dos três poderes da União, do Distrito Federal, dos estados e dos municípios. O OGD promove a transparência e a participação pública, sendo essencial para diversas aplicações tanto no setor público quanto no privado. Em particular, as licitações públicas – que envolvem uma ampla gama de documentos como atas, editais e erratas – são processos sensíveis a fraudes e irregularidades, uma vez que lidam diretamente com o uso de recursos públicos. Em resposta a esses desafios, pesquisas recentes vêm desenvolvendo aplicações orientadas a dados que fortalecem a inteligência de órgãos reguladores e facilitam o monitoramento das despesas públicas pelos cidadãos, promovendo a segurança no processo licitatório. Este trabalho apresenta um estudo de caso em Minas Gerais, focado em aplicações de *Natural Language Processing* (NLP) e *Deep Learning* (DL) para automatizar processos e detectar padrões latentes em documentos de licitações municipais. As licitações municipais trazem desafios adicionais, pois a ausência de padronização entre os portais de transparência dos municípios gera uma diversidade de formatos e modos de disponibilização dos documentos. Nosso trabalho se inicia com a construção do LiPSET, um conjunto de dados com 9.761 documentos de 18 municípios de Minas Gerais, dos quais 6.337 foram rotulados em 4 meta-classes e 13 tipos, elaborados com o apoio de especialistas. A caracterização do LiPSET permitiu observar os principais desafios do domínio de licitações públicas municipais, como o desbalanceamento de classes e a falta de padrão na distribuição dos documentos por município. Apresentamos também o LiBERT-SE, um modelo BERT adaptado e treinado especificamente para documentos de licitação pública, com potencial para servir de base para diversas aplicações. Duas aplicações práticas foram desenvolvidas: a primeira é uma classificação automática de documentos que usa métodos heurísticos e modelos LSTM, avaliando técnicas de pré-processamento e de representação textual para otimizar o desempenho; nesta aplicação, os métodos propostos alcançaram bons resultados, com valores para *F1-Macro* e *F1-Weighted* superiores a 96% na melhor configuração para na classificação por tipo de documento.

Para a classificação das meta-classes utilizando o metido heurístico o resultado foi 91% de F1-Macro. A segunda aplicação, de modelagem de tópicos latentes em documentos de licitação, avalia o potencial do LiBERT-SE para identificar padrões utilizando o BERTopic, uma metodologia de modelagem de tópicos baseada em técnicas de agrupamento e *sentence embeddings* contextualizados; neste contexto, o LiBERT-SE superou consistentemente os *baselines* com significância estatística nas métricas de coerência e diversidade de tópicos. Em geral, a avaliação qualitativa identificou temas gerais de licitação nos tópicos gerados pelo LiBERT-SE; contudo, observou-se muito ruído nas palavras que descrevem os tópicos. Em suma, as aplicações apresentadas demonstraram impactos práticos para a classificação de documentos, enquanto a modelagem de tópicos ainda requer refinamento para potencializar sua aplicação em consultas de interesse público. O conjunto de dados LiPSET e o modelo LiBERT-SE estão publicamente disponíveis e podem ser utilizados como base para futuras pesquisas voltadas à detecção de padrões em documentos de licitação pública.

Palavras-chave: processamento de linguagem natural;aprendizado profundo;classificação de documentos;modelagem de tópicos;dados abertos governamentais;licitações públicas.

Abstract

Open government data (OGD) encompasses data on government actions, expenditures, and investments, made accessible and transparent to the public. In Brazil, the Access to Information Law (Law No. 12,527 of 2011) guarantees citizens the right to access information from the three branches of government, Federal District, states and municipalities. OGD promotes transparency and public participation, and is essential for a variety of applications in both the public and private sectors. In particular, public procurement processes – which involve a wide range of documents such as minutes, notices, and errata – are processes that are susceptible to fraud and irregularities, since they directly involve the use of public resources. In response to these challenges, recent research has been developing data-driven applications that strengthen the intelligence of regulatory agencies and facilitate the monitoring of public expenditures by citizens, promoting security in the bidding process. This paper presents a case study of Minas Gerais, focused on applications of *Natural Language Processing* (NLP) and *Deep Learning* (DL) to automate processes and detect latent patterns in municipal bidding documents. Municipal bidding processes bring additional challenges, since the lack of standardization among the transparency portals of the municipalities generates a diversity of formats and ways of making the documents available. Our work begins with the construction of LiPSET, a dataset with 9,761 documents from 18 municipalities in Minas Gerais, of which 6,337 were labeled in 4 meta-classes and 13 types, developed with the support of experts. The characterization of LiPSET allowed us to observe the main challenges of the municipal public bidding domain, such as class imbalance and the lack of pattern in the distribution of documents by municipality. We also present LiBERT-SE, a BERT model adapted and trained specifically for public bidding documents, with the potential to serve as a basis for several applications. Two practical applications were developed: the first is an automatic classification of documents that uses heuristic methods and LSTM models, evaluating preprocessing and textual representation techniques to optimize performance; in this application, the proposed methods achieved good results, with values for *F1-Macro* and *F1-Weighted* higher than 96% in the best configuration for classification by document type. For the classification of meta-classes using the heuristic method, the result was 91% of F1-Macro. The second application, latent topic

modeling in bidding documents, evaluates the potential of LiBERT-SE to identify patterns using BERTopic, a topic modeling methodology based on clustering techniques and contextualized *sentence embeddings*; in this context, LiBERT-SE consistently outperformed the *baselines* with statistical significance in the metrics of topic coherence and diversity. In general, the qualitative evaluation identified general bidding themes in the topics generated by LiBERT-SE; however, noises was observed in the words describing the topics. In summary, the presented applications demonstrated practical impacts for document classification, while topic modeling still requires refinement to enhance its application in public interest queries. The LiPSET dataset and the LiBERT-SE model are publicly available and can be used as a basis for future research aimed at detecting patterns in public bidding documents.

Keywords: natural language processing;deep learning;document classification;topic modeling;open government data;public bidding.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Os avanços significativos nas Tecnologias de Informação impactaram profundamente quase todas as áreas da sociedade. Em particular, as iniciativas digitais redesenharam a dinâmica do poder governamental, permitindo maior participação popular e transparência nas ações públicas [2]. Neste contexto, um dos principais tópicos na literatura é o uso de Dados Abertos Governamentais (OGD, do inglês *Open Government Data*) [81]. O OGD é constituído por conjuntos de dados referentes a ações, gastos e investimentos governamentais distribuídos de forma acessível e transparente ao público [86]. Especificamente no Brasil, com a Lei de Acesso à 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 [49, 22, 57, 85, 87].

A área de *Natural Language Processing* (NLP) tem como objetivo desenvolver modelos computacionais capazes de capturar e interagir com as complexidades e nuances da comunicação humana [82]. Para OGD a NLP é uma ferramenta importante pois grandes partes dos dados desse campo são compostos de uma vasta gama de documentos textuais complexos com um vocabulário técnico de caráter jurídico [40, 73].

Nesse contexto, técnicas de Deep Learning (DL) têm sido proeminentes para o desenvolvimento de aplicações robustas de OGD [14, 73, 85, 87]. O DL é um campo específico do Machine Learning (ML) que emprega o uso de Artificial Neural Networks (ANNs) para extrair padrões e atributos complexos de grandes conjuntos de dados [33]. Especificamente, na presente dissertação nós exploramos os desafios existentes em documentos de licitações públicas através de modelos de DL.

Uma licitação pública é um processo administrativo usado pelo governo para con-

[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

tratar serviços, adquirir bens ou realizar obras, garantindo que a escolha seja feita de forma justa, transparente e econômica. As licitações são compostas por uma alta quantidade de diferentes documentos, incluindo o texto do edital, anexos, errata, ata, projeto, memorial descritivo, dentre outros. Cada um desses documentos possui um formato específico que traz informações diferentes do processo de licitação.

Em especial, o processo licitatório é campo sensível por tratar de uso de recursos públicos. Acontecem com frequência casos de corrupção e fraude, segundo a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) [28]. Para promover avanços na segurança e transparência do processo licitatório, este trabalho propõe o desenvolvimento de aplicações práticas que visam, simultaneamente, apoiar investigadores na detecção de padrões irregulares em documentos de licitação e explorar as características fundamentais desses dados para criar aplicações robustas baseadas em NLP e DL.

O trabalho realizado tem caráter multidisciplinar e contou com o apoio de especialistas do Ministério Público de Minas Gerais (MPMG) para seu desenvolvimento. A seguir, detalhamos a parceria com o MPMG, definimos os principais objetivos deste estudo e apresentamos a estrutura da dissertação.

1.1 Contextualização do trabalho

A dissertação foi desenvolvida como parte do programa de capacidades analíticas (PCA) do MPMG. O projeto é uma parceria entre a Universidade Federal de Minas Gerais (UFMG) e o MPMG e tem como objetivo desenvolver soluções e metodologias baseadas em Data Science (DS) para apoiar e expandir a inteligência do órgão em seus trabalhos de investigação.

Muitos problemas enfrentados no projeto envolvem grandes e complexos volumes de dados textuais não estruturados. Mais especificamente, o trabalho foi realizado por uma equipe de 7 membros constituída de 2 alunos de graduação, 3 alunos de pós-graduação incluindo o autor da presente dissertação, um bolsista de pós-doutorado e um gerente empregado pelo programa. O time que encerrou seu trabalho em setembro de 2023 e era responsável por parte do fluxo de alimentação das bases de dados e pesquisa utilizando documentos de licitações municipais do Estado de Minas Gerais.

Trabalhar com licitações municipais é uma tarefa difícil, pois não há padronização nem no texto dos documentos nem em sua disponibilização. Cada município desenvolve de maneira própria o portal da transparência, *site* onde ficam armazenados os documentos de licitação. Isso nos leva ao desafio de lidar com dados diversos, apresentados de forma heterogênea.

Por isso, a equipe desenvolveu uma série de trabalhos que abrangem desde a coleta de documentos até sua classificação, visando um fluxo eficiente de ingestão e organização de dados. Foram utilizadas técnicas avançadas, como *web crawlers*, para automatizar o fluxo de coleta, além de modelos de classificação e modelagem de tópicos (MT) baseados em DL e métodos mais simples baseados em heurísticas. Neste documento, apresentamos o trabalho realizado pela equipe e suas principais contribuições para a pesquisa no domínio de licitações públicas.

1.2 Objetivos

Com base na contextualização apresentada na seção anterior, definimos o objetivo geral desta dissertação como: *Realizar um estudo aprofundado de aplicações práticas de NLP que possam contribuir para o trabalho dos especialistas do MPMG na detecção de irregularidades em documentos e enfrentar os principais desafios no desenvolvimento dessas aplicações.*

Para alcançar nosso objetivo geral, empregamos principalmente modelos de DL que são o estado-da-arte em NLP para propor novas aplicações capazes de extrair padrões úteis nos documentos. Dessa forma, vamos avaliar o desempenho desses modelos em um tema ainda pouco explorado na literatura, que são os documentos de licitações públicas.

A seguir, definimos os objetivos específicos da dissertação:

1. Introduzir um novo conjunto de dados rotulado composto por diferentes tipos de documentos de licitações públicas extraídas dos portais da transparência de municípios do estado de Minas Gerais.

2. Caracterizar os documentos de licitação municipal e avaliar os principais desafios contidos nestes para o desenvolvimento de aplicações robustas utilizando DL.

3. Desenvolver uma aplicação prática supervisionada atráves de uma nova proposta de classificação automática de documentos de licitação a partir classes definidas por nós utilizando os documentos coletados.

4. Desenvolver uma aplicação prática não supervisionada para extração de padrões nos documentos de licitação atráves de uso de técnicas de agrupamento e modelagem dos tópicos latentes.

5. Propor novos modelos e métodos baseados em técnicas de NLP consolidadas para o domínio de licitações públicas.

1.3 Estrutura da dissertação

Nos capítulos seguintes serão descritos a revisão da literatura dos trabalhos relacionados, o referencial teórico base para a dissertação, a metodologia proposta, o design experimental aplicado, a discussão sobre os resultados e, por fim, a conclusão.

Primeiro, o Capítulo 2, Trabalhos Relacionados, discute os estudos sobre dados abertos governamentais e outros trabalhos que tratam dados de licitações no contexto brasileiro e internacional. Em seguida, o Capítulo 3, Referencial Teórico, apresenta a base teórica para a dissertação a partir do histórico das técnicas de representação textual baseadas em DL e os conceitos principais para o desenvolvimentismo de uma aplicação de NLP.

No Capítulo 4 de Metodologia, discutimos o processo de coleta, rotulação e estruturação dos documentos de licitações municipais e propomos duas aplicações práticas de classificação automática de documentos e MT. Na sequência, o Capítulo 5, Métodos Propostos e Design Experimental, avalia a escolha técnica dos métodos empregados e os procedimentos para avaliação das aplicações propostas. Por fim, no Capítulo 6 de Resultados Obtidos, analisamos os resultados práticos obtidos dos experimentos.

Finalmente, no Capítulo 7, Conclusão, discutimos criticamente os impactos reais da pesquisa, as contribuições para o domínio de licitações públicas e limitações do trabalho. Além disso, delineamos os trabalhos futuros a serem explorados em pesquisas subsequentes.

Capítulo 2

Trabalhos Relacionados

Para situar esta dissertação na literatura e enfatizar suas principais contribuições, revisamos, neste capítulo, a literatura existente e os principais trabalhos relacionados ao uso de *Open Governement Data* (OGD) e às técnicas de *Artificial Intelligence* (AI) aplicadas à análise de documentos legais e administrativos de instituições públicas.

Além disso, a crescente disponibilidade de dados governamentais tem impulsionado o desenvolvimento e a aplicação de técnicas de IA, especialmente no domínio do Natural Language Processing (NLP) para análise de grandes volumes de documentos legais e administrativos. Este capítulo apresenta uma revisão detalhada das metodologias e abordagens utilizadas na literatura para lidar com dados legais, incluindo técnicas para a detecção de fraudes em licitações públicas [46, 35, 39], sumarização de documentos [85, 47], e recuperação da informação [40, 74, 87, 66, 11], entre outras [49, 38, 84, 75, 11, 73, 22].

Dessa modo, organizamos este capítulo em três seções principais. A Seção 2.1 aborda os conceitos fundamentais e principais aplicações de OGD, com foco nos esforços internacionais e nas tendências globais. A Seção 2.2 concentra-se na implementação e nos desafios específicos do uso de dados jurídicos de OGD no Brasil, discutindo as iniciativas de pesquisa nacionais. Finalmente, a Seção 2.2.2 debate as contribuições da presente dissertação em relação ao campo OGD no geral.

2.1 Dados abertos governamentais

Diferentes esforços de pesquisa têm sido realizados para analisar os desafios e impactos do uso da Tecnologia da Informação (TI) pelas repartições públicas [56, 70, 68, 62].

Nesse contexto, em [81], foi realizada uma revisão sistemática da literatura, avaliando-se mais de 189 artigos publicados entre 2009 e 2021. O trabalho chama atenção para a importância do uso do OGD, que compõe a maior parte da literatura. No entanto, critica-se a falta de caráter interdisciplinar dos trabalhos publicados, que focam majoritariamente na questão dos dados e deixam de lado o caráter multidimensional do conceito de governos abertos. O trabalho conclui com um direcionamento para futuros estudos na área, sugerindo a promoção de ferramentas capazes de unir especialistas e o público dentro do processo governamental.

Em [86], OGD é definido como um conjunto de dados sobre ações governamentais, gastos e investimentos, disponibilizados de maneira acessível e transparente ao público. O estudo realizou uma consulta com 210 participantes sobre as expectativas em relação ao possíveis benefícios do OGD. Os resultados mostraram que os participantes têm sido otimistas quanto a novas propostas de novas ferramentas que serão desenvolvidas por pesquisadores e entidades públicas, e ressaltaram que a acessibilidade deve ser um fator-chave para que o OGD possa contribuir para governos abertos e mais democráticos.

2.1.1 Revisão da Literatura Internacional sobre OGD

O crescimento do OGD tem possibilitado o desenvolvimento de novas tecnologias de inteligência, que têm um impacto positivo ao direcionar corretamente as políticas governamentais e ao economizar recursos [2]. Falando sobre a capacidade do OGD, em [38], foram utilizados dados históricos de mediações para criar um modelo de DL capaz de guiar o usuário em formas de resolução de conflitos com custo mínimo. No mesmo contexto, em [40], foi proposto um modelo de extração automática de conhecimento do Código Federal do governo estadunidense para sua representação em forma de ontologias. Eles transformaram os dados textuais do Código Federal em uma estrutura de busca com vários tópicos e linhas, permitindo que empresas e agências não especializadas conduzissem seu trabalho de forma a manterem-se comprometidas com a regulamentação legal.

Além dos trabalhos supracitados, em [57] são abordadas diversas publicações que utilizam dados públicos para criar modelos de detecção de fraude. Essa literatura revela o uso extensivo de técnicas de *Machine Learning* (ML). O trabalho de [57] também

chama a atenção para o sucesso crescente das técnicas de NLP baseadas em DL, que são mais adequadas para lidar com dados textuais, que compõem a maior parte dos dados governamentais abertos disponíveis.

2.1.2 Dados abertos governamentais no Brasil

O Brasil aderiu de vez ao movimento internacional do OGD através da publicação da Lei de Acesso à Informação em 2011. Essa nova legislação estabeleceu diretrizes legais para a abertura dos dados governamentais brasileiros, incluindo o processo licitatório. A partir dessa lei, foi direcionado ao governo federal, aos estados e aos municípios o desenvolvimento de portais de transparência, que fornecem uma interface para o usuário final realizar consultas públicas a documentos governamentais.

Apesar dessa nova iniciativa, existe uma lacuna na avaliação dos portais de OGD no Brasil, uma lacuna que também pode ser observada em outros países em desenvolvimento [2]. Em resposta a esse desafio, o trabalho de [50] é um dos primeiros sobre o tema, utilizando um método denominado modelo de 5 estrelas para avaliar portais de transparência brasileira em nível federal, estadual e municipal. O estudo aponta que acessar esses portais para o desenvolvimento de novos projetos é algo difícil. Além disso, em [27] e [24], após avaliar portais em diferentes contextos, foram apontados os principais desafios enfrentados pelos pesquisadores que lidam com a OGD no Brasil, tais como: qualidade dos dados, dificuldade de uso dos portais, dados não padronizados em diferentes formatos, e fontes de dados múltiplas e descentralizadas, sem padrões para a publicação.

Finalmente, em [41] foi aplicado um questionário para conduzir a análise da usabilidade dos portais da transparência no Brasil. Os autores listaram os principais fatores que dificultam o uso desses portais, incluindo a ausência de uma cultura orgânica favorável a dados abertos e o fato de que muitos gestores públicos não têm conhecimento sobre o que são dados abertos ou não estão interessados no aumento do controle social sobre eles.

Em consequência dos problemas persistentes do OGD no Brasil, avanços de pesquisa têm sido realizados com o propósito de melhorar a qualidade e o acesso aos dados governamentais. Um exemplo é o trabalho QualiSuS [19], que disponibiliza um banco de dados relacional criado a partir da extração de dados do Portal DataSUS. Já na área legal, o

projeto JusBD [49] apresenta um conjunto de dados não rotulados de audições forenses no contexto judiciário.

2.2 Aplicações brasileiras de Dados Legais Abertos (DLA)

As repartições governamentais utilizam uma vasta gama de documentos legais para sua gestão, o que envolve diferentes tipos de informações, muitas vezes não padronizadas e disponibilizadas em formatos de texto aberto. Como mencionado na seção anterior, técnicas de NLP são especialmente úteis para lidar com esses tipos de documentos.

Desse modo, em [85] os autores abordam o problema da organização e o desafio de sumarização de documentos legais brasileiros, destacando a complexidade e o jargão jurídico, que incluem diversos termos e frases específicas. Complementando esse trabalho, em [84] os autores utilizam técnicas de pré-processamento combinadas com representações textuais baseadas em DL para criar representações dos documentos e aplicam técnicas de geração de tópicos para a organização automática desses documentos.

Também no campo da sumarização de documentos legais, o trabalho de [47] apresenta o modelo CLSJUR.BR, baseado em aprendizado contrastivo e avaliado com uma metodologia livre de referência. Para o treinamento e construção de um corpus de ouro, foi utilizado o conjunto de dados chamado Ruling.BR, composto por 10.623 sentenças da Suprema Corte Federal. A metodologia apresentada obteve resultados próximos do ótimo em comparação com o conjunto de dados de referência.

Por fim, o trabalho [87] utiliza técnicas de ranqueamento de texto baseadas em ML para filtrar documentos do Diário Oficial do governo brasileiro.

2.2.1 Modelos de Representação Textual e Recuperação de Informação

No contexto de recuperação de informação, o trabalho [74] emprega o BERTopic [34] para capturar os principais tópicos abordados em documentos legais. Outro exemplo, em [66] é apresentado um estudo de caso que avalia o método BM25 e modelos S-BERT [69] no campo da similaridade textual. Para tal, foram utilizadas 269 propostas de emendas relativas à PEC 6/2019 enviadas à Câmara Legislativa e ao Senado brasileiro, que foram usadas como busca textual para organizar propostas semelhantes. Cada documento foi rotulado por consultores em 28 tópicos, descritos por uma ou mais palavras. Os resultados demonstram que a combinação de técnicas de pré-processamento com o modelo BM25 obteve resultados superiores aos modelos SBERT. O trabalho conclui que essa maior performance do modelo BM25, em comparação aos outros modelos testados, se deve às características dos documentos, que muitas vezes contêm as palavras que descrevem o tópico demarcado.

Em [21] foram apresentados quatro conjuntos de dados para a tarefa de similaridade textual. Os dados são relativos a documentos extraídos dos portais sobre julgamentos do Supremo Tribunal de Justiça e votos do Tribunal de Contas da União. O trabalho inova ao propor uma técnica heurística para a rotulação dos documentos a partir de meta-dados disponíveis. A nova técnica foi validada estatisticamente a partir de um questionário com amostras dos pares rotulados pela heurística e perguntas relativas à sua verdadeira similaridade. Ao todo, foram 240 questionários respondidos por 27 estudantes de mestrado em Direito.

Avanços também têm sido feitos em relação à representação textual de dados legais brasileiros. Em [75] é apresentado o modelo pré-treinado LegalBERTPT, desenvolvido especialmente para o domínio jurídico em português. Baseado no modelo BERTimbau [79], que foi treinado em um corpus geral em português, o LegeBERT reforça a etapa de pré-treinamento na tarefa de *Masked Language Modeling* (MLM) utilizando um corpus especializado no domínio legal. O corpus textual utilizado inclui 1,5 milhões de documentos de 10 cortes brasileiras, coletados por meio de *web crawlers*. Além do reforço no pré-treinamento, o trabalho expande o vocabulário original do BERTimbau adicionando 5.977 novas palavras extraídas dos documentos legais, utilizando a biblioteca SentencePiece [42]. O novo modelo apresentou uma métrica de perplexidade menor em comparação ao seu

modelo base quando testado em dados legais (quanto menor, melhor), e resultados superiores em comparação com outros modelos aplicados aos principais conjuntos de dados disponíveis na literatura.

Próximo ao contexto da presente dissertação, uma outra equipe do Programa de Capacidades Analíticas (PCA) do Ministério Público do Estado de Minas Gerais (MPMG) também apresentou em [11] um estudo aprofundado sobre representação textual para agrupamento de itens similares de licitação. Os itens de licitação frequentemente não são padronizados, e o mesmo item pode ter diversas formas de escrita. Para enfrentar o problema da ambiguidade, a equipe avaliou desde modelos bag-of-words até modelos de embeddings baseados em transformers para agrupamento usando o algoritmo HDBSCAN. Para a construção e avaliação do modelo, foram utilizadas as descrições de mais de 2,1 milhões de itens de licitação em português. Os experimentos destacaram a efetividade da combinação de métodos heurísticos simples com modelos de representação baseados em BERT e SIF.

2.2.2 Detecção de Fraudes em documentos legais

No contexto de detecção de fraudes em [39] foca-se na detecção de padrões de risco de fraude/corrupção através da análise de um conjunto de dados de notas fiscais emitidas por empresas que vendem produtos e prestam serviços a órgãos públicos do estado de Mato Grosso. Os autores utilizam diferentes tipos de dados extraídos em estudos anteriores de notas fiscais eletrônicas emitidas por entidades públicas. A metodologia adotada inclui abordagens de pré-processamento, métodos de agrupamento e algoritmos de detecção de *outliers* para averiguar itens suspeitos de fraude.

Um outro da mesma linha foi apresentado em [46], onde foi desenvolvido um modelo biLSTM [36] para detecção de risco de fraudes em fragmentos publicados no Diário Oficial da União relacionados a licitações. Para o treinamento do modelo, os autores contaram com o auxílio de uma equipe de especialistas forenses que, entre os 15.132.968 artigos coletados, levantaram manualmente uma amostra de 1.907 artigos com risco de fraude. O trabalho apresenta resultados positivos do uso de modelos DL para a extração de padrões de irregularidades em dados textuais abertos de licitações.

Avaliado o desenvolvimento de aplicações de detecção de fraude em [45] foi proposto

um *framework* para avaliação de agrupamento textual utilizando seis abordagens diferentes aplicadas a textos legais. O foco desse trabalho é agrupar documentos por similaridade para aumentar a acurácia da análise realizada por investigadores.

2.2.3 Projetos de pesquisa relacionados

Além de todos os trabalhos mencionados até então, destacam-se na literatura brasileira sobre OGD projetos de pesquisa que têm expandido os horizontes para novas oportunidades de pesquisa. Nesse contexto, o projeto Ulysses tem trabalhado com o objetivo de aumentar a transparência de dados para a população e promover o desenvolvimento de novas aplicações em dados legais abertos. Em [1] é apresentado o UlyssesNER-Br, um conjunto de dados anotado para a tarefa de *Named Entity Recognition* (NER). Nesse trabalho, foram construídos três grandes corpos textuais com dados coletados do *site* oficial da Câmara dos Deputados do Brasil. Cada corpo abrange uma temática diferente relacionada ao trabalho realizado pelos deputados, sendo elas: (I) Projetos de Lei, (II) Solicitações de Trabalho e (III) Documentos de Gestão Interna da Câmara, fornecidos pela própria equipe de trabalho do local. Além disso, para melhorar a organização dos documentos, foram identificados, com a ajuda de especialistas, 18 tipos de entidades estruturadas em 7 categorias que abrangem temas jurídicos específicos.

De forma similar à abordagem não supervisionada desta dissertação, o trabalho [73], realizado no contexto do Projeto Ulysses, apresenta um novo modelo pré-treinado para análise e geração de tópicos em comentários de cidadãos sobre projetos de lei na Câmara Legislativa Brasileira. O trabalho caracteriza os principais tópicos discutidos pelos usuários através da metodologia BERTopic [34].

Expandindo os trabalhos anteriores do projeto Ulysses, [77] apresenta o Ulysses Tesemõ, um vasto conjunto de dados abertos não rotulados que abrange desde comentários de usuários, documentos de instituições legislativas e governamentais, até trabalhos acadêmicos. Os dados foram extraídos de 159 fontes oficiais, totalizando mais de 3,5 milhões de documentos, contemplando 30,5 GiB de dados em texto aberto. O conjunto foi organizado em uma árvore de diretórios, contendo o arquivo de texto principal e os metadados relacionados. Os metadados categorizam os documentos em 30 tópicos, de acordo com

as características de cada fonte, similaridade entre propostas, grau de formalidade, estrutura e escopo.

Além do Projeto Ulysses, destaca-se na literatura o Projeto VICTOR [22], que é focado no desenvolvimento de métodos de ML no domínio jurídico. O projeto disponibilizou um grande conjunto de dados textuais que compreende publicações relacionadas a recursos extraordinários apresentados ao Supremo Tribunal Federal, contendo 692.966 documentos diferentes, compreendendo mais de 4,5 milhões de páginas. Os arquivos coletados eram compostos de PDFs (Portable Document Format) disponibilizados oficialmente e imagens de documentos escaneados. Para extrair o texto dos documentos escaneados, foram aplicadas e validadas técnicas de *Optical character recognition* (OCR).

Para criar uma tarefa de classificação, um conjunto de 628,820 documentos foi rotulado por um time de funcionários do tribunal para duas tarefas de classificação, quais sejam: o tipo de documento, que compreende 7 classes, e o tema, que associa uma ou mais classes dentre 28 relacionadas à recuperação geral do documento. Para cada tarefa proposta, o trabalho oferece os resultados *baselines*.

2.3 A presente dissertação frente ao estado da arte

Considerando os trabalhos relacionados apresentados neste capítulo, discutimos, nesta seção, as contribuições desta dissertação para o estado da arte em OGD.

Como mencionado anteriormente, licitações públicas são um alvo sensível para fraudes. Embora existam alguns estudos na literatura sobre o tema [39] e [45], ainda há uma lacuna para o desenvolvimento de aplicações que considerem o texto presente nos documentos. O texto aberto compõe a maior parte das informações envolvidas em uma licitação, nos quais itens e processos importantes são descritos sem um padrão ou norma bem definidos.

Inspirados por outros estudos que utilizam textos legais, esta dissertação contribui ao propor um pipeline automatizado para a classificação de documentos de licitação e a análise exploratória dos principais tópicos latentes em licitações. A princípio, nosso estudo permite a organização dos documentos com base em critérios de similaridade e contexto, facilitando o trabalho de investigação dos especialistas. Além disso, ao propor essas apli-

Tabela 2.1: Tabela comparativa dos principais trabalhos relacionados e a atual dissertação.

Contribuições	**Principais trabalhos**					**Presente dissertação**
	[22]	[75]	[73]	[45]	[46]	
Domínio de licitações publicas					✓	✓
Disponibilização de dados	✓				✓	✓
Classificação de documentos	✓	✓			✓	✓
Representação textual		✓	✓			✓
Análise de tópicos			✓	✓		✓

cações clássicas de NLP, conseguimos estudar as especificidades impostas pelo domínio de licitações e desenvolver novas abordagens que podem ser base para novas aplicações.

Outra lacuna identificada é a falta de conjuntos de dados abertos de licitações públicas. Em resposta a essa necessidade de dados para o desenvolvimento das aplicações propostas, esta dissertação contribui disponibilizando um conjunto de dados aberto e de fácil acesso, agregando documentos textuais de licitações de diferentes municípios de Minas Gerais. A Tabela 2.1 apresenta uma comparação entre esta dissertação e os principais trabalhos da literatura que utilizam abordagens semelhantes.

Capítulo 3

Referencial Teórico

O textitDeep Learning (DL) tem sido um dos fundamentos básicos para o desenvolvimento de aplicações de *Natural Language Processing* (NLP) modernas [60] . Esse destaque deve-se aos métodos de representação textual, que conseguem extrair atributos sintáticos, semânticos e contextuais de grandes corpus textuais. Essas representações ricas servem de base para a construção de aplicações de NLP.

Neste capítulo, discutimos as bases teóricas para o desenvolvimento de aplicações modernas de NLP que utilizam extensivamente o aprendizado profundo. Desse modo, primeiro a Seção 3.1 descreve as principais arquiteturas de DL voltadas para NLP, com ênfase nos Modelos de Linguagem Neurais (MLNs). Em seguida, a Seção 3.2 discute os fundamentos das aplicações de NLP, apresentando um *pipeline* padrão para essas aplicações.

3.1 *Modelos de Linguagem Neurais (MLNs)*

A tarefa de *Language Modeling* representa objetivo base para criarmos representações computacionais da linguagem natural [48]. A LM pode ser descrita como uma distribuição de probabilidades $p(s)$ sobre um conjunto de sentenças que, por sua vez, são formadas por uma sequência de *palavras*. Desta forma, dada uma sequência de palavras s com $w_1, w_2, ..., w_t$, um modelo LM atribui uma probabilidade a s que representa a chance de sua ocorrência em corpo textual qualquer.

A abordagem mais tradicional para a LM consiste em métodos estatísticos para estimar $p(s)$ com base na frequência da palavra no conjunto de dados [16]. Nesse contexto, se tornou popular o uso de N-Gramas que empregam uma suposição Markoviana de grau

n para prever uma palavra w_i considerando seus n antecessores: $p(w_i|w_i-1, w_i-2...w_i-n)$. Desse modo, um corpus textual é representado pela matriz de probabilidades de co-ocorrência dos termos.

Até hoje, a abordagem estatística mantém um lugar importante na literatura, sendo a base de modelos de representação robustos como o *Global Vectors for Word Representation*(GloVE) [63]. Entretanto, a abordagem clássica apresenta limitações importantes por produzirem representações esparsas e terem uma alta complexidade computacional [58].

Para superar as limitações da abordagem clássica, surge na literatura a proposta de representações aprendíveis, mais conhecidas pelo termo em inglês *word embeddings* [5, 52]. Estes modelos mapeiam as palavras de um determinado vocabulário para um vetor N-dimensional, onde cada dimensão reflete atributos diferentes para cada palavra. Nesse espaço de alta dimensão, palavras que compartilham significados próximos tendem também a ficar próximas umas das outras.

O desenvolvimento dos *word embeddings* para LM foi impulsionado através do uso de redes neurais artificiais ou em inglês *Artifical Neural Networks* (ANNs) motivado pelo fato de esses modelos serem regressores universais [80].

3.1.1 *Feedforward neural network*

O modelo mais tradicional de ANN é chamado de *Feedforward neural network* (FFNN). Essa rede é composta por camadas de nós, ou neurônios, interligados, onde cada nó realiza uma soma ponderada dos valores de entrada e que são, por fim, utilizados em uma função de transformação não linear para gerar uma saída. Esse valor é então propagado para a próxima camada ou para a saída final do modelo. O ajuste dos pesos de cada nó ocorre em resposta aos erros calculados por uma função de perda que compara a saída da rede com o valor esperado. Esse processo de ajuste nos pesos é chamado de *backpropagation* [33].

O primeiro trabalho a empregar as ANNs para LM foi apresentado em [5]. Os autores, inspirados nos tradicionais N-Gramas, propuseram uma rede FFNN que tinha como entrada uma janela de contexto que era uma sequência de palavras N cada uma represen-

tada em um vetor *one-hot*[1] com o tamanho do vocabulário. A primeira camada da rede, chamada de projeção, transforma o vetor *one-hot* de cada palavra em vetores contínuos de alta dimensão por meio de neurônios compartilhados entre os termos. Para combinar esses vetores em uma única representação, a segunda camada, chamada de camada oculta, realiza uma transformação não linear das palavras projetadas. Por fim, a camada final prevê a próxima palavra por meio de uma função *softmax* aplicada a todo o vocabulário, calculando a distribuição de probabilidade para essa previsão.

Os resultados obtidos pela nova abordagem apresentados em [5] foram superiores em escala em relação aos modelos N-Gramas. Esse trabalho, posteriormente, serviu de base para a criação do famoso modelo de *word embedding word2vec* [52, 54] que gera representações estático para cada palavra, considerando os diferentes aspectos sintáticos e semânticos sem a influência da janela de contexto.

3.1.2 *Recurrent Neural Networks*

A aplicação de FFNN para LM apresenta uma limitação importante devido ao uso de uma janela de contexto fixa, pois restringe sua capacidade de capturar dependências de longo alcance no texto. É possível superar essa limitação utilizando *Recurrent neural networks* (RNNs) [33] no lugar de FFNNs.

Mais detalhadamente, a *Recurrent Neural Network* (RNN) é uma versão auto-regressiva da ANN que permite estender a janela de contexto de maneira arbitrária [82], possibilitando, assim, a modelagem de atributos contextuais mais complexos entre as palavras. Nos MLNs baseados em RNNs, cada palavra é processada sequencialmente, e a projeção contínua da palavra atual é combinada com a projeção da palavra anterior por meio de um conjunto de pesos treináveis, chamado de estado oculto.

A primeira implementação de uma MLN utilizando RNN foi apresentada em [53] onde foram discutidos os principais desafios para o treinamento dessas redes, como efeito do *Backpropagation through Time* [33] e a dificuldade para modelar relações de longa distância entre os termos [6]. Com a evolução da área de DL, surgiram mecanismos para superar esses limitadores nas RNNs, como o Gated Recurrent Unit (GRU) [17] e o Long Short-

[1] vetor binário onde todos valores são 0 exceto o índice a ser representado

Term Memory (LSTM) [36], que são capazes de reter informações relevantes no vetor de contexto enquanto "esquecem" partes menos importantes.

As RNNs foram a base para o *Embeddings From Language Model*(ElMo) [64] um modelo de *word embedding* capaz de criar múltiplas representações para uma mesma palavra, considerando o contexto em que aparece. Além de MLNs, o uso de RNNs se tornou popular em diversas aplicações de NLP, pois consegue modelar relações contextuais utilizando os atributos gerados pelos modelos de *word embeddings* [23, 64, 46, 10].

3.1.3 Mecanismo de atenção e *Transformer*

Para algumas tarefas de NLP, como *tradução por máquina*, *sumarização de textos* e *legenda automática*, a saída do modelo é um novo texto. Para esses fins, as aplicações de DL costumam seguir a arquitetura de codificador-decodificador [60].

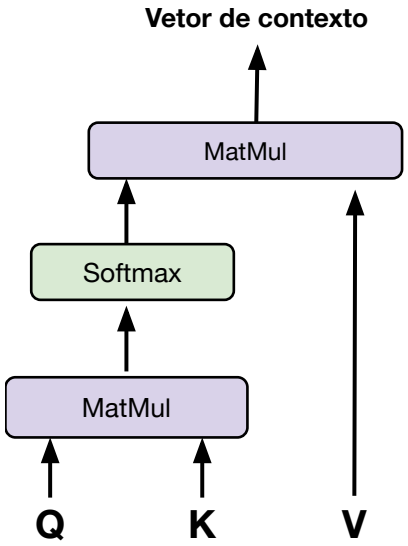
Nessa abordagem, uma rede codificadora transforma o texto original em uma representação vetorial de tamanho fixo, enquanto uma rede decodificadora utiliza esse vetor para gerar um texto de tamanho variável, adaptado ao contexto e conteúdo do texto de entrada.

Pesquisadores observaram que, além da avaliação de contexto possibilitada pelas RNNs, era importante considerar o alinhamento entre os elementos da entrada e da saída esperada. Por exemplo, imagine a tradução da frase em inglês "I ate a cheesecake" para o espanhol como "Comí una tarta de queso". Nesse caso, o mais relevante para gerar as palavras "tarta" e "queso" é o termo "cheesecake".

Essa observação foi a intuição inicial para a proposta dos mecanismos de atenção, primeiro apresentados em [3]. Esses mecanismos permitem à rede aprender a enfocar quais elementos da entrada são mais importantes para gerar a saída atual.

Conforme as operações apresentadas na Figura 3.1, o vetor de atenção pode ser descrito como um mapeamento entre uma consulta Q e dois conjuntos pareados K e V de chaves-valores, onde Q, K e V são todos vetores. Nos mecanismos de atenção, o estado oculto da entrada anterior é utilizado como Q, enquanto os estados ocultos da sequência de texto onde se busca o alinhamento com a entrada atual servem como V (valores), com K definido como $K = V$. O vetor K é particularmente útil para reforçar alinhamentos

Figura 3.1: Fluxograma das operações realizadas pelos mecanismos de atenção.



Fonte: Adaptado do artigo [83].

específicos dentro da sequência. A representação final do vetor de contexto é calculada ao multiplicar os valores de V pelas pontuações de atenção, obtidas a partir da multiplicação de Q e V, e normalizadas por uma função de atenção, como a *softmax*, que ajusta essas pontuações para valores entre 0 e 1.

Os mecanismos de atenção contribuíram para a melhoria de modelos de DL em tarefas generativas. Entretanto, os mecanismos de atenção marcaram a área de NLP através da introdução da arquitetura *transformers* [83] que propõe o uso exclusivo de mecanismos de atenção para criar um modelo auto-regressivo alternativo às RNNs. Para tal, é proposto um novo mecanismo chamado de *self-attention* que permite projetar a atenção para palavras presentes em uma mesma sentença. Dessa forma, os mecanismos de *self-attention* consideram quais palavras são mais importantes para construir o contexto atual. Vale ressaltar que o uso exclusivo de camadas de auto-atenção pode resultar na perda de informações sobre as relações temporais entre os elementos da sequência durante o aprendizado. Para contornar esse problema, a arquitetura *transformer* incorpora informações posicionais periódicas nas camadas iniciais de cada item da sequência, em um método conhecido como codificação posicional.

Em comparação com as RNNs, os *transformers* apresentam um tempo de treinamento significativamente menor, aproveitando de forma eficiente a capacidade de processamento paralelo em *hardwares* modernos, como GPUs [60]. Além disso, como os mecan-

ismos de *self-attention* ligam cada entrada às demais, é possível modelar o contexto de forma bidirecional.

O ganho de eficiência trazido pelos *transformers* marcou o início da era dos *Large Language Models* (LLMs), que são modelos de linguagem neural com bilhões de parâmetros treinados em enormes volumes de dados. Esses LLMs alcançaram o estado da arte em muitas tarefas de NLP [15].

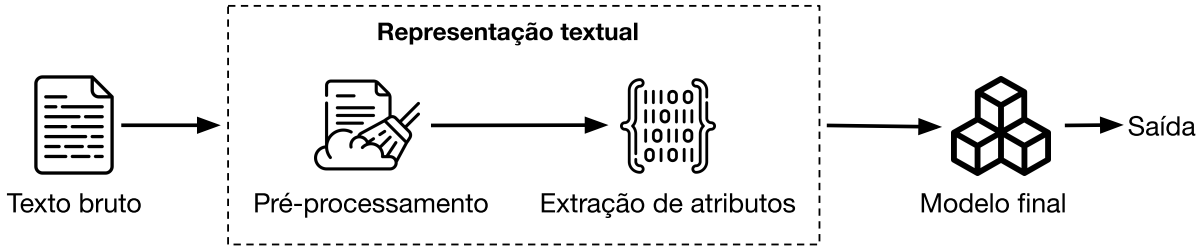
Em especial, o modelo *Bidirectional Encoder Representations for Transformers* (BERT) [25] é um modelo pré-treinado que oferece representações bidirecionais robustas para palavras e sentenças. Ele pode ser facilmente integrado a outros modelos e treinado para tarefas específicas por meio do processo chamado de *finetuning* [58].

Além disso, o BERT pode ser adaptado a diferentes domínios, como o de licitações públicas, usando dados não rotulados através de sua tarefa de aprendizado autosupervisionado chamada *Masked Language Modeling* (MLM). Diferente da tarefa original de (LM) descrita no início dessa seção, a MLM mascara palavras aleatórias do texto e o modelo é treinado para predizer as máscaras baseado no contexto circundante.

Hoje, diversos modelos BERT adaptados para variados domínios estão amplamente disponíveis na plataforma Hugging Face[2].

3.2 Aplicações de NLP

Figura 3.2: Pipeline geral para aplicações NLP baseadas em Machine Learning (ML).



Fonte: Autor.

[2] https://huggingface.co/models?pipeline_tag=fill-mask&sort=trending

As aplicações de NLP podem ser categorizadas de acordo com a tarefa que realizam. Nas tarefas supervisionadas, os modelos são treinados a partir de dados rotulados, aprendendo a distinguir classes previamente definidas. Por outro lado, nas tarefas não supervisionadas, não são utilizados rótulos e se busca identificar padrões latentes ou agrupamentos estruturais nos dados.

Independentemente da natureza da tarefa, tipicamente as aplicações de NLP baseadas em *Machine Learning* (ML) seguem uma *pipeline* similar [58]. O *pipeline* ilustrado na Figura 3.2 é composto por duas etapas principais: a representação do texto onde o texto bruto é normalizado e transformado para um formato estruturado e o modelo final que vai ser responsável pela saída final conforme a tarefa. A seguir, são discutidas mais detalhadamente cada uma das etapas do *pipeline* de uma aplicação de NLP.

3.2.1 Pré-Processamento

O pré-processamento é a etapa na qual são aplicadas transformações ao texto bruto, visando adequá-lo ao objetivo da aplicação. Técnicas como a remoção de termos irrelevantes, como *stopwords* e URLs, e a normalização textual, ajudam a reduzir a esparsidade dos dados textuais [10]. Além disso, as características textuais dos dados podem variar conforme o domínio da aplicação, envolvendo terminologias, jargões específicos e diferentes graus de formalidade.

Estudos como os de [4, 59] demonstraram que encontrar a combinação adequada de pré-processamento conforme o domínio específico contribui significativamente para o processo final de representação textual, além de melhorar o desempenho das aplicações.

3.2.2 Extração de atributos

A extração de atributos envolve converter o texto pré-processado em representações numéricas que capturam as relações textuais importantes para a tarefa alvo. Métodos clás-

sicos de extração de atributos como TF-IDF, *Bag-of-Words* e N-Gramas utilizam análises de frequência e coocorrência das palavras no conjunto de dados para gerar representações matriciais [58].

Como debatido na seção anterior, os *word embeddings* permitiram a criação de modelos pré-treinados cujas representações são utilizadas como método de extração de atributos. Entretanto, para aplicações de classificação e modelagem de tópicos, onde sentenças ou documentos completos são usados como entrada, surge o desafio de criar representações únicas para os documentos, capazes de capturar e descrever padrões entre eles [76].

Um *baseline* muito popular é utilizar a média da representação vetorial de cada palavra para formar uma única representação [69]. Em alternativa a essa abordagem, existem na literatura propostas de métodos de treinamento específicos para o desenvolvimento de modelos de *sentence embeddings*. Exemplos incluem métodos como SimCSE [31] e S-BERT [69], que adaptam os modelos de *word embeddings* para gerar representações únicas de sentenças através de uma nova etapa de treinamento.

Em resumo, a extração de atributos é fundamental para converter texto em representações numéricas eficazes para tarefas de NLP. Como a representação ideal pode variar conforme a tarefa, é essencial compreender bem os objetivos da aplicação para escolher os métodos de extração de atributos mais adequados.

3.2.3 Modelo Final

A etapa final envolve a escolha do modelo ou método que gerará a saída final para realizar a tarefa alvo da aplicação. Em NLP, há uma ampla variedade de algoritmos disponíveis, desde abordagens estatísticas como LDA [12], modelos clássicos de ML como Naive Bayes e Árvores de Decisão [58].

O uso de ANNs tem se tornado uma escolha predominante como modelo final. Novas camadas podem ser integradas facilmente a modelos *word embeddings* baseados em MLNs para serem treinadas em tarefas subjacentes como classificação e Modelagem de Tópicos (MT) em um processo conhecido como *finetuning*. Essa integração proporciona maior flexibilidade e precisão.

Por exemplo, nos trabalhos [59] e [43], redes LSTMs e convolucionais são, respec-

tivamente, utilizadas como modelos finais para distinguir diferentes classes de documentos a partir das representações geradas pelos *word embeddings*.

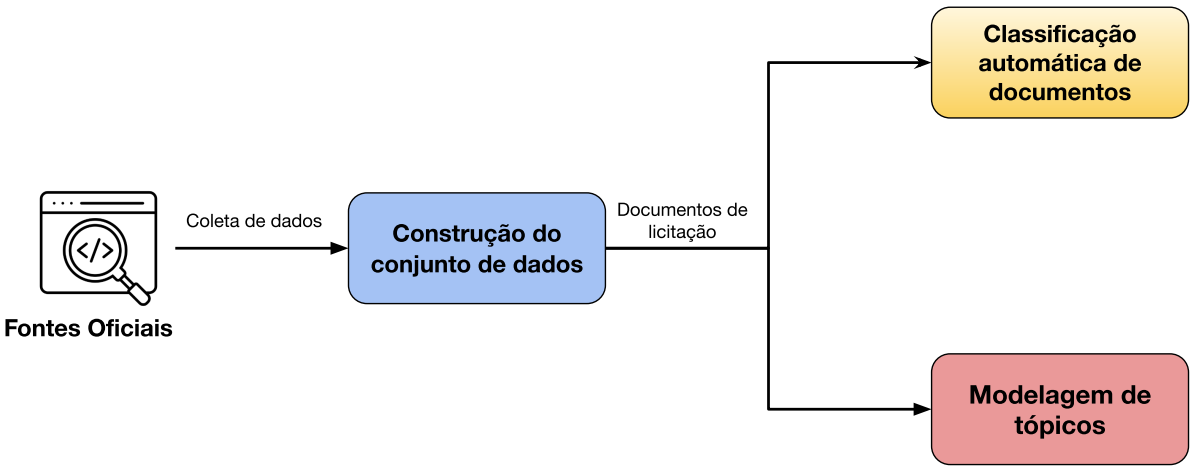
Capítulo 4

Metodologia

Considerando alcançar os objetivos delineados na Seção 1.2, esse capítulo descreve a metodologia adotada nesta dissertação. A metodologia geral da dissertação consiste em 3 etapas principais contempladas na Figura 4.1, quais sejam: Construção do conjunto de dados, proposta supervisionada para Classificação Automática de Documentos e proposta não supervisionada de Modelagem de Tópicos (MT).

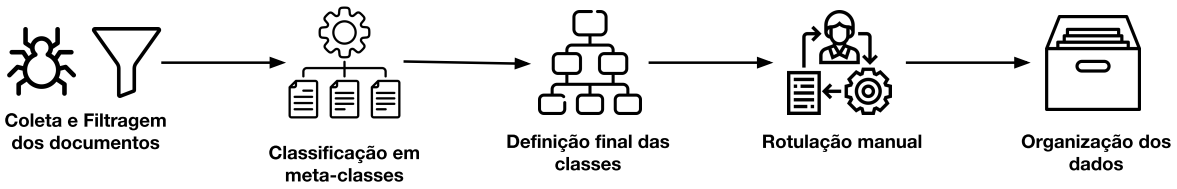
Metodologias específicas foram desenvolvidas para compor cada etapa. Em relação a isso, a Seção 4.1 apresenta o processo de construção de um conjunto de dados rotulado a partir de documentos de licitações públicas coletados dos portais de transparência de municípios do Estado de Minas Gerais. Em seguida, a Seção 4.2 propõe duas abordagens para a tarefa de classificação automática de documentos a partir dos rótulos propostos na etapa anterior. A primeira abordagem é baseada em DL, em particular, utilizando

Figura 4.1: Metodologia geral da dissertação.



Fonte: Autor.

Figura 4.2: Metodologia para construção do conjunto de dados de licitações. Fonte: Autor.



a arquitetura LSTM [36]. Já a segunda apresenta uma forma simples para classificação utilizando heurísticas criadas a partir da análise dos documentos.

Por fim, a Seção 4.3 trata da aplicação da metodologia BERTopic [34] para MT. O foco desta aplicação é avaliar a eficácia do nosso novo modelo pré-treinado LiBERT-SE [37], adaptado ao domínio de licitações públicas, para identificar padrões latentes e extrair informações relevantes dos documentos de licitação.

4.1 Construção do conjuntos de dados de licitações

Um processo licitatório envolve diferentes tipos de documentos de licitação, incluindo o texto do edital, anexos, errata, ata, projeto, memorial descritivo, dentre outros. Cada um desses documentos possui um formato específico que traz informações diferentes do processo de licitação. 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 analisar parte deles para ampliar a compreensão do tipo de documento sendo consultado) e para as máquinas (por ter que ser capazes de automatizar as tarefas de coleta, análise e processamento).

Esta seção tem por finalidade detalhar a metodologia desenvolvida para a extração, organização e processamento dos documentos de licitações públicas dos municípios do estado de Minas Gerais. A Figura 4.2 ilustra as principais etapas desse procedimento, que serão descritas em detalhes a seguir.

4.1.1 Extração e filtragem de documentos

A primeira etapa do processo de coleta dos dados é o levantamento dos portais que devem ter suas licitações coletadas a partir dos critérios de acessibilidade e qualidade do portal. Em seguida, a estrutura do *site* de cada município é analisada, visto que não há um padrão de disponibilização dos documentos comum à maioria deles.

Geralmente, portais de transparência de várias cidades hospedam *links* que fornecem acesso a documentos de licitação individuais. Para automatizar o processo de coleta, nossa proposta consiste no uso *web crawlers* desenvolvidos especificamente para os respectivos portais da transparência dos municípios. Os *web crawlers* simulam o comportamento de usuário de acessar os *links* dos documentos existentes para baixá-los e armazená-los.

Uma vez extraídos os documentos, realiza-se uma etapa de filtragem para separá-los em duas categorias: os processáveis e os não processáveis. A princípio, os documentos categorizados como processáveis são aqueles em formato *Portable Document Format* (PDF), que permitem a extração direta de seu conteúdo facilmente por meio de bibliotecas como o PDFPlumber[1] do Python. Já os documentos em outros formatos foram classificados como não processáveis.

Posteriormente, os documentos em formato PDF que não continham a informação textual pura — como documentos escaneados, corrompidos ou contendo apenas imagens — são reclassificados como não processáveis. Ao final, os documentos considerados processáveis são selecionados e seguem para as próximas etapas do fluxo desenvolvido.

4.1.2 Definição das classes dos documentos

Inicialmente, para os rótulos dos documentos, foram considerados os 56 tipos previstos nas leis de licitações (Lei nº 8666/1993, Lei nº 10520/2002 e Lei nº 14133/2021) para categorizar os documentos coletados de processos licitatórios.

Entretanto, com o apoio de especialistas membros do PCA e funcionários da Controladoria Geral do Estado de Minas Gerais (CGEMG) e do MPMG, realizamos uma

[1] https://github.com/jsvine/pdfplumber

segmentação hierárquica dos tipos de documentos mais relevantes para as aplicações, reduzindo o número final de rótulos para 13 classes organizadas em 4 meta-classes.

Mais detalhadamente, as meta-classes são uma categoria mais geral agregando documentos que são essenciais dentro do processo licitatório. Respectivamente, as características que definem os 4 rótulos dessa categoria são:

- **Adjudicação/Homologação**: Abrange documentos de adjudicação, homologação ou que apresentam ambas as informações em um mesmo documento;
- **Atas**: Cobre todos os documentos de ata disponíveis;
- **Editais**: Abrange documentos de edital e convites enviados em licitações da modalidade Convite;
- **Outros**: Contempla os arquivos pertencentes às demais classes de documentos. Esses documentos possuem funções mais especificas dentro do processo licitatório, por exemplo: erratas, anexos, contratos e memoriais descritivos.

Com a separação em meta-classes, um novo processo de análise manual de cada categoria foi realizado. Dessa forma, chegou-se à proposta de 13 diferentes tipos de documentos distribuídos dentro das meta-classes. A Tabela 4.1 apresenta os tipos de documentos identificados para cada meta-classe. Os rótulos finais foram utilizados como insumo para as tarefas seguintes, discutidas na próximas seções desse capítulo.

É importante destacar que não é viável considerar todos os tipos de documentos listados na lei de licitações como classes, pois foram identificados diversos casos em que tais documentos ocorrem muito raramente. Considerar esses documentos menos importantes sem o agrupamento pode impactar negativamente o desempenho de modelos para outras tarefas subsequentes de forma desnecessária.

4.1.3 Rotulação dos documentos

Para garantir um conjunto ouro a ser utilizado para serem insumos para treinamento e validação das aplicações propostas, todos os documentos coletados foram manualmente rotulados conforme as classes e categorias estabelecidas na Seção 4.1.2. Esse processo se deu

Tabela 4.1: Tabela das metas-classes e as classes identificadas nos documentos.

Meta-classe	**Classe**
Adjudicação/Homologação	Homologação/Adjudicação
Atas	Ata pregão presencial
	Ata registro preços
	Outras atas
	Ata dispensa Licitação
Editais	Edital
Outros	Errata
	Aditamento
	Outros
	Aviso
	Contrato
	Ratificação

a partir do alinhamento com os sete membros da equipe envolvida sobre as características e funções das classes definidas. Nesse alinhamento, os documentos foram divididos em lotes de tamanhos iguais para serem rotulados por cada um da equipe. A verificação da concordância entre os rotuladores foi opcional devido à simplicidade do processo de rotulagem e conhecimento do domínio por parte da equipe.

4.1.4 Armazenamento e organização dos dados

Para armazenar os dados relacionados a cada documento, foram utilizados arquivos no formato JSON[2], pois podem ser facilmente convertidos em dicionários. Essa escolha oferece várias vantagens, incluindo a capacidade de armazenar diferentes tipos de dados, proporcionando flexibilidade nas informações armazenadas. A Tabela 4.2 apresenta os campos contidos em cada arquivo JSON, juntamente com seus tipos de dados correspondentes e entradas de exemplo.

Em relação às informações armazenadas em cada campo, um código de identificação hexadecimal padronizado é armazenado no campo “file_id”, exclusivo para cada documento, com o nome original do documento no banco de dados de origem armazenado

[2]JavaScript Object Notation

Tabela 4.2: Dicionário de dados, contendo um exemplo de documento.

Campo	Tipo	Exemplo
`file_id`	*string*	"d2a0a04e5954c3095c1c1bbabcb5a107"
`original_name`	*string*	"d2a0a04e5954c3095c1c1bbabcb5a107.pdf"
`n_pages`	*int*	1
`text_content`	*array*	["\n \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"

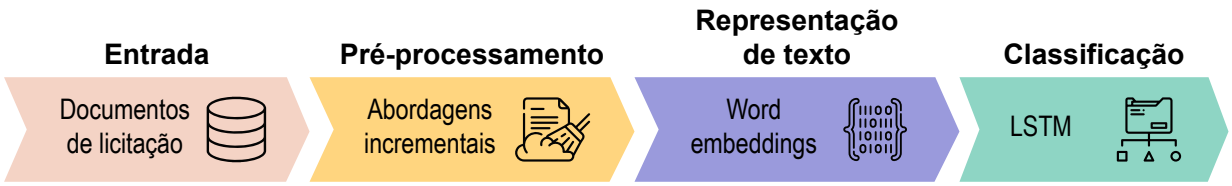
no "original_name" e o número de páginas inserido no campo "n_pages". Os campos "text_content" e "table_content" possuem cada um *array* dos textos e as tabelas no arquivo original do documento. Por fim, os campos "status", "city" e "text_preprocessed" armazenam o status do documento (processável ou não), a cidade de origem e o texto normalizado, respectivamente.

Para garantir a usabilidade do conjunto de dados para aplicações propostas na dissertação, cada arquivo JSON inclui os campos "meta_class" e "type_document". Esses campos armazenam a meta-classe e o tipo de documento obtidos do processo de rotulagem manual descrito na Seção 4.1.3.

4.2 Aplicação supervisionada: Classificação automática de documentos

Um dos objetivos definidos na Seção 1.1 foi a criação de uma aplicação supervisionada para a classificação automática de documentos. Mais especificamente, a classificação é uma tarefa supervisionada bem conhecida na área de aprendizado de máquina (ML). Ela é definida como o processo de categorizar ou rotular dados de entrada em classes ou

Figura 4.3: Metodologia para a classificação de tipo de documento.



Fonte: Extraído do artigo [10] que possui colaboração do autor.

categorias pré-definidas com base em suas características. Dado um conjunto de exemplos de treinamento rotulados, o objetivo de um modelo de classificação é aprender a associação entre os atributos dos exemplos e suas respectivas classes, de modo que possa prever corretamente a classe de novas amostras não vistas [60].

No contexto de licitações públicas, os documentos presentes possuem tipos de informações distintos sobre a licitação. Por exemplo, para a ingestão adequada dos dados, o processamento de uma Ata deve ser diferente do processamento de um Edital, pois, respectivamente, um se refere às decisões tomadas no processo e o outro aos itens e requisitos licitados. Além disso, o estado de Minas Gerais possui 853 municípios, o que torna a análise manual das licitações municipais uma tarefa extremamente desafiadora. Por isso, a automatização da classificação de documentos é essencial para otimizar o trabalho dos especialistas.

A tarefa de classificação também permite uma compreensão mais aprofundada das características dos tipos e meta-classes apresentados na Seção 4.1.2. As predições corretas dos modelos, assim como a análise de seus erros, permitem reavaliar as classes e compreender melhor a unicidade de cada uma das propostas.

Para atender ao desafio proposto, nesta seção apresentamos duas metodologias, uma para a classificação de tipo de documentos, e outra para a classificação das meta-classes. A primeira abordagem emprega métodos de classificação mais consolidadas, utilizando técnicas de DL. Para as meta-classes, por sua vez, palavras-chave foram identificadas, possibilitando a criação de uma heurística para sua classificação.

4.2.1 Classificação de tipo de documento

Uma vez que não foram identificados padrões simples como palavras-chave para identificar os tipos de documentos definidos por nós, optamos por adotar uma abordagem mais consolidada na literatura para a tarefa de classificação. Com isso, empregamos a combinação de técnicas de pré-processamento e modelos *word embeddings* estáticos (Seção 3.1.1) para a representação textual e realizamos a classificação dos tipos de documentos de licitação através de uma *Recurrent Neural Network* (Seção 3.1.2), especificamente, uma LSTM [36]. O uso dessas técnicas são particularmente úteis porque possuem um baixo custo computacional e conseguem lidar com a variação lexical e sintáticas, algo comum em textos extensos e sem padronização como os documentos de licitações.

A metodologia proposta é descrita na Figura 4.3 e tem como base a *pipeline* descrita na Seção 3.2. Em especial, os métodos de pré-processamento são avaliados em configurações incrementais, nas quais, a partir de um grupo base, os demais adicionam métodos de normalização mais complexos ao texto bruto. O objetivo é identificar qual combinação é mais útil para o contexto de licitações públicas.

No próximo capítulo de Métodos Propostos e Design Experimental serão detalhadas as técnicas de pré-processamento e modelos de *word embeddings* considerados para a tarefa de classificação de tipos de documentos.

4.2.2 Classificação das meta-classes

Para classificação somente das meta-classes, propomos uma abordagem heurística mais simples - porém elegante - em alternativa à abordagem anterior utilizando LSTM. A vantagem é que a proposta não requer modelos ou dados de treino para ser aplicada.

Durante o processo de rotulação (Seção 4.1.3) realizado pela equipe envolvida do PCA conforme descrito na Seção 4.1.3, foram observados certos padrões de palavras-chave que caracterizavam os documentos dentro das meta-classes. A partir desta constatação, desenvolvemos um meta-classificador heurístico em que conjuntos de palavras-chave identificadas são associados a cada meta-classe. A Tabela 4.3 apresenta amostras das palavras

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

Tabela 4.3: Meta-classes e palavras associadas.

Algorithm 1: Pseudo-código do Meta-classificador Heurístico

Input: Conjunto de documentos processáveis e um conjunto de Palavras-Chave para meta-classe
Output: A meta-classe predita de cada documento

```
1  begin
2      for Para documento do
3          Extraia o título e o conteúdo da primeira página do documento;
4          Declare countWordsTitle; // Ocorrência de palavra-chave no título, por metaclasse
5          Declare countWordsContent; // Ocorrência de palavra-chave no conteúdo da primeira
             página, por meta-classe
6          for cada meta-class do
7              Atualize countWordsTitle com o número de Palavras-Chave que ocorreram no
                 título;
8              Atualize countWordsContent com o número de Palavras-Chave que ocorreram no
                 conteúdo da primeira página;
9          end
10         if "Outros" meta-classe Palavras-Chave existir then
11             meta_classe ← "Outros"
12         end
13         if "Adjudicação/Aprovação" meta-classe Palavras-Chave existir then
14             meta_classe ← "Adjudicação/Aprovação"
15         end
16         Ordene countWordsTitle em ordem decrescente;
17         Ordene countWordsContent em crescente ;
18         if há uma ocorrência de palavra-chave no conteúdo da primeira página then
19             meta_classe ← meta-classe associada
20         end
21         meta_classe ← "Outros"
22     end
23 end
24 return Lista de documentos rotulados por meta-classe
```

identificadas para cada meta-classe, enquanto o Algoritmo 1 descreve o pseudocódigo do meta-classificador heurístico proposto.

Por ser simples e sem etapa de treinamento, a proposta pode ser facilmente implementada e avaliada. Ademais, ela pode ser incrementada com novas palavras-chave que venham a surgir a partir da análise contínua de novos documentos.

4.3 Aplicação não supervisionada: Modelagem de tópicos latentes em documentos de licitação

A modelagem de tópicos (MT) é uma tarefa tradicional de NLP não supervisionada, empregada para identificar automaticamente os principais temas ou tópicos em um grande conjunto de documentos. Essa técnica agrupa documentos com semântica semelhante, facilitando a compreensão do conteúdo e da estrutura de grandes volumes de texto.

A modelagem de tópicos tem sido amplamente utilizada para identificar temas predominantes em extensos corpus de documentos [18]. Os algoritmos tradicionais de modelagem de tópicos, como Latent Dirichlet Allocation (LDA) [7], podem nem sempre produzir resultados precisos e interpretáveis, especialmente ao lidar com dados de texto ruidosos. No contexto de dados de compras públicas, caracterizados por sua alta tecnicidade e jargão específico, processar e analisar os dados usando técnicas tradicionais pode ser desafiador.

Recentemente, tornou-se popular o uso de técnicas de agrupamento em combinação com as representações ricas fornecidas pelos modelos de *sentence embeddings* (Seção 3.2.2) para MT [26, 73]. As técnicas de agrupamento são métodos que buscam identificar grupos latentes nos dados, formando subconjuntos de amostras (*clusters*) com base em suas similaridades. Os métodos de modelagem de tópicos baseados em agrupamento partem do pressuposto de que os tópicos latentes correspondem aos *clusters*, utilizando modelos estatísticos para identificar as palavras mais representativas de cada grupo [18].

Considerando o objetivo definido na Seção 1.2, de desenvolver uma aplicação não supervisionada de MT para o domínio de licitações públicas, nesta seção apresentamos a metodologia proposta para essa tarefa. Para isso, empregamos o BERTopic [34], que utiliza representações geradas por modelos de *sentence embeddings* baseados em *transformers* [83], combinadas com técnicas de redução de dimensionalidade e agrupamento, para identificar tópicos latentes em conjuntos de documentos.

A motivação para o uso do BERTopic está em avaliar o desempenho do novo modelo, denominado LiBERT-SE, proposto pelo autor da dissertação em [37]. O novo modelo é baseado no *Bidirectional Encoder Representations from Transformers* (BERT) [25], amplamente reconhecido como estado da arte em representação textual.

4.3.1 Modelo LiBERT-SE

O LiBERT-SE é um modelo baseado nas representações disponibilizadas pelo *Large Language Model* (LLM) BERT mencionado na Seção 3.1.3. O LiBERT-SE foi proposto pelo autor da dissertação no âmbito do projeto PCA, onde adaptamos o modelo BERT em português brasileiro conhecido como BERTimbau [79] para o domínio de licitações públicas em português.

Seguindo as melhores práticas de adaptação de domínio descritas em [25], o LiBERT-SE foi treinado na MLM utilizando um conjunto de dados de 300.000 segmentos de diário oficial originados extraídos em [20]. Os segmentos abrangem artigos publicados nos diários oficiais de municípios de Minas Gerais, na temática de licitações. Esses artigos fornecem informações breves sobre outros documentos relacionados ao processo.

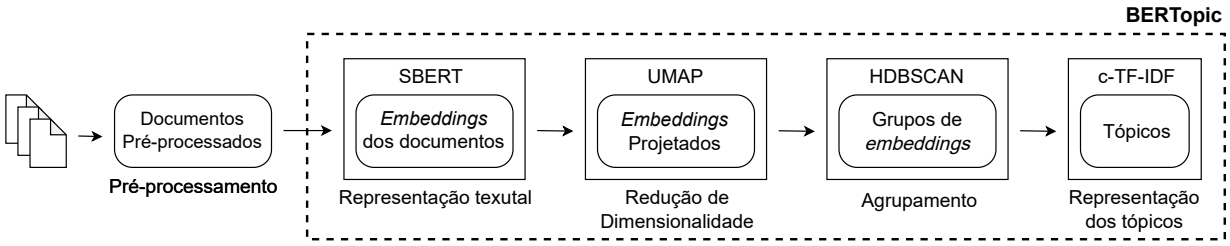
Para aprimorar o vocabulário do modelo, foi empregado um procedimento TF-IDF, que permite extrair os termos mais significativos dos segmentos de diário oficial. Além disso, foi incorporado manualmente jargões específicos de domínio de licitações públicas.

Além do aprendizado a nível de palavras, o LiBERT-SE foi treinado para gerar representações ricas de sentenças. Para isso, utilizou-se a técnica SimCSE [31], um método de aprendizado contrastivo aplicado a modelos baseados em *transformers*. Essa técnica trata as sentenças originais como amostras positivas e cria pares negativos ao adicionar ruídos aos *embeddings* das sentenças, utilizando a própria arquitetura dos *transformers*. O objetivo do SimCSE é minimizar a distância entre os pares positivos e maximizar a distância entre os pares negativos. Esse treinamento foi aplicado a um segundo conjunto de dados, composto por aproximadamente 300.000 segmentos extraídos de diários oficiais com foco no tema de licitações.

4.3.2 Metodologia BERTopic

A metodologia proposta pelo BERTopic [34], ilustrada na Figura 4.4, segue um fluxo similar à *pipeline* apresentada na Seção 3.2.Ela é composta por cinco etapas realizadas sequencialmente, as quais são detalhadas agora.

Figura 4.4: Visão da metodologia BERTopic.



Fonte: Extraído do artigo [37] escrito pelo autor.

Pré-processamento. Em primeiro lugar, como mencionado na Seção 3.2.1 o pré-processamento desempenha um papel fundamental nas tarefas de NLP, pois gera uma representação mais estruturada do texto e reduz o conjunto final de palavras (ou seja, o vocabulário) que será usado como entrada para modelos de aprendizado. De fato, aplicar uma etapa de pré-processamento específica para análise não supervisionada de aprendizado impacta diretamente a precisão dos modelos finais, reduzindo o ruído e melhorando a qualidade dos tópicos identificados.

Representação textual. Como discutido no Capítulo 3 mapear sentenças ou documentos em espaços vetoriais numéricos é um método eficiente para gerar representações de texto mais ricas em tarefas de NLP. Essas representações preservam informações semânticas, sintáticas e até contextuais nos documentos, levando a um melhor desempenho em modelos de aprendizagem que dependem da representação vetorial do texto como entrada. O BERTopic foi proposto para ser utilizado em conjunto com modelos Sentence-BERT (SBERT) [69] que são uma adaptação do modelo BERT para *sentence embeddings*. Os modelos SBERT utilizam uma camada adicional para agregar as representações das palavras em uma única representação. Um dos principais pontos fortes do SBERT é sua escalabilidade para processar grandes volumes de documentos, além de sua adaptabilidade, já que aproveita o conhecimento previamente adquirido por modelos BERT.

Redução da dimensionalidade. As representações geradas por modelos SBERT possuem alta dimensionalidade, o que dificulta o agrupamento semântico dos documentos devido à chamada "maldição da dimensionalidade". Para lidar com essa limitação, o BERTopic adiciona uma etapa de redução de dimensionalidade que prevê a aplicação do algoritmo *Uniform Manifold Approximation And Projection for Dimension Reduction* (UMAP) [51] para reduzir o tamanho das representações que serão utilizadas na etapa seguinte de agrupamento. O UMAP se destaca por preservar mais características locais e globais de dados

de alta dimensão projetados em dimensões menores. Utilizá-lo permite representar as informações de forma mais condensada usando menos dimensões, reduzindo o ruído causado por variáveis altamente correlacionadas.

Agrupamento. Para o agrupamento semântico dos vetores projetados dos documentos o BERTopic emprega o algoritmo HDBSCAN [13]. O HDBSCAN é conhecido por seu desempenho superior em comparação a outros algoritmos de agrupamento baseados em densidade, em termos de precisão e eficiência. O HDBSCAN pode lidar com agrupamentos de diferentes formas e tamanhos, tornando-o ideal para identificar tópicos dentro de um conjunto de documentos. O método pode identificar instâncias que não pertencem a nenhum grupo (ou seja, *outliers*), reduzindo o ruído na representação final do tópico. Como resultado, os tópicos detectados são mais coerentes e representativos da estrutura subjacente dos dados.

Representação dos tópicos. Como método de representação final dos tópicos, o BERTopic utiliza uma técnica própria denominada C-TF-IDF. Nesse método, cada grupo identificado na etapa de agrupamento é tratado como um único corpus, ao qual se aplica o TF-IDF. Esse procedimento permite calcular os termos mais importantes de cada cluster, gerando assim a distribuição dos tópicos. A técnica parte do pressuposto de que, ao extrair as palavras mais relevantes de cada cluster, obtém-se descrições representativas dos tópicos.

Capítulo 5

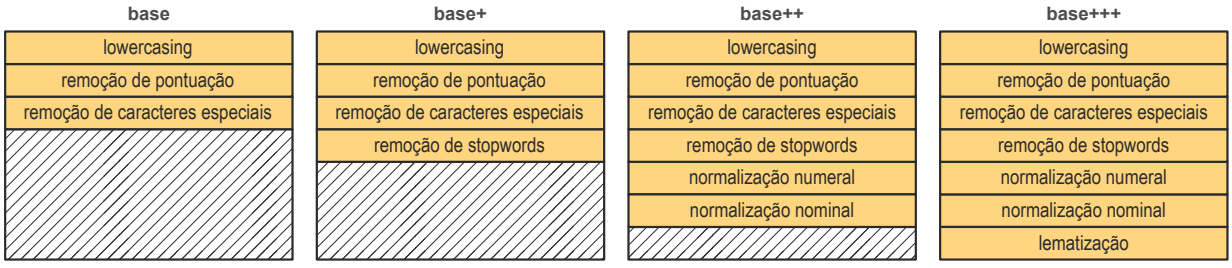
Métodos Propostos e Design experimental

O propósito deste capítulo é discutir as escolhas experimentais tomadas e os métodos escolhidos para a implementação correta e eficiente das aplicações propostas no capítulo anterior. Para tal, a Seção 5.1 aborda o pré-processamento escolhido para o trabalho. Em seguida, as Seções 4.2 e 4.3 abordam, respectivamente, os detalhes de implementação e configuração experimental para as aplicações de classificação automática de documentos de licitação pública e classificação apresentadas no capítulo anterior.

5.1 Técnicas de Pré-processamento avaliadas

Para encontrar a melhor estratégia de pré-processamento adequada ao domínio mais complexos como os de licitação pública é necessário avaliar diferentes técnicas. Para esse

Figura 5.1: Visão geral das quatro abordagens de pré-processamento.



Fonte: Extraído do artigo [10] que foi escrito com colaboração do autor.

estudo, são exploradas quatro abordagens distintas de pré-processamento para determinar a melhor estratégia, conforme ilustrado na Figura 5.1. Essas abordagens foram escolhidas com base nas técnicas de pré-processamento utilizadas em [59] e em análises preliminares dos dados coletados.

A primeira abordagem, chamada de "*base*", inclui três etapas principais: *lowercasing*, remoção de pontuação e remoção de caracteres especiais. As abordagens subsequentes expandem a abordagem "*base*", sendo que "*base+*" incorpora a remoção de *stopwords*, "*base++*" adiciona normalização numérica e nominal, e "*base+++*" inclui a lematização. Na sequência, as técnicas de pré-processamento avaliadas são descritas em detalhes.

Lowercasing. Esta técnica converte todos os caracteres do texto para minúsculas, ajudando a reduzir a variabilidade textual e a garantir que palavras idênticas não sejam tratadas como entidades diferentes devido a variações na capitalização. Além disso, contribui para a padronização dos dados textuais, evitando redundâncias, especialmente quando a mesma palavra pode aparecer em diferentes formatos (e.g., "ata" e "Ata"). No geral, é uma técnica de pré-processamento útil para dados textuais, especialmente quando o texto não é estruturado e pode ter variações de capitalização.

Remoção de pontuação. Consiste na eliminação de todos os sinais de pontuação do texto, como vírgulas, pontos, dois-pontos, ponto-e-vírgula, entre outros. Essa técnica simplifica os dados textuais e reduz o número de palavras únicas, eliminando sinais de pontuação que não possuem significado ou contexto significativo.

Remoção de caracteres especiais. Implica a exclusão de caracteres não alfanuméricos do texto, incluindo símbolos como *hashtags*, arrobas, cifrões e outros caracteres especiais que não sejam letras ou números. Esta técnica também simplifica o texto e reduz o número de palavras únicas, eliminando caracteres especiais que não possuem significado ou contexto significativo.

Remoção de *stopwords*. Envolve a eliminação de palavras comuns do texto, como artigos (e.g., "o", "a"), preposições (e.g., "em", "de", "para") e conjunções (e.g., "e", "ou"). O objetivo desta técnica é reduzir o ruído nos dados e melhorar a precisão das tarefas de análise subsequentes, removendo palavras que não possuem significado ou contexto significativo. Utilizamos uma lista de *stopwords* do português brasileiro, fornecida pela biblioteca NLTK.[1]

Normalização numeral. Consiste na conversão de todos os numerais no texto para um formato padrão. Isso pode incluir a substituição de dígitos por suas palavras correspon-

[1]NLTK:https://www.nltk.org/howto/portuguese_en.html#stopwords

dentes (e.g., "7" torna-se "sete") ou a substituição de todos os valores numéricos por um símbolo genérico (e.g., "1.000" torna-se "NUM"). O objetivo é reduzir a variabilidade dos dados textuais e simplificar as tarefas de análise subsequentes, tratando todos os valores numéricos de forma consistente. Neste estudo, seguindo o mesmo procedimento realizado em [59], todos os numerais foram substituídos por zero.

Normalização nominal. Envolve a conversão de nomes próprios no texto para um formato padrão, reduzindo o vocabulário e, consequentemente, a esparsidade das representações. A normalização desses nomes pode aumentar a precisão da classificação e garantir que informações relevantes sejam corretamente identificadas. Para essa tarefa, utilizamos um dicionário de nomes próprios comuns no Brasil, mapeando todos os nomes pelo termo *proper_name*, conforme a abordagem descrita em [59]. Além disso, os nomes das cidades presentes em cada documento foram removidos para evitar que informações de localização interfiram no modelo de classificação e prejudiquem seu desempenho.

Lematização. Esta técnica reduz as palavras do texto à sua forma básica ou de dicionário, conhecida como lema. Envolve identificar a raiz de uma palavra e mapear todas as suas formas flexionadas para o mesmo lema (e.g., "caminhar", "caminhou", "caminhando" são reduzidos a "caminhar"). O objetivo da lematização é diminuir a variabilidade dos dados e simplificar as tarefas de análise subsequentes, tratando todas as formas flexionadas de uma palavra como uma única entidade. Para esta tarefa, utilizamos a biblioteca spaCy para a língua portuguesa.[2]

5.2 Design Experimental: Classificação automática de documentos de licitação

Para o design experimental da tarefa de classificação automática, além das três abordagens de pré-processamento descritas na seção, foram considerados três métodos de *word embeddings* estáticos amplamente utilizados na literatura: *Global Vectors*(GloVe) [63], *word2Vec* [54] e *wang2Vec* [65]. O GloVe gera representações vetoriais de palavras com base em matrizes de coocorrência, enquanto o *word2Vec* e o *wang2Vec* utilizam técnicas

[2]spaCy:https://spacy.io/models/pt

baseadas em *FeedForward Neural Networks* (Seção 3.1.1). O *wang2Vec*, em particular, é uma extensão recente do *word2Vec*, que introduz um mecanismo de compartilhamento de pesos, melhorando a eficiência e escalabilidade do treinamento.

Todos os 3 modelos supracitados neste estudo foram obtidos do repositório *NILC-Embeddings*,[3] que oferece modelos treinados a partir de um grande corpus em português do Brasil e de Portugal. Para uma avaliação justa, todos os 3 modelos escolhidos foram configurados com 600 dimensões, e para o *word2Vec* e *wang2Vec* foi escolhida a abordagem *SKIP-GRAM*.

Para construção do classificador final, foi escolhida uma rede Long Short-Term Memory (LSTM) [36] com três camadas de recorrência. Adicionalmente, uma camada de *dropout* com probabilidade de 20% foi inserida para prevenir o *overfitting*. O modelo foi treinado com o otimizador Adam, utilizando uma taxa de aprendizado inicial de 0,001 e uma taxa de decaimento de 1e-6. O número de épocas de treinamento foi fixado em 8, e o tamanho do lote de treinamento foi definido como 64.

5.2.1 Configuração experimental

A tarefa de classificação automática de documentos propostas na Seção 4.2 avalia diferentes categorias de documentos, quais sejam: meta-classes e tipos de documento. Para a primeira categoria, foi proposta uma abordagem heurística e determinística de palavras-chave. Por consequência, sua avaliação pode ser realizada de forma mais simples, utilizando todo o conjunto de dados e comparando a resposta do classificador com os rótulos de referência.

Agora, ao tratarmos da segunda abordagem através de técnicas mais complexas de *Deep Learning* (DL), é necessária uma configuração experimental que consiga avaliar de forma robusta os impactos do pré-processamento, as diferenças textuais entre diferentes municípios e a divisão do conjunto em treino, teste e validação.

Com esse objetivo, foi empregada a validação cruzada estratificada para cada combinação de modelo e pré-processamento. A validação cruzada é uma técnica amplamente utilizada para avaliar o desempenho de modelos de aprendizado de máquina de forma ro-

[3]NILC-Embeddings: http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

busta. Nessa abordagem, o conjunto de dados é dividido em k subconjuntos, ou *folds*. A cada iteração, um dos subconjuntos é reservado como conjunto de teste, enquanto os demais são usados para treinar o modelo. Essa técnica reduz o viés na avaliação e oferece uma estimativa mais confiável do desempenho do modelo em dados não observados.

Particularmente, a validação cruzada estratificada divide os *folds* mantendo a mesma distribuição dos rótulos especificados. Neste estudo, dois modos de estratificação são utilizados: (1) Estratificação por classe, que considera a frequência das classes de tipos de documentos de licitação; e (2) Estratificação por classe e cidade, que leva em conta o fato de que documentos de licitação de uma mesma cidade podem compartilhar características similares. Assim, a divisão é feita de modo a garantir que a proporção de cada classe e cidade seja preservada em cada *fold*.

No total, foram realizados 25 experimentos: um para o meta-classificador e 24 para o classificador de tipos de documentos. A validação cruzada foi conduzida com 5 *folds*, uma vez que o uso de um número maior de *folds* não foi viável devido ao elevado tempo de processamento. Além disso, um subconjunto de validação foi reservado exclusivamente para o ajuste de hiperparâmetros dos modelos avaliados.

5.2.2 Métricas de avaliação

A avaliação do desempenho é uma etapa crucial no desenvolvimento de um classificador. A escolha de métricas apropriadas para avaliar a performance de um classificador é um desafio, pois diferentes métricas podem levar a resultados conflitantes e a escolha errada da métrica pode levar a conclusões equivocadas. Nesse contexto, a métrica *F1-Score* foi escolhida para a tarefa de classificação proposta na presente dissertação. A *F1-Score* é considerada uma das métricas mais populares para avaliar a performance de classificadores, sejam eles binários ou multi-classe [78].

Especificamente, a *F1-Score* leva em consideração a Precisão e a Revocação. A Precisão é a fração de verdadeiros positivos (VP) entre todos os positivos preditos (VP + FP) e a Revocação é a fração de verdadeiros positivos de todos os positivos reais (VP + FN). A F1-score representa a média harmônica entre as duas métricas supracitadas. De

maneira mais formal, cada uma das métricas é dada da seguinte forma:

$$Revocao = \frac{VP}{VP + FN}, \tag{5.1}$$

$$\text{Precisão} = \frac{VP}{VP + FP}, \tag{5.2}$$

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}}. \tag{5.3}$$

A *F1-Score* varia de 0 a 1, onde uma pontuação de 1 representa precisão e revocação perfeitas e uma pontuação de 0 representa desempenho ruim. Uma *F1-Score* alta indica que o modelo está funcionando bem em termos de precisão e recuperação, enquanto uma *F1-Score* baixa indica que o modelo tem baixo desempenho em pelo menos uma das duas outras métricas.

Em contrapartida, a acurácia é uma métrica mais simples e direta, que mede a proporção de previsões corretas do classificador. Por outro lado, o *F1-Score* também considera o tipo de erro cometido, penalizando mais fortemente os vieses de classificação. Essa característica é especialmente útil em contextos desbalanceados, onde o classificador tende a favorecer as classes majoritárias [67]. Portanto, o *F1-Score* se mostra mais adequado ao contexto desta dissertação, na qual estamos explorando o domínio de licitações públicas.

Para o trabalho, utilizamos 2 variações da métrica F1 no contexto multi-classe, quais sejam: *F1-Macro*, *F1-Weighted* [61]. A métrica *F1-Macro* calcula a *F1-Score* para cada classe separadamente e, em seguida, obtém a média não ponderada dessas pontuações. Em contraste, a métrica *F1-Weighted* calcula a média ponderada do *F1-Score* para cada classe, considerando o número de amostras em cada classe. Dessa forma, *F1-Weighted* reflete melhor o impacto das classes majoritárias no desempenho geral, enquanto *F1-Macro* dá igual importância a todas as classes, independentemente de seu tamanho no conjunto de dados.

5.3 Design Experimental: Modelagem de Tópicos

Nesta seção, discutimos os detalhes de implementação e configuração experimental para a aplicação de Modelagem de Tópicos (MT) proposta no capítulo anterior. Utilizamos

a versão do BERTopic [34] disponibilizada para a linguagem Python através da biblioteca de mesmo nome[4].

Para averiguar a performance do novo modelo LiBERT-SE descrito na Seção 4.3.1 na tarefa de MT, optamos por variar apenas os métodos de *sentence embeddings*. Dessa forma, em todos os experimentos, foi utilizada a configuração de pré-processamento 'base+', proposta na Seção 5.1.

5.3.1 Modelos de *sentence embddings* comparados

Para comparar a performance do modelo LiBERT-SE consideramos os principais modelos de *sentence embeddings* baseados na arquitetura *transformers* [83] disponibilizados na literatura. Dessa forma, avaliamos quatro modelos como *baselines*: *Multilingual Universal Sentence Encoder* (USE) [89], *Language-agnostic BERT Sentence Embedding* [30] (LaBSE), S-BERTimbau [69, 79] e *Portuguese Legislative Sentence Embedding* (Legal-BERTPTbr) [73]. Ademais, a Tabela 5.1 resume as informações dos modelos *baseline* empregados.

5.3.2 Configuração experimental

A configuração experimental é uma etapa fundamental na avaliação de tópicos gerados por métodos não supervisionados, que apresentam desafios únicos devido à falta de uma categorização previamente estabelecida. Por isso, é imprescindível utilizar abordagens robustas de validação para assegurar a precisão e a relevância dos resultados obtidos.

Para avaliação dos tópicos gerados adotamos duas abordagens populares, quais sejam: a avaliação interna, que resume os resultados em uma única pontuação de qualidade, com foco na coerência e diversidade dos tópicos; e a avaliação externa, que compara os resultados obtidos com uma "verdade básica" preexistente [29]. Essas abordagens fornecem

[4] https://maartengr.github.io/BERTopic/index.html

Tabela 5.1: Modelos de sentença escolhidos para este trabalho.

Modelo	Descrição
USE [89]	Projetado para gerar representações universais de sentenças que capturam o significado semântico do texto em vários idiomas. O modelo alavanca uma arquitetura de rede neural profunda com camadas transformadoras para codificar sentenças em vetores densos de comprimento fixo.
LaBSE [30]	Modelo de incorporação de sentenças multilíngue que suporta 109 idiomas, visando superar as limitações de modelos específicos de idioma. Ele é baseado na arquitetura popular BERT. O LaBSE incorpora objetivos de treinamento agnósticos a idioma para gerar representações de sentenças de alta qualidade que capturam informações semânticas multilíngues.
S-BERTimbau [69, 79]	Modelo pré-treinado adaptado para gerar *embeddings* de sentenças para o idioma português brasileiro. Ele se baseia na estratégia de ponderação proposta pelo *framework* Sentence-BERT (S-BERT).
LegalBERTPTbr [73]	É um modelo especializado de incorporação de sentenças, adaptado para documentos legais de comentários políticos brasileiros. Ele é treinado usando SimCSE [31] acoplado com BERTimbau. Os dados foram extraídos de dois projetos de emenda constitucional das plataformas oficiais.

uma visão ampla sobre a qualidade dos tópicos gerados, permitindo averiguar se os resultados produzidos são resultados coerentes e úteis.

5.3.2.1 Avaliação interna

O design experimental é uma etapa fundamental na avaliação de tópicos gerados por métodos não supervisionados, que apresenta desafios únicos devido à falta de uma categorização previamente estabelecida. Por isso, é imprescindível utilizar abordagens robustas de validação para assegurar a precisão e a relevância dos resultados obtidos.

Para os experimentos da avaliação interna, os resultados foram calculados como a média de 10 iterações, variando o número de tópicos entre 10 e 50, com um passo de 10. Essa estratégia permite avaliar o desempenho do BERTopic em diferentes quantidades de tópicos a fim de averiguar um número ideal no contexto de licitações públicas. Para comparação entre o desempenho dos modelos, empregamos duas métricas principais que serão descritas a seguir.

Em primeiro lugar, empregamos a coerência do tópico no inglês chamada de *topic*

coherence, uma medida baseada na medida de *Normalized pointwise mutual information* (NPMI) [8], que quantifica a associação entre palavras de um tópico. A métrica varia entre -1 e 1, onde quanto maior o valor, melhor a representação. A NPMI é conhecida por se aproximar da análise humana para avaliação da qualidade semântica dos tópicos [44]. Formalmente, a NPMI é definida pela seguinte fórmula:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \tag{5.4}$$

Onde w_i e w_j são duas palavras distintas sendo comparadas e $P(w_k)$ é a probabilidade marginal de w_k aparecer no corpus. Para a *topic conherence*, é calculada a média das NPMIs para todos os pares das palavras representativas identificadas, levando à seguinte fórmula:

$$\text{Topic Coherence} = \frac{1}{|W| \cdot (|W| - 1)} \sum_{i=1}^{|W|} \sum_{j=i+1}^{|W|} \text{NPMI}(w_i, w_j) \tag{5.5}$$

Sendo W o conjunto de palavras que descrevem um tópico.

A segunda métrica examinada é a diversidade do tópico, em inglês *topic diversity*, que avalia a porcentagem de palavras únicas entre todos os tópicos gerados. A diversidade do tópico é calculada como:

$$\text{Topic Diversity} = \frac{\text{Número de palavras únicas nos tópicos}}{\text{Número total de palavras nos tópicos}} \tag{5.6}$$

A diversidade varia de 0 a 1, onde próximo de 0 indica tópicos altamente redundantes e 1 representa tópicos mais diversos [26].

Por fim, para uma avaliação mais abrangente, combinamos as pontuações de *topic coherence* e *topic diversity*, ponderando a *normalized cohenrece* (NC) e a *normalized diversity* (ND) de acordo com sua importância relativa para a tarefa. A nomeada Weighted ponderada é dada por:

$$\text{Pontuação Final} = 0,8 \times NC + 0,2 \times ND, \tag{5.7}$$

Esta combinação convexa permite uma avaliação mais completa da modelagem de tópicos, garantindo que os tópicos gerados sejam não apenas coerentes, mas também diversos.

5.3.2.2 Avaliação externa

Na avaliação externa, o desempenho dos modelos foi analisado com base em dados adicionais não utilizados na modelagem de tópicos, como rótulos de classe previamente conhecidos. Para essa etapa, utilizou-se o modelo de *sentence embedding*, que apresentou os melhores resultados na avaliação interna, para ajustar os hiperparâmetros do BERTopic.

Dessa forma, foram ajustados os hiperparâmetros *nr_topics_list*, que contempla uma lista de números potenciais de tópicos, e *min_topic_sizes*, que define o tamanho mínimo de cada tópico. Os valores testados foram, respectivamente, (10, 13, 14, 15, 16, 17, 19, 20, auto) para *nr_topics_list* e (10, 20, 30, 40, 50, 60, 70, 80, 90, 100) para *min_topic_sizes*. O objetivo do ajuste é otimizar a geração de tópicos significativos e coerentes.

Além disso, o parâmetro *n_gram_ranges* foi definido como (1, 1), indicando que apenas palavras individuais (unigramas) foram consideradas no processo de modelagem de tópicos, focando nas palavras-chave mais importantes sem considerar combinações de palavras.

Por fim, os tópicos gerados pelo BERTopic com o modelo ajustado foram comparados com os rótulos de classe conhecidos, que serviram como referência para a avaliação. Ao realizar essa comparação e caracterização dos tópicos, buscou-se avaliar o grau de correspondência entre os tópicos gerados e os rótulos de classe previamente conhecidos, bem como a utilidade desses tópicos. Esse processo permite uma análise detalhada da capacidade do BERTopic de identificar padrões semânticos coerentes e de representar fielmente a estrutura e o conteúdo dos documentos de licitação.

Capítulo 6

Resultados obtidos

Esse capítulo discute os resultados obtidos pela dissertação considerando a metodologia apresentada no Capítulo 4 utilizando os métodos e experimentos propostos no Capítulo 5. Cada resultado foi analisado considerando o esclarecimento dos objetivos propostos na dissertação e a comparação das técnicas, com o intuito de aprofundar a compreensão do domínio de licitações públicas e de como abordá-lo de forma eficaz.

Diante disso, primeiro, na Seção 6.1, é apresentada a caracterização do conjunto de dados coletado no contexto da presente dissertação. Em seguida, as Seções 6.2 e 6.3 debatem os resultados obtidos na tarefa de classificação automática de documentos e modelagem de tópicos. Por fim, na Seção 6.4 são apresentados os artigos publicados envolvendo a dissertação e com contribuição do autor.

6.1 LipSet - O conjunto de dados de licitações

Nesta seção, são apresentados os resultados do processo de coleta dos documentos de licitação, incluindo a data e a origem da coleta. Além disso, é feita uma caracterização dos documentos com base em meta-classe, tipo e cidade de origem. As visualizações apresentadas avaliam a distribuição dos documentos em cada categoria e destacam as características relevantes presentes nos documentos de licitação coletados.

Tabela 6.1: Distribuição dos documentos PDF coletados em julho e dezembro de 2021.

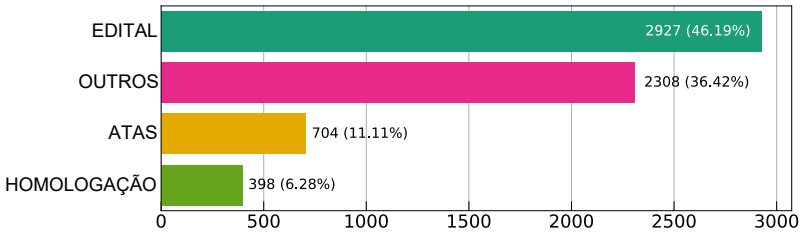
Mês da Coleta	#	**Cidade**	**Número de documentos**
Julho	1	Arantina	937
	2	Coqueiral	1,528
	3	Cristais	1,737
	4	Ijaci	455
	5	Itamarati de Minas	1,111
	6	Olaria	42
	7	Passa-Vinte	412
	8	Pirapetinga	1,108
	9	Ribeirão Vermelho	686
	10	São Bento Abade	275
Dezembro	1	Bias Fortes	159
	2	Cana Verde	402
	3	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**

6.1.1 Documentos coletados

Para coleta dos documentos de licitação pública, as fontes de dados utilizadas foram portais de transparência e/ou licitações de municípios de Minas Gerais. Cada portal foi analisado manualmente pela equipe do projeto PCA. Os portais foram ranqueados de acordo com a similaridade, disponibilidade de API para extração e complexidade para o desenvolvimento dos *web crawlers* responsáveis pela coleta. Em alguns casos, foi identificada a impossibilidade da coleta por conta do sistema de *CAPTCHA* implementados pelos portais.

Ao todo, foram coletados documentos de licitação de 18 municípios. O processo de coleta de dados ocorreu em dois períodos distintos: julho e dezembro de 2021. A Tabela 6.1 fornece uma visão geral do número total de arquivos coletados para cada município e o mês de coleta correspondente. O conjunto de dados consiste em um total de 9.761 documentos. Vale ressaltar que os arquivos em formato HTML e CSV contêm principalmente informações extraídas diretamente das páginas da web visitadas. Sua inclusão no conjunto de dados é impulsionada principalmente pelos aspectos estruturais das respectivas páginas da web, em vez de seu conteúdo. Portanto, para o propósito do desenvolvimento dos objetivos desta dissertação, esses arquivos foram desconsiderados e a análise é limitada a

Figura 6.1: Número de documentos por Meta-Classe.



Fonte: Extraído do artigo [71] que possui colaboração do autor.

documentos PDF (Seção 4.1).

6.1.2 Distribuição das Classes e Meta-Classes

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 marcado como *FAILED* no campo de status.

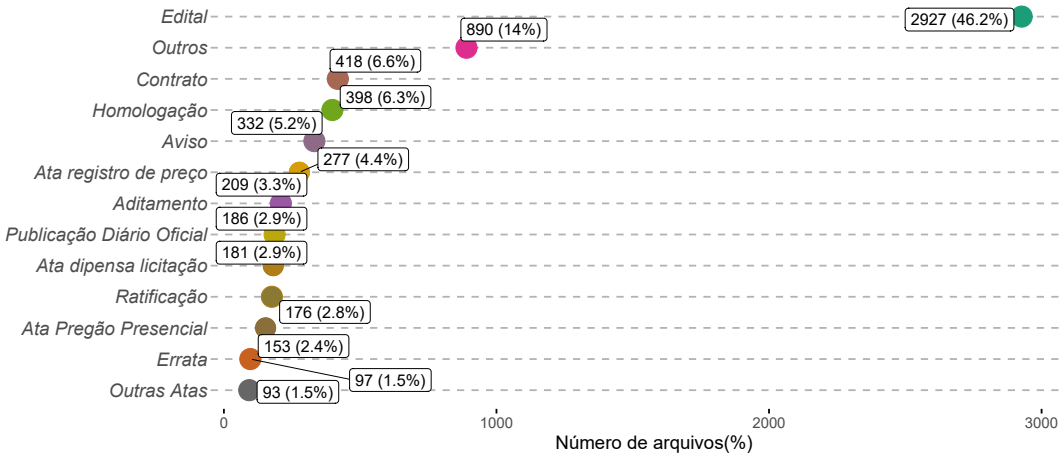
Além disso, 1.201 documentos são do tipo PDF, mas não foram rotulados, pois são imagens, plantas ou fotografias anexadas. Portanto, o LiPSet compreende 6.337 documentos de licitação pública classificados em uma das quatro meta-classes e 14 classes definidos na Seção 4.1.2.

A Figura 6.1 apresenta a distribuição do número de documentos por meta-classe, revelando que aproximadamente 83% dos documentos pertencem às meta-classes Edital (46,19%) e Outros (35,4%). As demais meta-classes Ata e Homologação, respectivamente, compõem as classes minoritárias, correspondendo a 11,11% e 6,28% dos documentos coletados.

Para os 13 tipos de documentos divididos nas 4 meta-classes, a Figura 6.2 apresenta a distribuição dos documentos. A imagem fornece uma visão geral das classes majoritárias e minoritárias em todos os documentos rastreados. A classe Edital corresponde a 46,2% dos documentos, enquanto a classe Outras Atas representa apenas 1,5%. As demais categorias relacionadas a atas, embora compreendam menos de 10% do conjunto de dados, possuem uma frequência duas a três vezes maior do que o tipo Outras Atas.

É notável que, no subconjunto de tipos pertencentes à meta-classe majoritária Out-

Figura 6.2: Distribuição dos documentos por classe.



Fonte: Extraído do artigo [71] que possui colaboração do autor.

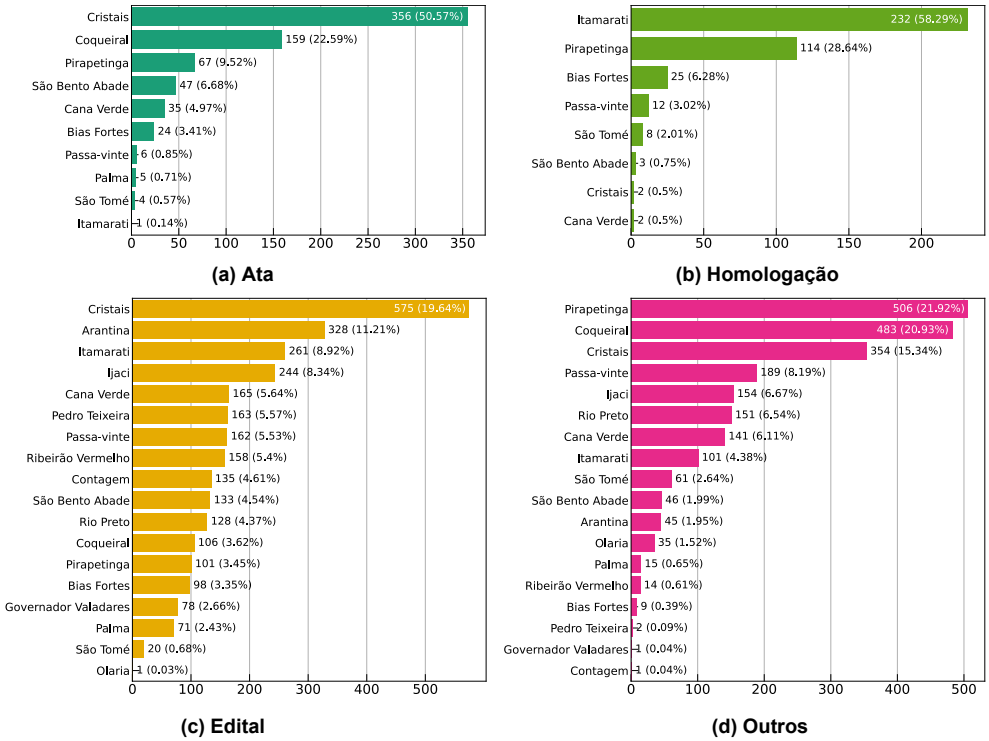
ros, os documentos estão distribuídos em quatro classes que, analisadas individualmente, apresentam menor representatividade no conjunto de dados. A maior delas é o tipo Outros, que compõe 14% dos documentos, enquanto a menor é a Errata, com apenas 153 documentos (2,4%). Por se tratar de uma categoria que abrange uma gama mais ampla de documentos, esse resultado indica que é importante reavaliar o tipo Outros, a fim de explorar melhor seu conteúdo e verificar se existe algum subtipo que faça sentido ser considerado na meta-classe Outros.

De maneira complementar, as Figuras 6.3a, 6.3b, 6.3c e 6.3d descrevem a distribuição de documentos para cada meta-classe em diferentes municípios. A Figura 6.4, por sua vez, apresenta a distribuição da categoria em tipos de documentos por município.

Observando a meta-classe Ata por município, dos 704 documentos, estes estão distribuídos por dez cidades, entretanto, cerca de 51% desses documentos são provenientes do município de Cristais (Figura 6.3a). Além disso, um padrão semelhante pode ser observado em relação à meta-classe Homologação, conforme mostrado na Figura 6.1b. Dos 398 documentos dessa meta-classe, aproximadamente 87% são oriundos das cidades de Itamarati (232 documentos) e Pirapetinga (114 documentos).

Os documentos da meta-classe/tipo majoritário Edital são maioria dentre os arquivos encontrados nos portais dos municípios. Exemplos incluem: Palma, em que corresponde a 78% dos documentos, Arantina (88%), Governador Valadares (99%), Pedro Teixeira (99%), Contagem (99%), Bias Fortes (63%), Ribeirão Vermelho (92%), São Bento Abade (58%) e Ijaci (61%). Já para a meta-classe Outros, também é notável sua pre-

Figura 6.3: Número de documentos por meta-classe considerando os municípios coletados.



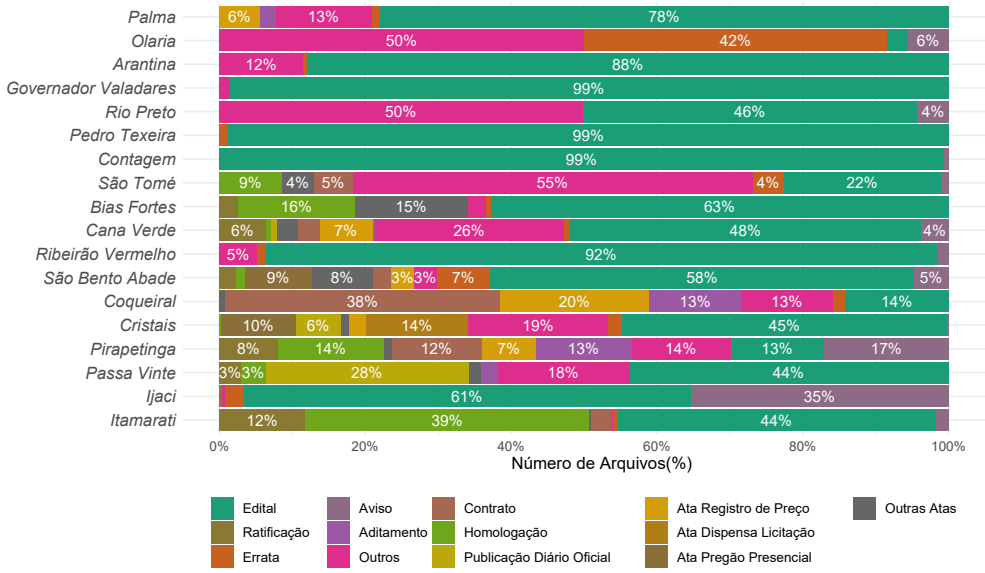
Fonte: Extraído do artigo [71] que possui colaboração do autor.

sença na maioria das cidades, como Palma (13%), Olaria (50%), Arantina (12%), Rio Preto (50%), São Tomé (55%), Cana Verde (26%), Coqueiral (13%), Cristais (19%), Pirapetinga (14%) e Passa-Vinte (18%). Entretanto, vale notar que apenas o tipo subjacente Outros é presente na maior parte das cidades.

Em geral, os municípios apresentaram grandes variações na distribuição dos documentos, com a maioria das cidades possuindo apenas um conjunto restrito de tipos. Por exemplo, no município de Arantina foram identificados apenas documentos dos tipos Editais e Outros. Em Governador Valadares, Pedro Teixeira e Contagem, 99% dos documentos coletados são editais. Em contraste, alguns tipos minoritários tiveram uma presença significativa em certos municípios, estando ausentes nos demais. Um exemplo é a cidade de Coqueiral, onde documentos do tipo Contrato representam 38% do total coletado, enquanto em Ijaci, o tipo Outras Atas compõe 35% dos documentos.

Esse cenário de desbalanceamento de amostras entre as diferentes meta-classes e tipos de documentos e falta de padrão entre os municípios representa um desafio signi-

Figura 6.4: Distribuição dos documento por tipo e cidade.



Fonte: Extraído do artigo [71] que possui colaboração do autor.

ficativo para esta dissertação, que propõe o uso de modelos de *Deep Learning* (DL) no contexto das licitações públicas. A baixa frequência de determinadas classes pode prejudicar o desempenho dos modelos, que normalmente dependem de um volume maior de dados para extrair padrões que definam cada classe de forma eficaz. Além disso, a classe "Outros" aparece com alta frequência nos documentos coletados, indicando que ela engloba uma ampla variedade de documentos com diferentes tipos de informações sobre o processo licitatório. Essa heterogeneidade pode também comprometer as aplicações de DL que utilizam o LipSet, pois dificulta a identificação precisa de padrões.

Outro ponto de destaque são as dificuldades enfrentadas durante a coleta dos documentos, uma vez que os portais de transparência frequentemente apresentam problemas de acesso, o que dificulta o acesso abrangente aos documentos por especialistas e pesquisadores.

6.1.3 Características dos Documentos de Licitações Públicas

As características levantadas nesta análise visam compreender a dimensão dos desafios presentes nos documentos de licitação ao considerar possíveis aplicações de *Natural Language Processing* (NLP). Nesse contexto, primeiro avalia-se a distribuição do número de páginas dos documentos, a fim de identificar o intervalo de número de páginas mais relevante para tais aplicações. Compreender a variação no tamanho dos documentos pode ajudar na definição de parâmetros mais adequados, otimizando o desempenho em termos de tempo de processamento.

De forma complementar, também foram analisadas a dispersão dos documentos coletados segundo o modelo LiBERT-SE descrito na Seção 4.3.1. Para isso, foram geradas visualizações em duas dimensões, baseadas nos *embeddings* produzidos pelo modelo. A redução da dimensionalidade foi realizada com o método de projeção *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) [51], amplamente reconhecido por preservar a estrutura esparsa das amostras originais ao projetá-las em dimensões menores. A implementação do UMAP usada neste experimento foi fornecida pela linguagem Python, por meio da biblioteca *umap-learn*[1]. Durante a execução do experimento, foram mantidas as configurações de hiperparâmetros padrão da biblioteca.

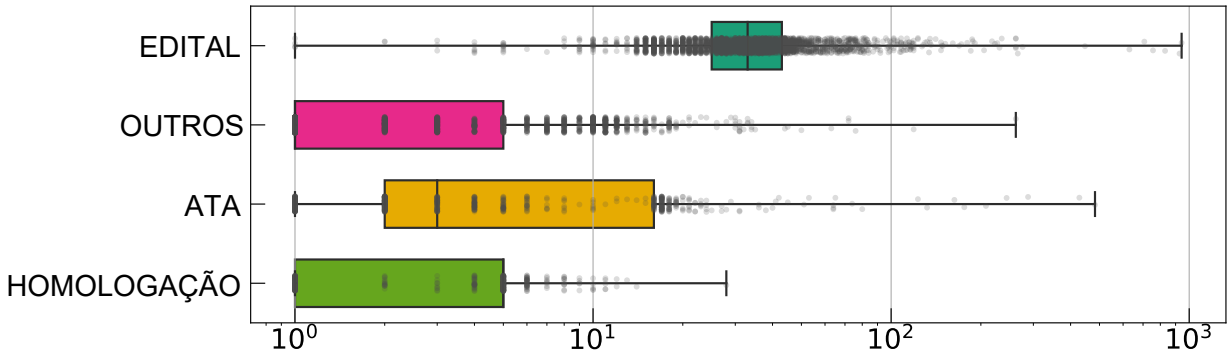
6.1.3.1 Avaliação do número de Páginas dos documentos

A Figura 6.5 apresenta os resultados da avaliação de número de páginas. Para a meta-classe Edital, observa-se um maior número de páginas com presença de muitos *outliers*. O número de páginas varia de 1, o menor valor, até 943, o maior valor registrado. O valor mais frequente de páginas para Edital é 27.

Para a meta-classe Outros, o número de páginas vai de 1, que é o valor mais recorrente, até 262 páginas, sendo esse considerado um *outlier*. Já na meta-classe Homologação, que possui uma distribuição próxima à meta-classe Outros, os documentos possuem de 1 a 28 páginas, sendo 5 o valor mais frequente. Por fim, a meta-classe Atas apresenta uma

[1] https://umap-learn.readthedocs.io/en/latest/

Figura 6.5: Distribuição do número de paginas dos documentos em escala logarítmica.



Fonte: Extraído do artigo [72] com colaboração do autor.

dispersão notável, também com muitos *outliers* presentes. O comprimento dos documentos varia de 1 até 483 páginas, sendo 1 o valor mais recorrente. É importante ressaltar que a meta-classe Ata é minoritária, com 704 amostras (11,11% do total de documentos).

Considerando todos os documentos, a maioria dos documentos não apresenta mais de 10 páginas. Optamos por considerar esse intervalo (de 1 a 10) por ser suficiente para a maior parte dos exemplos e por reduzir o custo computacional nas aplicações consideradas em relação a escolhas mais abrangentes.

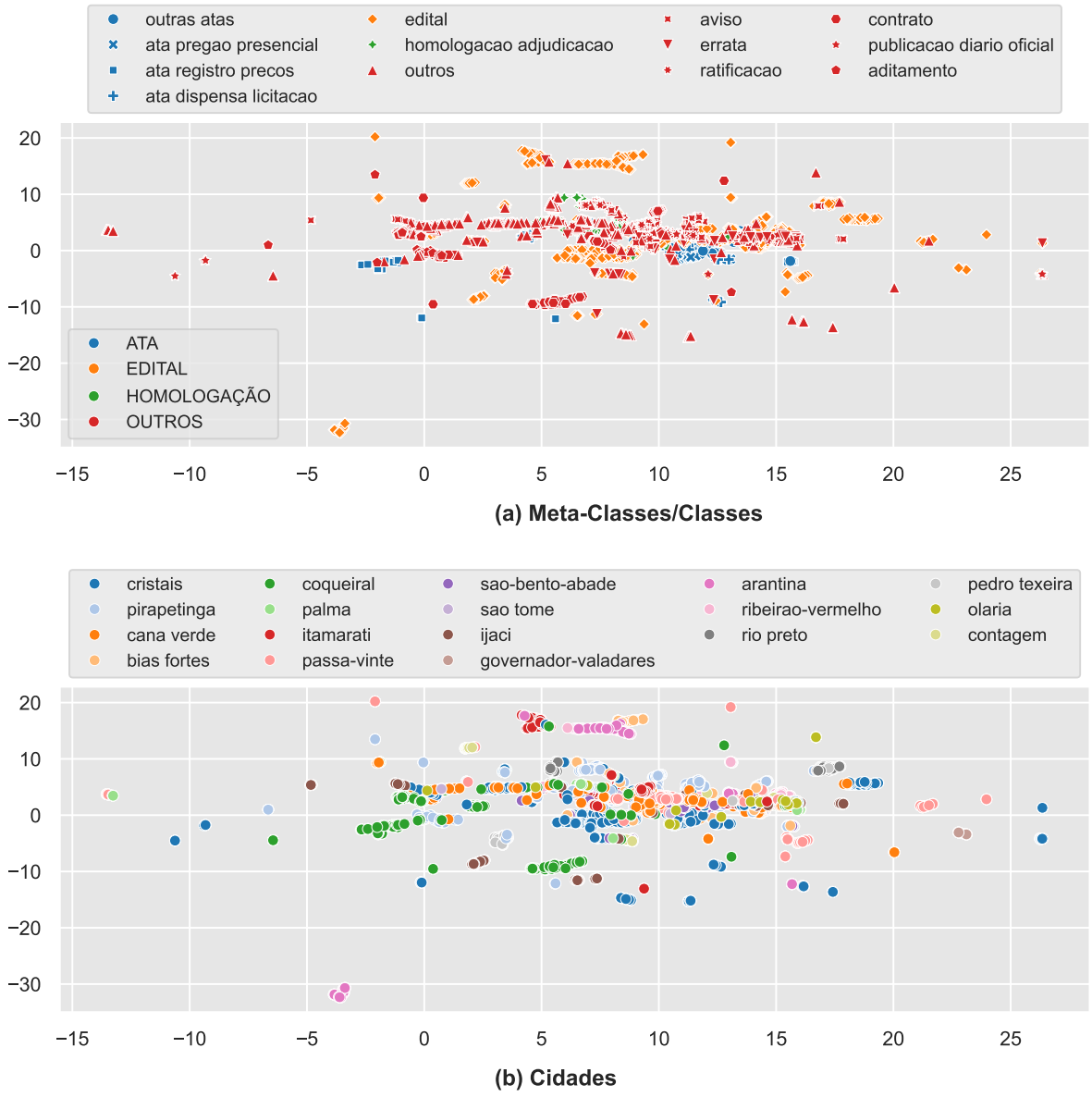
6.1.3.2 Avaliação da dispersão dos documentos

O espaço criado pelo *embddings* do modelo LiBERT-SE é ilustrado na Figura 6.6 onde os eixos x e y representam as projeções geradas pelo método UMAP dos atributos do modelo. Para destacar as variações dos documentos conforme suas categorias, a Imagem 6.6 a identifica as meta-classes através das cores e os tipos de documentos através dos símbolos. Agora, para explorar o impacto das diferenças textuais entre municípios, a Imagem 6.6 b exibe cada município por uma cor.

De maneira geral, a distribuição visualizada revela uma forte sobreposição entre os documentos, especialmente na região central. Esse núcleo reúne principalmente documentos das meta-classes Outros e Edital.

No entanto, também é possível observar a formação de grupos mais afastados, compostos predominantemente por um único tipo de documento de uma cidade específica. Por

Figura 6.6: Gráfico de dispersão dos documentos segundo o modelo LiBERT-SE.



Fonte: Autor.

exemplo, no intervalo $x = (-5, 0)$ e $y = (-30, -40)$, há um grupo de editais do município de Arantina. Além disso, entre os intervalos $x = (15, 20)$ e $y = (10, 0)$, três grupos próximos são formados principalmente por editais e avisos pertencentes à meta-classe Outros, com cada grupo sendo dominado por documentos de Rio Preto, Cristais e Ijaci, respectivamente. Isso sugere que o modelo identificou especificidades relacionadas às licitações

realizadas por esses municípios.

Mesmo entre os documentos que apresentam sobreposição próximos ao centro, é possível observar a formação de alguns grupos com características bem definidas. No intervalo $x = (5, 10)$, destaca-se um grupo de editais provenientes do município de Cristais. Um pouco abaixo, no mesmo intervalo, há um agrupamento de documentos do tipo Adiantamento, pertencente à meta-classe Outros, advindos do município de Coqueiral. Logo acima, outro grupo é formado por editais da cidade de Arantina. Esse fato é interessante, pois, conforme mencionado anteriormente, o município de Arantina já havia mostrado um agrupamento de editais mais afastado na distribuição. Isso sugere a existência de padrões consideravelmente distintos nos textos de edital, mesmo quando provenientes da mesma fonte.

Os documentos de Adjudicação/Homologação apresentaram forte sobreposição com documentos do tipo Outros próximos ao centro da distribuição, com vários municípios também sobrepostos nessa região. Isso pode indicar que os documentos rotulados como o tipo Outros podem representar uma nova modalidade para a meta-classe de Adjudicação/Homologação. Por fim, a meta-classe Outros demonstrou uma maior dispersão em relação às demais. Nessa classe, destaca-se a formação de grupos isolados entre si.

Para concluir, a visualização criada por meio do UMAP nos permitiu compreender de maneira mais clara a complexidade do domínio das licitações, que apresenta características bem específicas, levando em conta o município e o tipo de documento. Além disso, a análise realizada também possibilitou identificar possíveis melhorias que podem ser propostas para os tipos de documentos, ao examinar a sobreposição entre as categorias.

6.2 Resultados classificação automática de documentos

Na presente sessão são debatidos os resultados dos 25 experimentos propostos na Seção 5.2.1 para avaliação da tarefa de classificação automática de documentos de licitações públicas.

Tabela 6.2: Comparação das 24 configurações experimentais na classificação dos documentos de licitação, utilizando a LSTM.

		Estratificação por classe		Estratificação por classe e cidade	
Pré-processamento	**Word Embedding**	***F1-Macro***	***F1-Weighted***	***F1-Macro***	***F1-Weighted***
base	*word2vec*	0.863 ± 0.203	0.955 ± 0.068	0.893 ± 0.064	0.971 ± 0.015
	wang2vec	0.954 ± 0.003	0.984 ± 0.001	0.942 ± 0.044	0.983 ± 0.010
	GloVe	0.950 ± 0.017	0.985 ± 0.004	0.908 ± 0.128	0.973 ± 0.031
base+	*word2vec*	0.957 ± 0.007	0.986 ± 0.002	0.953 ± 0.003	0.985 ± 0.002
	wang2vec	0.960 ± 0.002	0.986 ± 0.001	0.964 ± 0.002	0.987 ± 0.002
	GloVe	0.971 ± 0.012	0.989 ± 0.004	0.969 ± 0.005	0.989 ± 0.003
base++	*word2vec*	0.937 ± 0.016	0.981 ± 0.003	0.932 ± 0.009	0.977 ± 0.003
	wang2vec	0.943 ± 0.016	0.981 ± 0.005	0.929 ± 0.045	0.976 ± 0.016
	GloVe	0.960 ± 0.016	0.986 ± 0.005	0.954 ± 0.009	0.985 ± 0.004
base+++	*word2vec*	0.925 ± 0.037	0.979 ± 0.009	0.914 ± 0.041	0.976 ± 0.012
	wang2vec	0.928 ± 0.024	0.977 ± 0.010	0.946 ± 0.022	0.983 ± 0.004
	GloVe	0.939 ± 0.025	0.981 ± 0.009	0.963 ± 0.005	0.987 ± 0.001

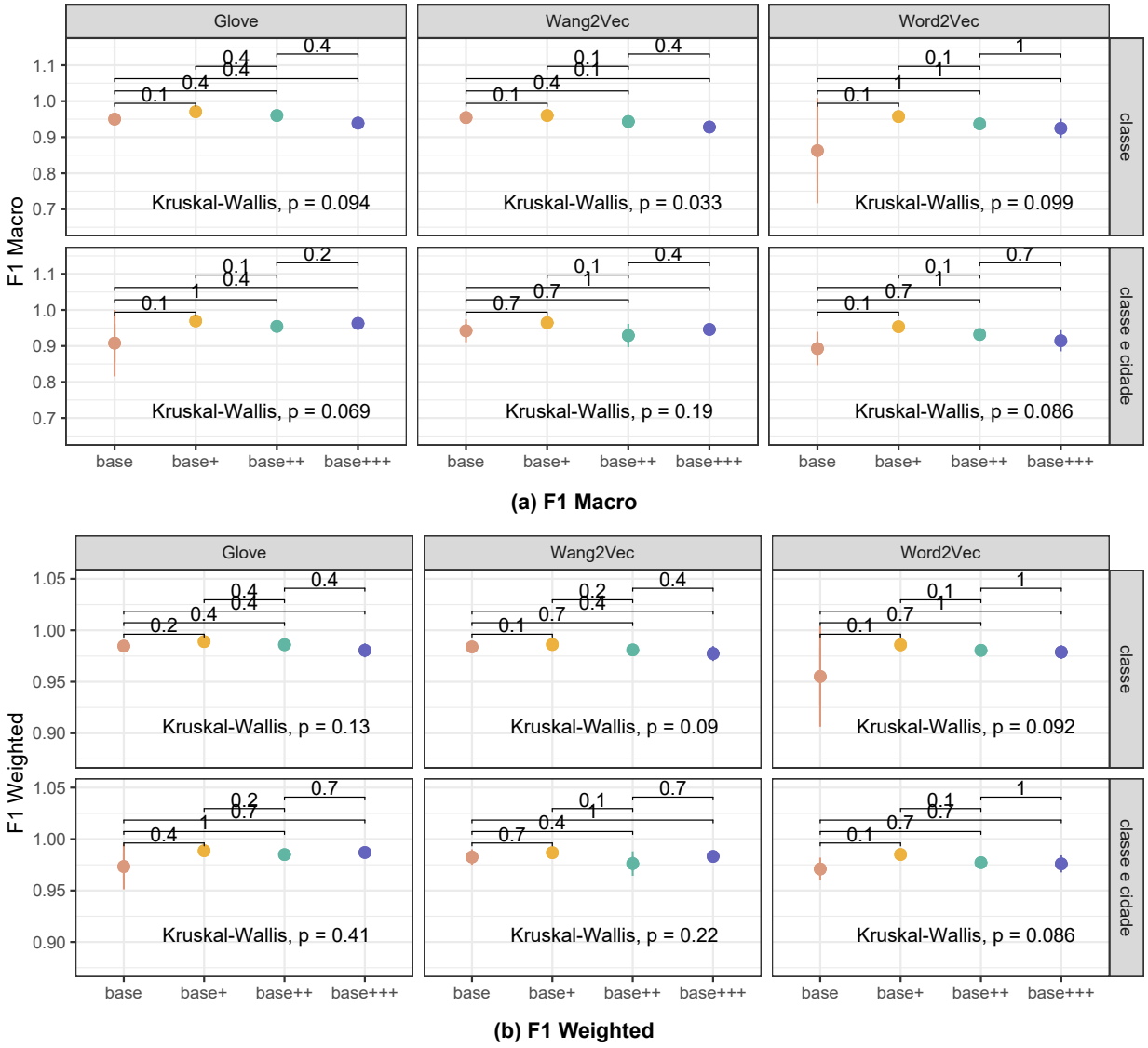
6.2.1 Classificação de Tipo de Documento

A Tabela 6.2 expõe os resultados obtidos para os 24 experimentos realizados com o modelo de classificação com LSTM [36]. Os resultados apresentados são bastante consistentes entre as diferentes combinações experimentais analisadas. O melhor desempenho foi obtido com a configuração que utilizou estratificação por classe, pré-processamento *base+*, e a representação textual baseada no modelo GloVe, alcançando 0,971 para *F1-Macro* e 0,989 para *F1-Weighted*.

Para avaliar se há diferença significativa entre as configurações experimentais, as Figuras 6.7a e 6.7b apresentam, respectivamente, os resultados da *F1-Macro* e*F1-Weighted*, acompanhados dos testes de Kruskal-Wallis e Wilcoxon pareado para cada configuração experimental. Ambos os testes são não paramétricos e adequados para comparar amostras independentes. Enquanto o teste de Wilcoxon pareado é utilizado para comparar duas amostras, o Kruskal-Wallis permite a comparação entre três ou mais amostras. A análise dos valores de *p-value* dos testes de Kruskal-Wallis nas Figuras 6.7a e 6.7b indica que não é possível rejeitar a hipótese nula de que as medianas da *F1-Macro* e *F1-Weighted* entre os experimentos são iguais, já que o *p-value* obtido é maior que 0,05 (ou seja, uma probabilidade superior a 5%). Assim, apesar dos bons resultados na tarefa, os testes sugerem que as diferenças observadas entre os experimentos podem ser atribuídas ao acaso.

Uma exceção notável é a comparação entre os pré-processamentos aplicados com o modelo *wang2vec* e a estratificação por classe na métrica *F1-Macro*, onde o *p-value* foi

Figura 6.7: Resultado da classificação das configurações experimentais conforme (a) *F1-Macro* e (b) *F1-Weighted*.

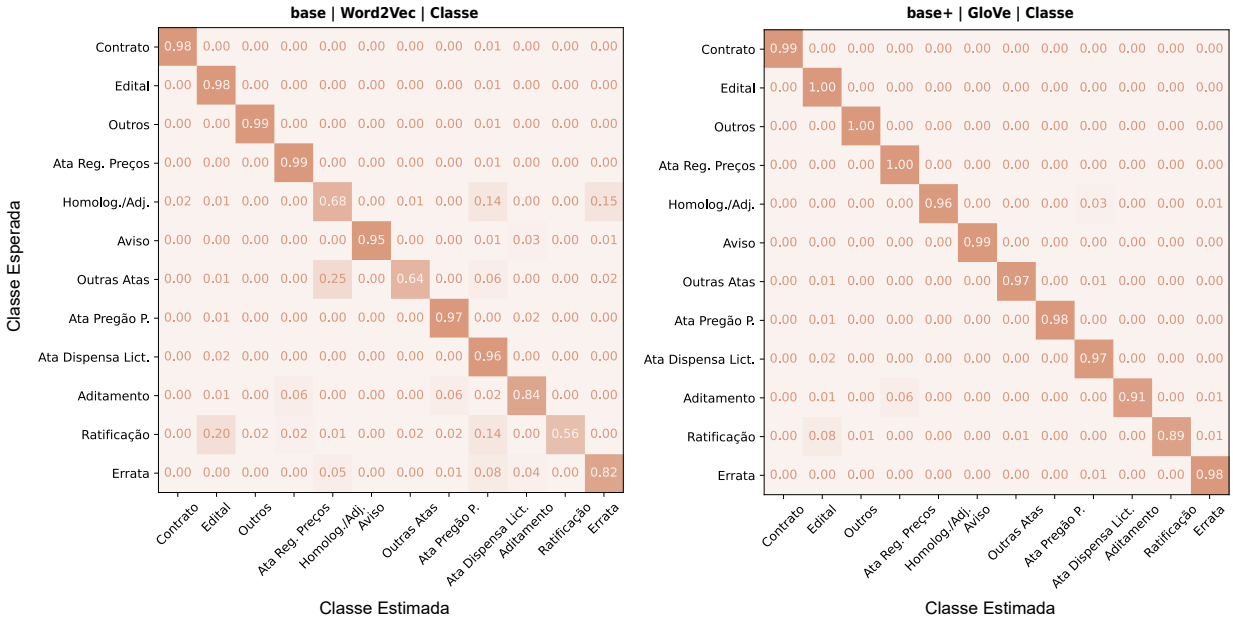


Fonte: Extraído do artigo [9] que possui colaboração do autor.

de 0,033, ligeiramente abaixo do limite de 0,05. No entanto, para a mesma configuração experimental na métrica *F1-Weighted*, o *p-value* apresentou um valor mais elevado, de 0,09. Por isso, conclui-se que não há evidências suficientes para afirmar que existe uma diferença significativa entre as diferentes abordagens de pré-processamento nas configurações experimentais avaliadas, tanto na *F1-Macro* quanto na *F1-Weighted*.

Quanto aos resultados do teste de Wilcoxon pareado, todos os *p-values* obtidos

Figura 6.8: Matriz de confusão da pior e melhor configuração experimental.



(a) Experimento com menor valor para F1-Macro **(b) Experimento com maior valor para F1-Macro**

Fonte: Extraído do artigo [10] que possui colaboração do autor.

foram superiores a 0,05, indicando que também não houve diferença significativa entre as configurações experimentais quando comparadas duas a duas. Esse resultado corrobora os achados do teste de Kruskal-Wallis, sugerindo que as variações nas combinações de pré-processamento e nas representações textuais não impactaram significativamente o desempenho da rede LSTM na classificação dos documentos de licitação.

Para facilitar a compreensão dos resultados, a Figura 6.8a apresenta duas matrizes de confusão referentes às 12 classes utilizadas na classificação dos documentos de licitação. A Figura 6.8a ilustra o experimento com o menor desempenho na métrica *F1-Macro*, cuja configuração experimental incluiu estratificação apenas por classe, pré-processamento do tipo *base* e a representação textual com o modelo *word2vec*. Os dados indicam que a classe mais desafiadora de classificar foi a *ratificação*, onde 20% dos documentos dessa classe foram erroneamente classificados como *edital*, e 14% como *ata de pregão presencial*. Também se destaca para a classe Outras Atas onde 25% dos exemplos foram preditos como sendo do tipo Outros.

Em contraste, a Figura 6.8b, que corresponde ao experimento com o melhor desempenho na métrica *F1-Macro*, usa a configuração com estratificação por tipo, e o pré-processamento *base+* e a representação textual com o modelo GloVe. É notável como a

maior parte dos erros cometidos pelo classificador foram corrigidos nessa configuração. Entretanto, o tipo Ratificação ainda apresenta o maior número de erros de classificação. Por outro lado, a taxa de confusão com Edital foi reduzida para 8%, indicando uma melhora significativa no desempenho geral.

Isso nos indica que, apesar do bom resultado das métricas de F1 considerando o cenário prático, existe um problema significativo de representação e sobreposição entre as classes definidas. Especificamente, examinando as duas matrizes de correlação é possível ponderar que há um problema na representação das classes Ratificação e Edital devido a serem os tipos onde o modelo mais erra. Esse resultado reforça a análise anterior feita na Seção 6.1.3.2, onde foi discutida a sobreposição dos tipos Edital e Ratificação observada pelo modelo LiBERT-SE.

Assim, é importante melhor analisar de maneira mais profunda como os textos estão representados nas diferentes classes dos documentos de licitação. Além disso, é importante averiguar o uso da LSTM que, embora seja eficaz para capturar dependências em sequências longas, apresenta limitações no manejo de entradas extensas, o que pode comprometer a performance na predição [90].

6.2.2 Classificação Meta-Classes

Os resultados obtidos pelo meta-classificador, apresentados na matriz de confusão da Figura 6.9, superaram as expectativas iniciais. Apesar de sua simplicidade, o modelo alcançou 91% na *F1-Macro*.

O desempenho geral do método foi promissor para todas as meta-classes, com taxas de acerto variando entre 77% e 98%. Um destaque é a meta-classe Atas, que apresentou a maior taxa de erro, com 23% dos documentos sendo incorretamente classificados como pertencentes à meta-classe Outros.

Em contrapartida, as meta-classes Homologação e Edital alcançaram taxas de acerto de 98% e 92%, respectivamente. Esses resultados comprovam a eficácia do meta-classificador em identificar documentos com alta precisão dentro dessas meta-classes.

No entanto, para as meta-classes Outros e Atas, os erros observados indicam que ainda há espaço para melhoria na classificação, especialmente considerando a confusão

Figura 6.9: Matriz de confusão da meta-classificação.

Fonte: Extraído do artigo [71] que possui colaboração do autor.

entre essas duas categorias. As melhorias podem ser implementadas através de um método mais complexo, como foi o caso utilizado para os tipos de documento ou no levantamento de novas palavras-chave.

O resultado apresentado pelo meta-classificador mostra que, apesar do desbalanceamento e dos grupos sobrepostos de documentos, é possível determinar, em um nível mais alto, as características de cada meta-classe.

6.3 Resultados Modelagem de Tópicos

Esta seção apresenta os resultados do experimentos do estudo não supervisionado no domínio de licitações públicas através da tarefa de Modelagem de Tópicos. Seguindo o design experimental exposto na Seção 5.3.

Tabela 6.3: Comparação de modelos de sentenças com base em métricas nas avaliação interna. O melhor resultado para cada métrica está sublinhado.

	topic coherence	***topic diversity***	***weighted score***
USE	0.073 ± 0.006	0.814 ± 0.008	0.414 ± 0.022
LaBSE	0.081 ± 0.007	0.833 ± 0.008	0.457 ± 0.026
S-BERTimbau	0.108 ± 0.005	0.815 ± 0.007	0.537 ± 0.021
LegalBERTPTbr	0.034 ± 0.008	0.737 ± 0.012	0.229 ± 0.034
LiBERT-SE	<u>0.139 ± 0.008</u>	0.843 ± 0.010	<u>0.664 ± 0.032</u>
LiBERT-SE + SimCSE	0.036 ± 0.006	0.842 ± 0.011	0.307 ± 0.024

6.3.1 Avaliação interna

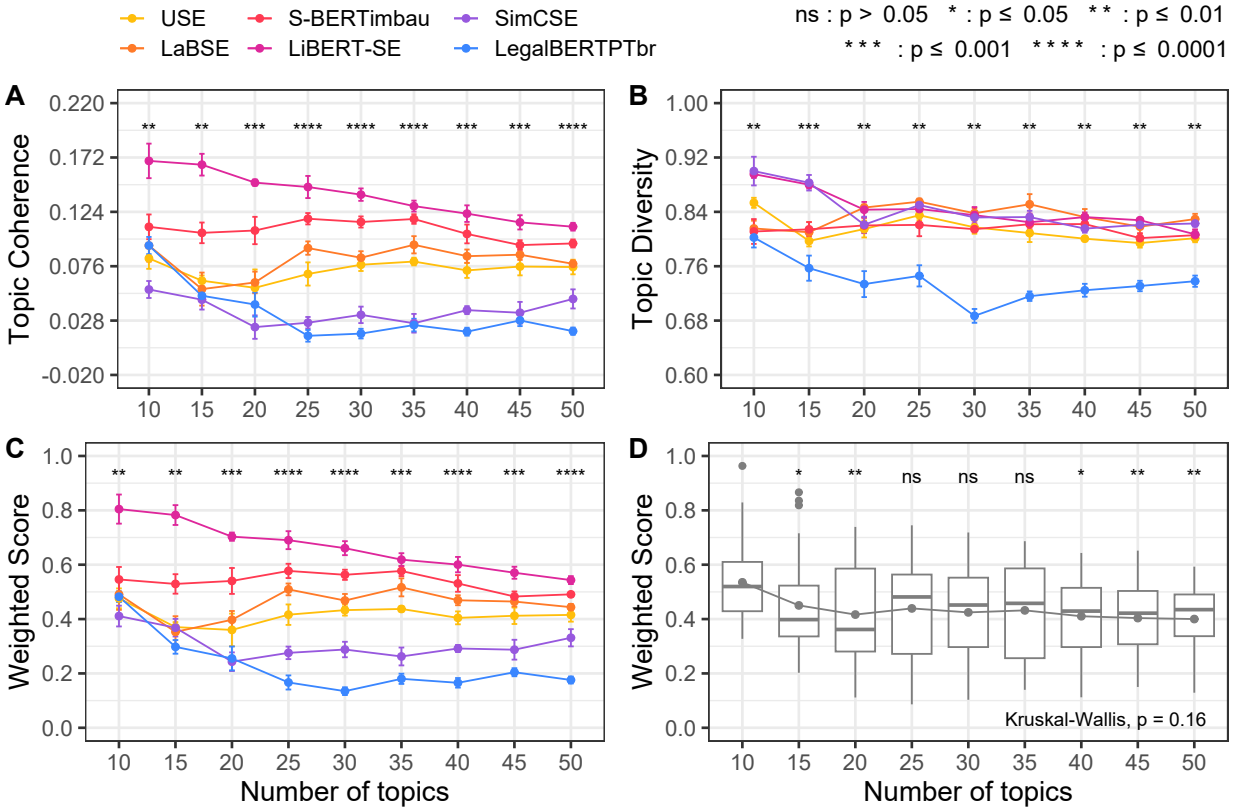
Os resultados da avaliação interna são apresentados na Tabela 6.3, fornecendo uma análise comparativa do desempenho dos modelos avaliados. Os resultados indicam que o modelo LiBERT-SE, proposto no contexto da presente dissertação, supera consistentemente os outros cinco modelos na maioria das métricas de avaliação interna. Essas descobertas ressaltam a eficácia do LiBERT-SE na representação de documentos de licitações públicas.

Além disso, foi observado que o uso da técnica SIMCSE [31] levou a uma considerável queda na métrica de *topic coherence*. De mesmo modo, o modelo LegalBERT-PTBR que também utiliza o SIMCSE teve um resultado igualmente baixo para a *topic coherence*. Ambos possuem os piores resultados entre os experimentos realizados.

A Figura 6.10 complementa a avaliação dos resultados, descrevendo a distribuição da *topic coherence* (**A**), *topic diversity* do tópico (**B**), *weighted score* (**C**), variando entre os diferentes modelos e números de tópicos. Além disso, a Figura 6.10 (**D**) mostra a distribuição agrupada da *weighted score*, fornecendo uma visão geral concisa do desempenho de cada modelo em todo o intervalo do número de tópicos. Os testes de Kruskal-Wallis e de Wilcoxon pareado foram conduzidos para determinar a significância estatística das diferenças observadas. Esses testes nos permitem verificar se há diferenças significativas entre as configurações consideradas, fornecendo informações adicionais sobre seu desempenho relativo.

As Figuras 6.10(**A**–**C**) corroboram as descobertas apresentadas na Tabela 6.3, reforçando ainda mais o desempenho superior do LiBERT-SE em comparação com os outros modelos na maioria das métricas avaliadas. O modelo alcançou pontuações mais altas que os demais em termos de *topic coherence* em diferentes números de tópicos. No entanto,

Figura 6.10: Avaliação interna. (**A**–**C**) Distribuição das métricas de avaliação interna, variando entre os diferentes modelos e número de tópicos. O teste de Kruskal-Wallis é aplicado para a comparação das médias dos modelos. (**D**) Distribuição agrupada da *weighted score*.

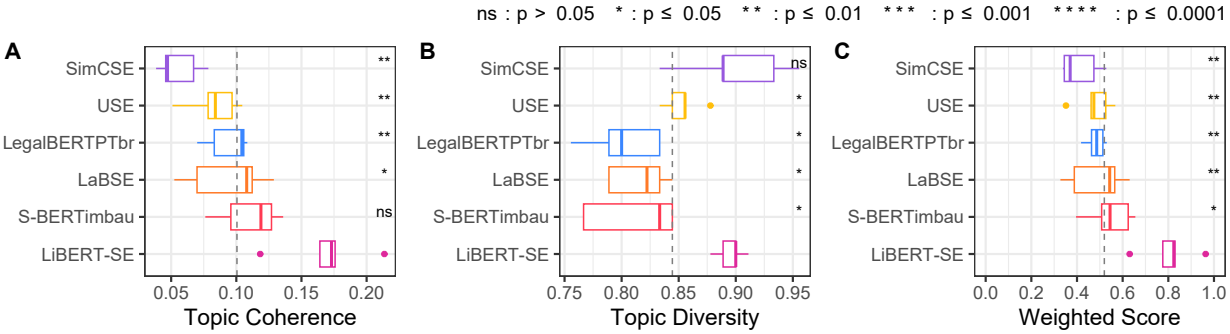


Fonte: Extraído do artigo [37] escrito pelo autor.

em relação à *topic diversity*, o desempenho do modelo foi comparável aos modelos SimCSE e LaBSE. Portanto, embora o modelo tenha se destacado na *topic coherence*, ele também manteve um alto nível de *topic diversity*, o que é crucial para gerar tópicos significativos e variados.

A Figura 6.10(**D**) fornece uma visão geral abrangente do desempenho dos modelos, mostrando a distribuição agrupada da *weighted score* em todo o intervalo de tópicos. Em média, considerar dez tópicos resultou em um alto desempenho geral. No entanto, conforme o número de tópicos aumentou, o desempenho variou entre os modelos e diminuiu ligeiramente. Essa descoberta sugere que o número de tópicos pode impactar o desempenho dos modelos. Enquanto um número menor de tópicos tende a produzir um desempenho geral mais alto, aumentar o número de tópicos introduz mais granularidade, mas pode

Figura 6.11: Comparativo de (**A**) *topic coherence*, (**B**) *topic diversity* e (**C**) pontuação Ponderado para cada modelo ao considerar dez tópicos. A linha tracejada vertical representa o valor mediano para cada métrica de avaliação interna.



Fonte: Extraído do artigo [37] escrito pelo autor.

resultar em uma ligeira diminuição no desempenho.

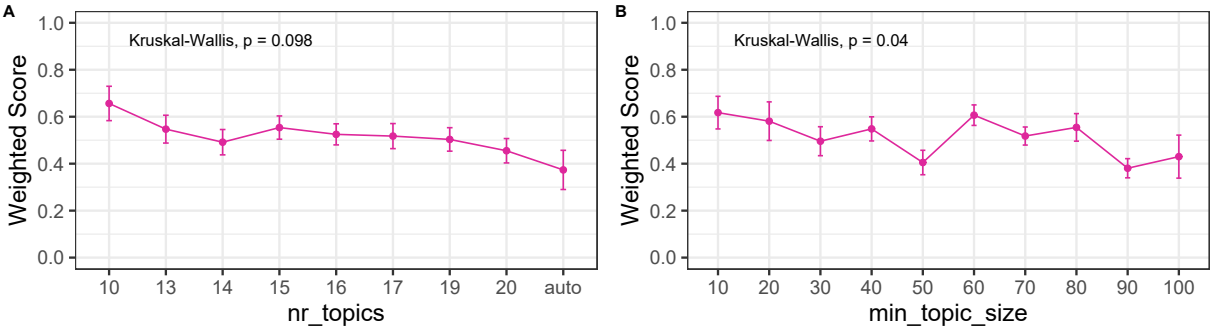
Para concluir a avaliação interna, a Figura 6.11 fornece uma análise mais aprofundada com foco explícito em (**A**) *topic coherence*, (**B**) *topic diversity* do tópico e (**C**) *weighted score* para cada modelo ao considerar dez tópicos. O teste de Wilcoxon pareado é aplicado para a comparação de médias dos modelos, referindo-se ao modelo LiBERT-SE.

Ao examinar essas métricas individualmente, é possível avaliar os pontos fortes e fracos de cada modelo na captura dos temas subjacentes nos documentos de licitações públicas. Alinhado às descobertas anteriores, o desempenho do LiBERT-SE supera, de forma estatisticamente significativa, todos os modelos avaliados considerando o *weighted score*. No entanto, para a métrica de *topic coherence*, apenas em relação ao S-Bertimbau não foi possível afirmar que a média do LiBERT-SE foi superior. Além disso, nosso modelo exibe desempenho competitivo em relação à *topic diversity* do tópico, sugerindo que ele pode gerar tópicos variados e distintos.

Considerando apenas a *weighted score*, que combina *topic coherence* e *topic diversity*, os modelos que apresentam bom desempenho após o LiBERT-SE são S-BERTimbau e LaBSE. O alto desempenho do S-BERTimbau pode ser atribuído ao seu design específico para o idioma português. O LaBSE, por outro lado, é um modelo independente de idioma que aproveita dados de treinamento multilíngues em larga escala. Seu desempenho competitivo pode ser explicado por sua capacidade de lidar com vários idiomas, incluindo o português.

Por fim, o desempenho superior do LiBERT-SE pode ser atribuído a seu vocabulário mais rico e adaptado a jargões típicos do domínio de licitações públicas. Não obstante, o

Figura 6.12: Comparação de desempenho com base na pontuação Ponderado em diferentes variações dos hiperparâmetros *nr_ topics_ list* e *min_ topic_ sizes*.



Fonte: Extraído do artigo [37] escrito pelo autor.

pré-treinamento em dados de licitação pública em larga escala permitiu ao modelo capturar a compreensão contextual do domínio, melhorando assim a qualidade dos grupos de tópicos gerados.

6.3.2 Avaliação externa

Para realizar a avaliação externa, foram ajustados os dois hiperparâmetros específicos do BERTopic usando o modelo de linguagem de melhor desempenho na etapa de avaliação, ou seja, LiBERT-SE. Esses hiperparâmetros são: *nr_ topics_ list* e *min_ topic_ sizes*. A Figura 6.12 mostra o desempenho com base na *weighted score* em diferentes variações de ambos os hiperparâmetros. Para determinar a significância estatística das diferenças obtidas ao variarmos, isoladamente, *nr_ topics_ list* e *min_ topic_ sizes*, aqui também foi empregado o teste de Kruskal-Wallis. Isso nos permitiu identificar as valorações que levaram ao melhor desempenho: 10 e 60, respectivamente.

A Figura 6.13 mostra as nuvens de palavras dos dez tópicos identificados pelo BERTopic, usando os hiperparâmetros ideais. Cada nuvem de palavras representa visualmente as palavras-chave mais significativas associadas a um tópico específico. O primeiro tópico (-1) representa os documentos *outliers* que não se alinham com nenhum tema ou tópico em particular.

Foi observada a presença de muito conteúdo redundante entre os tópicos, isto é,

Tabela 6.4: Tabela com os nomes dos tópicos identificados e suas principais palavras.

ID	Nome do Tópico	Palavras principais	Número de documentos
-1	Outliers	licitacao, edital, ativo, efetivos, propostas, precos, publico, pregao	1629
0	Edital	centro, licitacao, rua, pregao, presencial, edital, situada, local	4979
1	Atas	ata, precos, administracao, total, lei, fornecedor, federal, prazo	2991
2	Orçamento	bdi, sinapi, concreto, thome, total, letras, sao, financeiro	252
3	Contrato	contrato, termo, clausula, aditivo, inscrito, doravante, sob	225
4	Homologação	prefeito, homologacao, moreti, aviso, filho, resultado, vencedoras	124
5	Acreditação	credenciamento, edital, publico, presente, secretaria, praca	107
6	Concurso	impugnacao, empresa, edital, impugnante, processo, nao, razoes	101
7	Aquisição	familiar, alimenticios, agricultura, generos, resolucao, fnde	75
8	Outras Atas	mg, br, ml, prati, cloridrato, ata, oral, donaduzzi	70

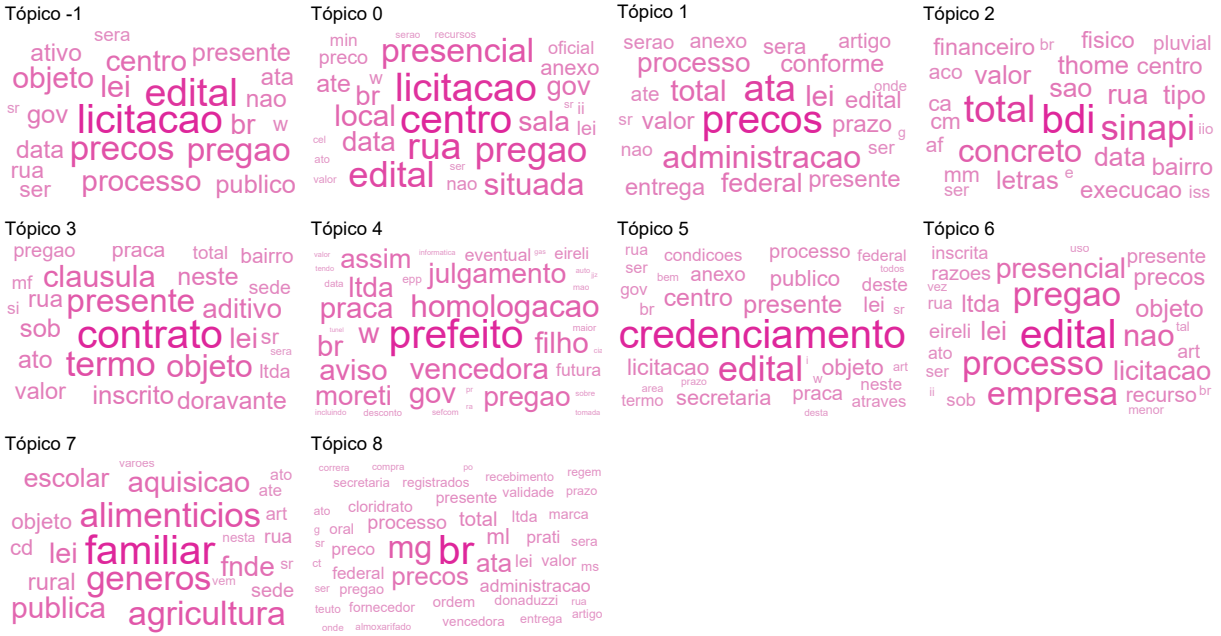
palavras que em geral estão presentes em maior frequência nos documentos de licitação. Notoriamente, o grupo -1 formado pelos *outliers* é o que mais contém palavras de uso geral no contexto de licitação, como preços, pregão público, etc. Os tópicos restantes capturam temas e conteúdos distintos dentro dos documentos. Esses tópicos variam entre domínios ou subcampos específicos das matérias tratadas em uma licitação, como projetos de construção, cadeia de suprimentos, regulamentações legais ou aspectos financeiros.

Para facilitar a compreensão dos tópicos, a Tabela 6.4 apresenta um mapeamento manual, realizado pelo autor a partir da análise das principais palavras de cada tópico, atribuindo um nome semântico final aos tópicos. Esse mapeamento exibe o identificador (ID) de cada tópico, um nome que o caracteriza, além das principais palavras e suas frequências dentro do tópico.

Com base nos nomes mapeados na Tabela 6.4, foi gerado um mapa de calor para analisar a ocorrência de cada documento entre tópicos e rótulos de meta-classe (Figura 6.14). O mapa de calor fornece uma representação visual das associações entre os tópicos identificados e os rótulos verdadeiros das meta-classes atribuídas aos documentos de licitação. A intensidade das cores no mapa de calor reflete a frequência ou ocorrência do par tópico meta-classe. Cores mais escuras indicam uma frequência mais alta, enquanto cores mais claras representam uma frequência mais baixa ou a ausência do par tópico meta-classe.

Os tópicos apresentaram um forte alinhamento com as meta-classes majoritárias Outros e Edital, o que está de acordo com as distribuições de meta-classes apresentada na Seção 6.1.2. Especificamente, os tópicos relacionados à descrição dos itens licitados e ao

Figura 6.13: Nuvem de palavras para cada tópico.

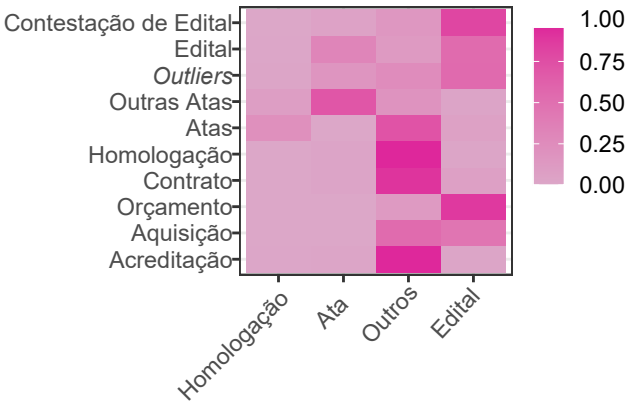


Fonte: Extraído do artigo [37] escrito pelo autor.

orçamento, que geralmente estão presentes em editais, mostraram maior associação com sua respectiva meta-classe. Já para a meta-classe Outros, observa-se que pelo menos cinco tópicos estão majoritariamente relacionados a essa categoria, destacando-se a presença de tópicos identificados como relacionados a contratos, que são um tipo de documento pertencente à meta-classe Outros. As demais classes são minoritárias.

Entretanto, no geral, as associações esperadas para os tópicos identificados como Atas, Outras Atas, Homologação e Edital não se alinham com suas respectivas meta-classes. Este desalinhamento pode ser atribuído a vários fatores. Primeiro, o BERTopic é baseado nos padrões subjacentes e coocorrências de palavras dentro dos documentos, que podem nem sempre se alinhar precisamente com os rótulos semânticos das meta-classes. Em segundo lugar, as especificidades do domínio de licitações públicas podem contribuir para a variação e dispersão de tópicos em diferentes meta-classes. Consequentemente, os tópicos podem se sobrepor ou abranger várias meta-classes, levando a uma correspondência um-para-um menos direta. Além disso, a qualidade e a representatividade dos dados de treinamento podem influenciar os resultados da modelagem de tópicos. Se os dados de treinamento não estiverem cobrindo totalmente a diversidade das categorias presentes nos documentos de licitação, isso pode afetar a precisão das associações tópico e meta-classe.

Figura 6.14: Mapa de calor das associações Tópicos e Meta-classes.



Fonte: Extraído do artigo [37] escrito pelo autor.

Em resumo, o modelo LiBERT-SE apresentou desempenho superior em comparação aos demais modelos. No entanto, melhorias na descrição dos tópicos precisam ser implementadas para que os tópicos identificados pelo agrupamento dos documentos possam ser adequadamente interpretados. Essas melhorias podem ser alcançadas por meio da exploração de técnicas de pré-processamento mais robustas e da avaliação de alternativas ao método C-TF-IDF proposto pelo BERTopic.

6.4 Artigos Publicados

O esforço de pesquisa realizado para a produção desta dissertação produziu também artigos científicos que apresentam resultados parciais do trabalho contemplado por completo neste documento. Nessa seção são abordados esses artigos que contaram com a participação do autor desta dissertação e da equipe envolvida no projeto PCA (ver Seção 1.1). Os artigos são apresentados em ordem cronológica de publicação, destacando como um deles contribuiu para a construção desta dissertação.

- *Lipset: Um conjunto de dados com documentos rotulados de licitações públicas* [72]: Artigo publicado no Simpósio Brasileiro de Banco de Dados (SBBD) no ano de 2022. Como um dos primeiros trabalhos de nosso grupo, ele apresenta a versão inicial do

conjunto explorado nessa dissertação, o LIPSet. Nele é apresentado o desenvolvimento e os resultados da metodologia utilizada nessa dissertação para extração e rotulação dos documentos licitatórios.

- *Impacto do pré-processamento e representação textual na classificação de documentos de licitações* [10]: O artigo prossegue com o desenvolvimento do trabalho da equipe e desta dissertação, apresentando a metodologia para a escolha dos procedimentos de pré-processamento das licitações e modelos de *word embeddings*. Ele foi publicado na edição do SBBD de 2023.

- *Evaluating contextualized embeddings for topic modeling in public bidding domain* [37]: Outro artigo, também publicado em 2023 na *Brazilian Conference on Intelligent Systems*, apresenta a metodologia utilizada nesta dissertação para análise dos tópicos latentes nos documentos de licitação. Além disso, o artigo discute os avanços em relação aos modelos pré-treinados avaliados pela equipe, onde foi desenvolvido um modelo específico para o domínio de licitações públicas, baseado na arquitetura estado da arte *Transformers*.

- *PLUS: A Semi-automated Pipeline for Fraud Detection in Public Bids* [9]: Artigo publicado em 2024 na revista internacional *Digital Government: Research and Practice*. O artigo apresenta todo o trabalho realizado pela equipe até então, descrevendo seus objetivos e os resultados obtidos no contexto da detecção de fraudes em licitações públicas. Além de fornecer um histórico detalhado da equipe, o artigo destaca de maneira crítica os problemas enfrentados e as soluções desenvolvidas.

- *LiPSet: A Comprehensive Dataset of Labeled Portuguese Public Bidding Documents* [71]: Artigo publicado em 2024 na revista Journal of Information and Data Management como extensão dos artigos [72] e [10]. Nesse trabalho e equipe apresenta os resultados da segunda coleta de documentos realizado pelo *web crawlers* e revisa o trabalho e a metologia até então aplicada pela equipe na tarefa de classificação de documentos de licitação.

Capítulo 7

Conclusão e Trabalhos Futuros

Nesta dissertação, foram explorados os desafios presentes no domínio de licitações públicas para o desenvolvimento de aplicações robustas de *Natural Language Processing* (NLP) baseadas em *Deep Learning* (DL). O trabalho foi conduzido no contexto do projeto PCA, com o objetivo de desenvolver ferramentas práticas para apoiar especialistas do Ministério Público do Estado de Minas Gerais (MPMG) na análise de licitações realizadas nos municípios.

Para a elaboração das aplicações, nossa metodologia enfrentou o desafio da coleta dos documentos, culminando na apresentação do LipSET, um conjunto de dados estruturado coletado por meio de *web crawlers*, contendo 9.761 documentos de licitação extraídos dos respectivos portais de transparência de 18 municípios. Dentre os documentos coletados, 6.337 foram rotulados manualmente em 4 meta-classes e 13 tipos (Seção 4.1.2 e Seção 4.1.3).

Além dos dados, propusemos um novo modelo pré-treinado especificamente desenvolvido para tratar documentos de licitações públicas, com potencial para ser facilmente adaptado a novas aplicações. Nomeada LiBERT-SE, a proposta é baseada no modelo estado da arte *Bidirectional Encoder Representations for Transformers* (BERT)[25]. O vocabulário original do modelo foi expandido para incluir terminologias típicas do domínio de licitações, e seu treinamento foi realizado em um conjunto auxiliar composto por 300.000 artigos relacionados a licitações públicas, extraídos do diário oficial de Minas Gerais. Adicionalmente, exploramos o modelo para gerar representações únicas dos documentos por meio da técnica SIMCSE [31].

A caracterização do LiPSET evidenciou os principais desafios de trabalhar com fontes tão heterogêneas, como a diversidade nos formatos e conteúdos dos portais de transparência. Em particular, a análise da distribuição dos documentos revelou um grande desbalanceamento entre as meta-classes e os tipos de documentos, além da ausência de um padrão comum entre os municípios, sendo que cada um apresenta um desbalancea-

mento característico e nem sempre contempla todas as classes propostas. Para visualizar a similaridade entre os documentos, foi utilizado um gráfico de dispersão gerado a partir das representações do LiBERT-SE, o que possibilitou análises importantes, como a forte sobreposição e a esparsidade da meta-classe 'Outros" em relação às demais. Além disso, destacou-se a formação de grupos influenciados principalmente pelos municípios de origem.

Em relação às tarefas propostas neste trabalho, a primeira tarefa desenvolvida foi a aplicação supervisionada de classificação automática de documentos utilizando os rótulos do LiPSET. Foram apresentados dois modelos: um modelo heurístico mais simples, voltado para as quatro meta-classes, que se baseou nas palavras-chave identificadas durante o processo de rotulação; e um segundo modelo, destinado aos 13 tipos de documentos, que combinou técnicas de pré-processamento e *word embeddings* para a construção de um modelo baseado em *Long Short-Term Memory* (LSTM) [36]. Apesar do alto tempo de treinamento devido ao uso de Redes Neurais Recorrentes (RNNs), essa solução apresenta baixo custo de *hardware* em função dos modelos e técnicas utilizados. Para o método heurístico, foi utilizado todo o conjunto de dados para avaliar sua eficácia. Já no modelo LSTM, aplicou-se validação cruzada estratificada por tipo e cidade, simulando o cenário prático de classificar documentos de novos municípios com base em subconjuntos de documentos previamente disponíveis.

Apesar dos desafios destacados na caracterização, os modelos de classificação propostos apresentaram resultados satisfatórios. O modelo LSTM, para a classificação de tipos de documentos, obteve, na melhor configuração utilizando o pré-processamento *base+*, uma média superior a 96% nas métricas F1 para ambos os cenários avaliados. Contudo, os testes estatísticos indicaram que não foi possível observar um impacto significativo entre as diferentes abordagens de pré-processamento e métodos de representação considerados. Já o método heurístico para a classificação das meta-classes alcançou uma métrica de F1-Macro de 91%, demonstrando que, apesar de sua simplicidade, padrões básicos podem ser explorados na classificação de documentos. Além disso, a análise das matrizes de confusão dos classificadores reforçou as limitações das classes propostas para os tipos das meta-classes Outros e Edital, nas quais há uma forte sobreposição de documentos.

Para a segunda tarefa, foi empregada a modelagem de tópicos (MT), uma abordagem não supervisionada. Nesse contexto, avaliamos a performance do LiBERT-SE na identificação de padrões latentes em documentos de licitação. Para isso, utilizamos a metodologia BERTopic [34], proposta especificamente para a tarefa de modelagem de tópicos com base em modelos BERT. Foram realizadas uma avaliação interna, que comparou

a eficácia do LiBERT-SE com outros modelos disponíveis na literatura, e uma avaliação externa, que examinou a qualidade dos tópicos identificados pela metodologia BERTopic.

Na tarefa de MT, o modelo LiBERT-SE superou consistentemente os *baselines*, apresentando diferenças estatisticamente significativas nas principais métricas de avaliação interna. Os hiperparâmetros do BERTopic foram ajustados especificamente utilizando o LiBERT-SE, o que permitiu identificar o número ideal de 10 tópicos para os documentos. Na avaliação qualitativa do melhor cenário com 10 tópicos, observou-se uma grande quantidade de ruído, com termos gerais de licitação sendo rotulados como *outliers*. Além disso, não foi possível observar um alinhamento claro entre as meta-classes e os tópicos identificados. No entanto, foi possível identificar as principais temáticas relacionadas aos tópicos, que abrangiam assuntos gerais de licitação.

Em suma, as aplicações desenvolvidas neste trabalho permitiram explorar o domínio de licitações públicas e entender melhor suas peculiaridades e possíveis aplicações. Atualmente, o classificador de tipos de documentos está sendo utilizado por sistemas desenvolvidos no contexto do projeto PCA. Por outro lado, para a tarefa de modelagem de tópicos (MT), ainda são necessários refinamentos para seu uso prático por parte dos especialistas. Melhorias na descrição dos tópicos precisam ser consideradas, o que pode envolver a aplicação de técnicas mais avançadas de pré-processamento ou o uso de métodos alternativos ao c-TF-IDF proposto pelo BERTopic. O objetivo é proporcionar descrições mais precisas para os tópicos identificados a partir dos agrupamentos gerados pelas representações do LiBERT-SE.

7.1 Limitações e trabalhos futuros

Nesta seção, discutimos os limites da pesquisa realizada e damos direcionamento a novas oportunidades que podem ser exploradas para expandir os objetivos da dissertação.

Em primeiro lugar, foram observadas importantes limitações nos documentos coletados. A baixa quantidade de amostras de algumas meta-classes e tipos de documento dificulta uma análise precisa e também compromete o treinamento de novos modelos. Além disso, os *web crawlers* desenvolvidos realizaram a coleta apenas uma única vez por portal, o que poderia ser aprimorado para permitir uma coleta periódica programada, possibilitando

que o conjunto de dados de licitações seja continuamente atualizado. Como trabalho futuro relacionado aos dados, é necessário estender a coleta para mais cidades e comparar os documentos coletados com as licitações estaduais de Minas Gerais, que podem ser obtidas de uma fonte única.

Quanto ao classificador, apesar do desempenho elevado, o tempo de treinamento da LSTM ainda é um grande limitador para a evolução e manutenção do modelo com novos dados. Além disso, devido ao número limitado de amostras, não foi possível determinar com precisão quais técnicas são mais adequadas ao domínio. Estudos futuros podem investigar o uso de dados de mais municípios ou licitações estaduais. Além disso, pode-se avaliar o uso do LiBERT-SE na tarefa de classificação, uma vez que ele já está adaptado para o domínio de licitações. Embora tenha um custo de *hardware* mais elevado, o LiBERT-SE pode ser ajustado com facilidade usando documentos provenientes de novas coletas.

Em relação à modelagem de tópicos, notou-se uma quantidade significativa de ruídos nos tópicos identificados. Como trabalho futuro, seria relevante estudar mais a fundo os impactos do pré-processamento e da redução de dimensionalidade para gerar tópicos mais coesos na metodologia BERTopic. Além disso, metodologias alternativas ao BERTopic podem ser exploradas para melhorar a qualidade dos tópicos gerados. Técnicas que combinam o clássico LDA com modelos de *word embeddings* poderiam ser investigadas, conforme discutido em [12]

Por fim, direcionamos como aplicações futuras, baseadas nos métodos implementados, o uso de técnicas de *Out-of-Distribution* [88, 55], que utilizam as pontuações de modelos de DL para identificar dados que diferem significativamente da distribuição original. Esses métodos poderiam ser empregados em conjunto com dados não rotulados para avaliar documentos fora da distribuição e ajudar na detecção de possíveis padrões irregulares entre os documentos [32].

Referências

[1] Hidelberg O Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia FF da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, et al. Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In *International Conference on Computational Processing of the Portuguese Language*, pages 3–14. Springer, 2022.

[2] Miguel Arana-Catania, Felix-Anselm van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. Citizen participation and machine learning for a better democracy. *Digit. Gov. Res. Pract.*, 2(3):27:1–27:22, 2021.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Yanwei Bao, Changqin Quan, Lijuan Wang, and Fuji Ren. The role of pre-processing in twitter sentiment analysis. In *Intelligent Computing Methodologies: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10*, pages 615–624. Springer, 2014.

[5] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[6] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[8] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.

[9] Michele A Brandão, Arthur PG Reis, Bárbara MA Mendes, Clara A Bacha De Almeida, Gabriel P Oliveira, Henrique Hott, Larissa D Gomide, Lucas L Costa, Mariana O Silva, Anisio Lacerda, et al. Plus: A semi-automated pipeline for fraud detection in public bids. *Digital Government: Research and Practice*, 5(1):1–16, 2024.

[10] Michele A Brandão, Mariana O Silva, Gabriel P Oliveira, Henrique R Hott, Anísio M Lacerda, and Gisele L Pappa. Impacto do pré-processamento e representação textual na classificação de documentos de licitações. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 102–114. SBC, 2023.

[11] Pedro PV Brum, Mariana O Silva, Gabriel P Oliveira, Lucas GL Costa, Anisio Lacerda, and Gisele Pappa. Unsupervised grouping of public procurement similar items: Which text representation should i use? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17176–17185, 2024.

[12] Stefan Bunk and Ralf Krestel. Welda: Enhancing topic models by incorporating local word context. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pages 293–302, 2018.

[13] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *PAKDD*, volume 7819, pages 160–172. Springer, 2013.

[14] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics.

[15] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[16] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[18] Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35, 2022.

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

[20] Kattiana Constantino et al. Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. In *SBBD*, pages 304–316. SBC, 2022.

[21] Daniel da Silva Junior, Paulo Roberto dos Santos Corval, Daniel de Oliveira, and Aline Paes. Datasets for portuguese legal semantic textual similarity. *Journal of Information and Data Management*, 15(1):206–215, 2024.

[22] Pedro Henrique Luz De Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. Victor: a dataset for brazilian legal documents classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458, 2020.

[23] Wim De Mulder, Steven Bethard, and Marie-Francine Moens. A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language*, 30(1):61–98, 2015.

[24] Emilio Feliciano de Oliveira and Milene Selbach Silveira. Open government data in brazil a systematic review of its uses and issues. In Marijn Janssen, Soon Ae Chun, and Vishanth Weerakkody, editors, *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, DG.O 2018, Delft, The Netherlands, May 30 - June 01, 2018*, pages 60:1–60:9. ACM, 2018.

[25] Jacob Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics, 2019.

[26] Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguistics*, 8:439–453, 2020.

[27] Kellyton dos Santos Brito, Marcos Antônio da Silva Costa, Vinicius Cardoso Garcia, and Silvio Romero de Lemos Meira. Experiences integrating heterogeneous government open data sources to deliver services and promote transparency in brazil. In *IEEE 38th Annual Computer Software and Applications Conference, COMPSAC 2014, Vasteras, Sweden, July 21-25, 2014*, pages 606–607. IEEE Computer Society, 2014.

[28] Nicola Ehlermann-Cache et al. *Bribery in public procurement: methods, actors and counter-measures*. OECD, 2007.

[29] Ronen Feldman and James Sanger. *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[30] Fangxiaoyu Feng et al. Language-agnostic BERT sentence embedding. In *ACL*, pages 878–891. Association for Computational Linguistics, 2022.

[31] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910. Association for Computational Linguistics, 2021.

[32] Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[34] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[35] Bambang Leo Handoko and Ameliya Rosita. The effect of skepticism, big data analytics to financial fraud detection moderated by forensic accounting. In *2022 6th International Conference on E-Commerce, E-Business and E-Government*, pages 59–66, 2022.

[36] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[37] Henrique R. Hott, Mariana O. Silva, Gabriel P. Oliveira, Michele A. Brandão, Anisio Lacerda, and Gisele Pappa. Evaluating contextualized embeddings for topic modeling in public bidding domain. In Murilo C. Naldi and Reinaldo A. C. Bianchi, editors, *Intelligent Systems*, pages 410–426, Cham, 2023. Springer Nature Switzerland.

[38] Hsun-Ping Hsieh, JiaWei Jiang, Tzu-Hsin Yang, Renfen Hu, and Cheng-Lin Wu. Predicting the success of mediation requests using case properties and textual information for reducing the burden on the court. *Digit. Gov. Res. Pract.*, 2(4):30:1–30:18, 2021.

[39] Ruben Interian, Igor Carpanese, Bruno Mello, and Celso C Ribeiro. Red flag algorithms for brazilian electronic invoices: outlier detection and price risk classification. *International Transactions in Operational Research*, 2023.

[40] Karuna Pande Joshi and Srishty Saha. A semantically rich framework for knowledge representation of code of federal regulations. *Digit. Gov. Res. Pract.*, 1(3):21:1–21:17, 2020.

[41] Ilka Kawashita, Ana Alice Baptista, and Delfina Soares. Open government data use in the brazilian states and federal district public administrations. *Data*, 7, 2022.

[42] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics, 2018.

[43] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[44] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, 2014.

[45] João Pedro Lima and José Alfredo Costa. Comparing clustering techniques on brazilian legal document datasets. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 98–110. Springer, 2022.

[46] Marcos Lima, Roberta Silva, Felipe Lopes de Souza Mendes, Leonardo R de Carvalho, Aleteia Araujo, and Flavio de Barros Vidal. Inferring about fraudulent collusion risk on brazilian public works contracts in official texts using a bi-lstm approach. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1580–1588, 2020.

[47] Alex Aguiar Lins, Cecilia Silvestre Carvalho, Francisco das Chagas Jucá Bomfim, Daniel de Carvalho Bentes, and Vládia Pinheiro. Clsjur. br-a model for abstractive summarization of legal documents in portuguese language based on contrastive learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 321–331, 2024.

[48] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[49] Weverton RR Mata, DS Boechat, and Michele A Brandao. Jusbd: Um banco de dados para obtenção de informações do poder judiciário. In *Anais do II Dataset Showcase Workshop, DSW*, pages 398–407, 2019.

[50] Ricardo Matheus, Manuella Maia Ribeiro, and José Carlos Vaz. New perspectives for electronic government in brazil: the adoption of open government data in national and subnational governments of brazil. In David Ferriero, Theresa A. Pardo, and Haiyan Qian, editors, *6th International Conference on Theory and Practice of Electronic Governance, ICEGOV '12, Albany, NY, USA, October 22-25, 2012*, pages 22–29. ACM, 2012.

[51] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[52] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[53] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *2011 IEEE in-*

ternational conference on acoustics, speech and signal processing (ICASSP), pages 5528–5531. IEEE, 2011.

[54] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[55] John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7721–7735. PMLR, 2021.

[56] Sarah Moore. Digital government, public participation and service transformation: the impact of virtual courts. *Policy & Politics*, 47:495 – 509, 2019.

[57] Roberto Nai, Emilio Sulis, and Rosa Meo. Public procurement fraud detection and artificial intelligence techniques: a literature review. In *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*, Bozen-Bolzano, Italy, 2022.

[58] Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(5), June 2021.

[59] Mariana Y. Noguti, Eduardo Vellasques, and Luiz S. Oliveira. Legal document classification: An application to law area prediction of petitions to public prosecution service. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE, 2020.

[60] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

[61] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent

Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[62] Samuli Pekkola, Maija Ylinen, and Nicholas B. Mavengere. Consortium of municipalities co-tailoring a governmental e-service platform: What could go wrong? *Digit. Gov. Res. Pract.*, 3(1):6:1–6:16, 2022.

[63] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[64] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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, 2018.

[65] Mikael Poetsch, Ulisses Brisolara Correa, and Larissa Astrogildo de Freitas. A word embedding analysis towards ontology enrichment. *Res. Comput. Sci.*, 148(11):153–164, 2019.

[66] Diany Pressato, Pedro Lucas Castro de Andrade, Flávio Rocha Junior, Felipe Alves Siqueira, Ellen Polliana Ramos Souza, Nádia Félix Felipe da Silva, Márcio de Souza Dias, André Carlos Ponce de Leon Ferreira, et al. Natural language processing application in legislative activity: a case study of similar amendments in the brazilian senate. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 614–619, 2024.

[67] Foster J Provost, Tom Fawcett, Ron Kohavi, et al. The case against accuracy estimation for comparing induction algorithms. In *ICML*, volume 98, pages 445–453, 1998.

[68] Gabriel Puron-Cid, Dolores E. Luna, Sergio Picazo-Vela, J. Ramón Gil-Garcia, Rodrigo Sandoval-Almazan, and Luis F. Luna-Reyes. Improving the assessment of digital services in government websites: Evidence from the mexican state government portals ranking. *Government Information Quarterly*, 39(1):101589, 2022.

[69] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*, pages 3980–3990. Association for Computational Linguistics, 2019.

[70] Hans Jochen Scholl. Digital government: Looking back and ahead on a fascinating domain of research and practice. *Digit. Gov. Res. Pract.*, 1(1):7:1–7:12, 2020.

[71] Mariana O Silva, Gabriel P Oliveira, Henrique Hott, Larissa D Gomide, Bárbara MA Mendes, Clara A Bacha, Lucas L Costa, Michele A Brandão, Anisio Lacerda, and Gisele L Pappa. Lipset: A comprehensive dataset of labeled portuguese public bidding documents. *Journal of Information and Data Management*, 15(1):196–205, 2024.

[72] Mariana O Silva, Amanda F Paula, Gabriel P Oliveira, Iago AD Vaz, Henrique Hott, Larissa D Gomide, Arthur PG Reis, Bárbara MA Mendes, Clara A Bacha, Lucas L Costa, et al. Lipset: Um conjunto de dados com documentos rotulados de licitações públicas. In *Dataset Showcase Workshop (DSW)*, pages 13–24. SBC, 2022.

[73] Nádia FF da Silva, Marília Costa R Silva, Fabíola SF Pereira, João Pedro M Tarrega, João Vitor P Beinotti, Márcio Fonseca, Francisco Edmundo de Andrade, and André CP de LF de Carvalho. Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 104–120. Springer, 2021.

[74] Raquel Silveira, Carlos G Fernandes, João A Monteiro Neto, Vasco Furtado, and José Ernesto Pimentel Filho. Topic modelling of legal documents via legal-bert. *Proceedings http://ceur-ws org ISSN*, 1613:0073, 2021.

[75] Raquel Silveira, Caio Ponte, Vitor Almeida, Vládia Pinheiro, and Vasco Furtado. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Brazilian Conference on Intelligent Systems*, pages 268–282. Springer, 2023.

[76] Ksh Singh, H Mamata Devi, Anjana Kakoti Mahanta, et al. Document representation techniques and their effect on the document clustering and classification: A review. *International Journal of Advanced Research in Computer Science*, 8(5), 2017.

[77] Felipe A Siqueira, Douglas Vitório, Ellen Souza, José AP Santos, Hidelberg O Albuquerque, Márcio S Dias, Nádia FF Silva, André CPLF de Carvalho, Adriano LI

Oliveira, and Carmelo Bastos-Filho. Ulysses tesemõ: a new large corpus for brazilian legal and governmental domain. *Language Resources and Evaluation*, pages 1–20, 2024.

[78] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[79] Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. BERTimbau: Pretrained BERT models for brazilian portuguese. In *BRACIS*, volume 12319, pages 403–417. Springer, 2020.

[80] Donald F Specht et al. A general regression neural network. *IEEE transactions on neural networks*, 2(6):568–576, 1991.

[81] Kuang-Ting Tai. Open government research over a decade: A systematic review. *Government Information Quarterly*, 38:101566, 2021.

[82] Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana. *Practical natural language processing: a comprehensive guide to building real-world NLP systems*. O'Reilly Media, 2020.

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[84] Daniela Vianna, Edleno Silva de Moura, and Altigran Soares da Silva. A topic discovery approach for unsupervised organization of legal document collections. *Artificial Intelligence and Law*, pages 1–30, 2023.

[85] Daniela Vianna and Edleno Silva de Moura. Organizing portuguese legal documents through topic discovery. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3388–3392, 2022.

[86] Bernd W. Wirtz, Jan C. Weyerer, and Michael Rösch. Citizen and open government: An empirical analysis of antecedents of open government data. *International Journal of Public Administration*, 41(4):308–320, 2018.

[87] Henrique S Xavier. Overseeing government with ai: Lessons learned from a brazilian experience. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE, 2023.

[88] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: benchmarking generalized out-of-distribution detection. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[89] Yinfei Yang et al. Multilingual universal sentence encoder for semantic retrieval. In *ACL*, pages 87–94. Association for Computational Linguistics, 2020.

[90] Jiarui Zhang, Yingxiang Li, Juan Tian, and Tongyan Li. Lstm-cnn hybrid model for text classification. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1675–1680. IEEE, 2018.