Impacto do Pré-processamento e Representação Textual na Classificação de Documentos de Licitações

Michele A. Brandão[1,2], Mariana O. Silva[1], Gabriel P. Oliveira[1], Henrique R. Hott[1], Anísio M. Lacerda[1], Gisele L. Pappa[1]

[1]Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte - MG

[2]Instituto Federal de Minas Gerais (IFMG) - Ribeirão das Neves - MG

{michele.brandao, mariana.santos, gabrielpoliveira,

henriquehott, anisio, glpappa}@dcc.ufmg.br

Abstract. *Classifying public bidding documents is relevant for public and private bodies seeking accurate information about such processes. In this work, we investigate the impact of different preprocessing approaches and textual representation models of word embeddings on the effectiveness of the classification of bidding documents. The results show that the preprocessing does not significantly impact the classification result and that the textual representation is essential for the document classes to be more representative.*

Resumo. *A classificação de documentos de licitações públicas é uma tarefa relevante para órgãos públicos e privados que buscam informações precisas sobre tais processos. Neste trabalho, investigamos o impacto de diferentes abordagens de pré-processamento e modelos de representação textual por word embeddings na eficácia da classificação de documentos de licitação. Os resultados evidenciam que o pré-processamento não impacta significativamente no resultado da classificação e que a representação textual é um aspecto importante para que as classes de documentos sejam mais representativas.*

1. Introdução

O uso de dados abertos governamentais é uma política que tem sido cada vez mais difundida ao redor do mundo como uma forma de promover a transparência e a responsabilidade das instituições públicas perante a sociedade em geral. No Brasil, a Lei de Acesso à Informação (Lei n° 12.527, de 18 de novembro de 2011)[1] é considerada um importante marco na democracia, pois permite uma maior participação da sociedade nas ações governamentais. No entanto, trabalhar com esse grande volume de dados governamentais traz uma série de desafios, incluindo a diversidade e a complexidade das fontes de dados. Além disso, a produção constante de novas informações evidencia a necessidade do uso de abordagens automatizadas para lidar com tais dados.

Nesse sentido, a classificação de documentos de licitações públicas é uma tarefa importante para organizações governamentais e empresas privadas que buscam informações precisas e relevantes sobre processos de licitação. No entanto, a análise

[1]Lei de Acesso à Informação: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

manual de grandes volumes de dados pode ser demorada e sujeita a erros humanos [Oliveira et al. 2022]. Assim, a utilização de algoritmos de classificação se torna cada vez mais importante, uma vez que pode tornar o processo mais eficiente e preciso, permitindo a análise de grandes volumes de dados e reduzindo a possibilidade de erros humanos.

No entanto, para que os algoritmos de classificação sejam eficazes, é necessário realizar etapas de pré-processamento e representação de texto adequadas. O pré-processamento envolve a padronização do texto e a redução do vocabulário, visando tornar os dados mais representativos e menos esparsos. Já a representação de texto consiste em converter o texto em uma forma que possa ser utilizada como entrada em algoritmos de classificação, como vetores numéricos. Essas etapas são críticas para garantir que os algoritmos possam capturar nuances complexas e informações contextuais presentes nos documentos, o que pode melhorar significativamente a precisão da classificação.

A natureza não estruturada dos dados textuais pode tornar a tarefa de classificação ainda mais desafiadora, uma vez que palavras comuns, termos técnicos e variações linguísticas podem levar a ambiguidades e a interpretações diferentes do mesmo texto. Como consequência, geralmente, é necessário utilizar abordagens mais sofisticadas, como as redes neurais, que conseguem capturar nuances e relações complexas entre as palavras em um texto, melhorando a eficácia da classificação de documentos. Este trabalho, portanto, tem como objetivo avaliar o impacto de diferentes técnicas de pré-processamento e modelos de representação de texto por *word embeddings* na classificação de documentos de licitação pública, por meio do uso de redes neurais artificiais. É importante destacar que a técnica de *word embeddings* foi escolhida para representação textual por ser bem estabelecida e amplamente utilizada em problemas de processamento de linguagem natural [Albalawi et al. 2021, Poetsch et al. 2019].

Além de discutir os trabalhos relacionados na Seção 2, nossas contribuições incluem a descrição de uma metodologia para classificação dos documentos de licitação com ênfase no pré-processamento e representação de texto (Seção 3), um detalhamento sobre as configurações realizadas na experimentação (Seção 4) e a apresentação dos resultados da classificação de documentos de licitação utilizando dados reais (Seção 5). Finalmente, são discutidas as principais conclusões e limitações deste trabalho, bem como são apresentadas as oportunidades para trabalhos futuros.

2. Trabalhos Relacionados

A classificação de documentos no domínio jurídico é uma tarefa desafiadora devido ao extenso vocabulário e jargões técnicos presentes nos documentos jurídicos, principalmente em português, onde há falta de conjuntos de dados disponíveis [Bambroo and Awasthi 2021, de Araujo et al. 2023]. Dentre as iniciativas presentes na literatura, o projeto VICTOR [Luz de Araujo et al. 2020] apresenta um conjunto de dados rotulados de documentos do Supremo Tribunal Federal. Os documentos foram digitalizados e estruturados através de processos não supervisionados. O conjunto de dados suporta duas tarefas: classificação de documento por tipo e classificação multi-rótulos. Mais próximo ao contexto desse trabalho de licitações públicas, o LiPSet [Silva et al. 2022] é um conjunto de dados que pode ser utilizado para classificação com documentos de licitações públicas de 16 municípios do estado de Minas Gerais. Nesse trabalho, os documentos foram coletados do portal da transparência de cada município e disponibilizados de

maneira estruturada. Além disso, os documentos foram manualmente rotulados de acordo com a sua função no processo licitatório.

Outro trabalho relevante é o de [Lima et al. 2020], que propõe uma nova metodologia para detecção de fraudes em procurações públicas utilizando redes neurais recorrentes. Para tal objetivo, foi construído um conjunto de dados de procurações públicas extraídos de documentos publicados no Diário Oficial da União. Os documentos foram rotulados como suspeitos de fraudes ou não a partir da análise de especialistas. Além da contribuição do novo conjunto de dados o modelo de classificação proposto alcançou resultados competitivos em relação às métricas de precisão, revocação e F1 em comparação a outros modelos do estado da arte indicando a eficácia do uso de modelos de aprendizado profundo para a tarefa.

Além dos trabalhos mencionados anteriormente, o artigo [Coelho et al. 2022] aborda o problema de identificar o valor do dano moral em pareceres jurídicos como um problema de classificação. Foi utilizado um conjunto de técnicas de pré-processamento e *word embeddings* para gerar atributos representativos dos documentos, que foram usados para treinar vários modelos de classificação. Os resultados indicaram que os modelos baseados em *word embeddings* superaram os baselines que utilizavam TF-IDF para geração de atributos. Na literatura, é possível observar trabalhos visando analisar quais são as técnicas de pré-processamento e representação dos dados de forma mais adequada e mensurar seu impacto para uma tarefa alvo. Em [Noguti et al. 2020], é realizada uma comparação entre diferentes abordagens de representação textual para categorizar as descrições dos serviços realizados pelo Ministério Público do Estado do Paraná. Reforçando a importância das técnicas de pré-processamento, em [Belém et al. 2022], é possível obter melhorias expressivas para a detecção de entidades nomeadas e extração de relações com a adição de etapas especializadas de pré e pós-processamento.

Considerando tanto a representação quanto os impactos do pré-processamento, o estudo de [Albalawi et al. 2021] investiga o impacto do pré-processamento em textos em língua árabe relacionados à saúde, utilizando técnicas de pré-processamento e *word embeddings*. Foi constatado que apenas quatro das 26 técnicas de pré-processamento tiveram impacto significativo na performance dos modelos classificadores avaliados. Além disso, o uso de técnicas de normalização textual específicas para o idioma do problema mostrou-se mais eficaz. Modelos baseados em aprendizado profundo obtiveram resultados superiores aos modelos tradicionais, independentemente da configuração de *word embeddings* e pré-processamento.

O artigo [Souza Júnior et al. 2022] avaliou diferentes metodologias de pré-processamento na modelagem de tópicos para o português brasileiro. Foram aplicados três modelos de representação de documentos, incluindo duas novas propostas baseadas no modelo *CluWords* adaptadas para o português. O aumento da complexidade da metodologia de pré-processamento teve um impacto positivo, embora não significativo para o método de representação baseado em TDF-IDF. As novas propostas de representação obtiveram resultados bem mais altos na métrica de coerência. Quando combinadas com o pipeline de pré-processamento, foi possível obter um resultado cerca de nove vezes maior que com o modelo baseline. A nova proposta apresenta os melhores resultados na literatura para o idioma português brasileiro, considerando os conjuntos de dados utilizados.

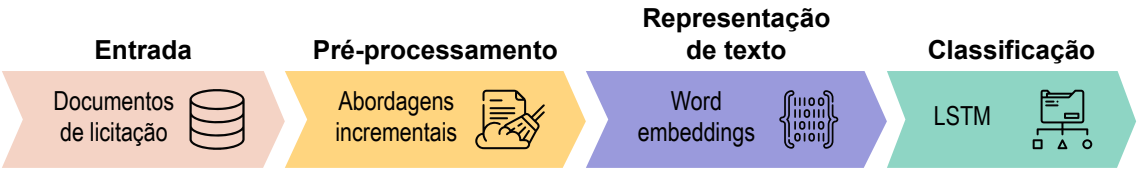


Figura 1. Metodologia para classificação de documentos de licitação.

Os trabalhos relacionados evidenciam a importância de considerar o pré-processamento de textos para modelos de linguagem natural e a necessidade de usar técnicas específicas para representação textual de cada idioma para realização de uma tarefa alvo. Este trabalho propõe uma nova abordagem para avaliar diferentes métodos de pré-processamento em conjunto com modelos de representação de documentos de texto por *word embeddings* para a tarefa de classificação de documentos de licitações públicas no idioma português brasileiro. Essa avaliação pode trazer novas perspectivas, já que não foram encontrados estudos semelhantes na literatura com ênfase em documentos de licitação (não padronizados com textos longos e não estruturados).

3. Metodologia

Esta seção apresenta as principais etapas para classificação dos documentos de licitação. A Figura 1 apresenta as quatro principais etapas da metodologia que inclui a extração de texto dos documentos de licitação (Seção 3.1) para serem utilizados como entrada de dados para a etapa de pré-processamento (Seção 3.2). Após a aplicação de quatro abordagens de pré-processamento, o texto é então representado por meio de *word embeddings* (Seção 3.3) para ser utilizado como entrada para um modelo de classificação. Aqui, foi utilizada a LSTM (Seção 3.4).

3.1. Definição dos Dados e das Classes de Documentos de Licitação

Os documentos de licitação utilizados neste trabalho são os mesmos presente no conjunto de dados chamado de LiPSet [Silva et al. 2022]. Esse conjunto de dados é formado por 6.337 documentos que foram rotulados manualmente quanto ao tipo do documento de licitação identificado (por exemplo, edital, ata, aviso). Ao todo, foram identificados 56 tipos de documentos que referem-se a documentos mais específicos, incluindo errata, notificação, ratificação, publicação em jornal oficial, contrato e emenda. Como a classe real da maioria dos documentos está presente no título do arquivo, o processo de rotulagem manual foi realizado verificando os títulos de cada documento.

Após processados, os tipos de documentos de licitação passaram por uma nova análise manual, onde documentos de tipos similares foram agrupados, resultando assim em 12 classes de documentos de licitação. Os agrupamentos para gerar as classes foram realizados de forma cuidadosa e criteriosa levando em consideração os seguintes critérios: a meta-classe, por permitir uma separação apropriada para os documentos; a quantidade de documentos coletados, por não ser viável treinar um classificador com poucos documentos o representando; e a importância desse documento em um processo licitatório, pelo fato da quantidade de tipos de documentos ser bastante elevado (só na amostra coletada, 56 tipos foram identificados) foi necessário elaborar classes mais representativas para os documentos de um processo licitatório.

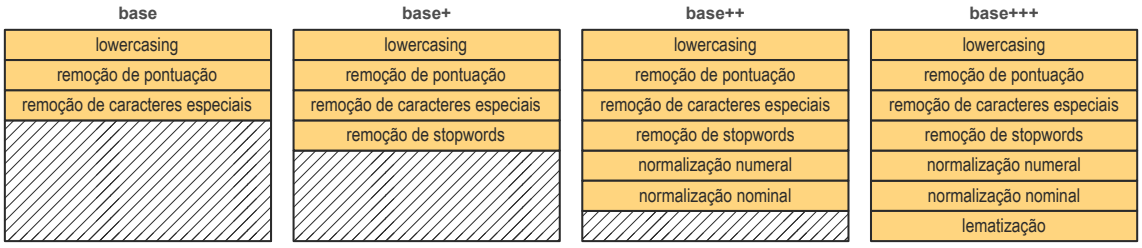


Figura 2. Visão geral das quatro abordagens de pré-processamento.

As classes identificadas neste trabalho, a partir da análise dos documentos de licitação, por meta-classe são: ata dispensa licitação, ata pregão presencial, ata registro preços e outras atas para a meta-classe Ata; edital para a meta-classe Edital; adjudicação/homologação para a meta-classe Homologação/Adjudicação; e errata, aditamento, outros, aviso, contrato e ratificação para a meta-classe Outros. É importante destacar que todas as metodologias e resultados apresentados nas seções seguintes foram desenvolvidos considerando essas 12 classes.

3.2. Descrição do Pré-processamento dos Textos

O pré-processamento é uma etapa fundamental para a representação adequada dos documentos, garantindo que os dados de entrada do classificador sejam relevantes e representativos para a tarefa de classificação em questão. Esse processo geralmente envolve a padronização do texto, que inclui a remoção de acentos e a transformação do texto em caixa baixa, entre outras técnicas. Além disso, é importante reduzir o vocabulário, identificando e tratando/retirando termos irrelevantes, como stopwords, o que torna os dados menos esparsos e facilita o processamento computacional. A combinação adequada dessas técnicas de pré-processamento pode ter um impacto significativo na precisão da classificação automática de documentos de licitações.

Neste estudo, quatro abordagens diferentes são avaliadas para verificar a melhor estratégia de pré-processamento, conforme detalhado na Figura 2. Essas abordagens foram selecionadas com base nas técnicas de pré-processamento empregadas em [Noguti et al. 2020] e estudos preliminares dos dados coletados. A primeira abordagem, referida como "base", consiste em três etapas principais: *lowercasing*, *remoção de pontuação* e *remoção de caracteres especiais*. As quatro abordagens subsequentes são construídas a partir da abordagem "base", com a "base+" incluindo uma etapa adicional de *remoção de stopwords*, a "base++" incluindo duas etapas de *normalização numérica e nominal*, e a "base+++" incluindo também a *lematização*. Cada técnica de pré-processamento é melhor descrita a seguir.

***Lowercasing*.** Converte todos os caracteres do texto em minúsculas. Essa etapa é útil para reduzir a variabilidade do texto e garantir que as mesmas palavras não sejam tratadas como entidades diferentes devido a variações na capitalização. Além disso, também pode ajudar a padronizar os dados do texto e evitar redundância, principalmente em situações em que a mesma palavra pode ser usada de formas diferentes (e.g., "ata" e "Ata"). No geral, é uma técnica de pré-processamento útil para dados textuais, principalmente nos casos em que o texto não é estruturado e pode conter variações na capitalização.

Remoção de pontuação. Envolve a remoção de todos os sinais de pontuação do texto,

incluindo símbolos como vírgulas, pontos, dois-pontos, ponto-e-vírgula e outros. Essa técnica simplifica os dados textuais e reduz o número de palavras únicas, removendo sinais de pontuação que não transmitem nenhum significado ou contexto significativo.

Remoção de caracteres especiais. Envolve a remoção de caracteres não alfanuméricos do texto, incluindo símbolos como hashtags, arrobas, cifrões e outros caracteres especiais que não sejam letras ou números. Essa técnica também simplifica o texto e reduz o número de palavras únicas, removendo caracteres especiais que não transmitem significado ou contexto significativo.

Remoção de stopwords. Envolve a remoção 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 dessa técnica é reduzir o ruído nos dados e melhorar a precisão das tarefas de análise subsequentes, removendo palavras que não transmitem nenhum significado ou contexto significativo. Aqui, usamos uma lista de stopwords do português brasileiro, disponibilizado pela biblioteca NLTK.[2] Além disso, também foram removidos os nomes das cidades pertencentes a cada documento, para evitar que informações de localização sobrecarreguem o modelo de classificação e prejudiquem o desempenho.

Normalização numeral. Envolve a conversão de todos os numerais no texto para um formato padrão. Isso pode incluir a substituição de dígitos por suas palavras correspondentes (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 desta técnica é reduzir a variabilidade dos dados de texto e simplificar as tarefas de análise subsequentes, tratando todos os valores numéricos de forma consistente. Aqui, seguindo [Noguti et al. 2020], substituímos todos os numerais por zero.

Normalização nominal. Envolve a conversão de nomes próprios no texto para um formato padrão. O objetivo é reduzir a variação de escrita de nomes que podem aparecer de forma diferente, como abreviações, erros ortográficos ou variações na grafia. A normalização desses nomes pode aumentar a precisão da classificação e garantir que as informações relevantes sejam corretamente identificadas. Para isso, utilizamos um dicionário com nomes próprios brasileiros comuns e mapeamos todos os nomes pelo termo *proper_name*, seguindo a abordagem de [Noguti et al. 2020].

Lemmatization. Envolve a redução de palavras no texto à sua forma básica ou de dicionário, conhecida como lema. Isso envolve identificar a forma da raiz de uma palavra e mapear todas as formas flexionadas dessa palavra para o mesmo lema (e.g., "caminhar", "caminhou", "caminhando" mapeiam para "caminhar"). O objetivo da lematização é reduzir a variabilidade dos dados e simplificar as tarefas de análise subsequentes, tratando todas as formas flexionadas de uma palavra como uma única entidade. Aqui, usamos a biblioteca spaCy (para a língua portuguesa)[3] para realizar a lematização dos textos.

3.3. Representação Textual com *Word Embeddings*

Após a etapa de pré-processamento, passamos para a fase de representação do texto. Dentre as diversas formas de representação disponíveis na literatura, uma das mais populares são as baseadas em *word embeddings*. Um *embedding* é um tipo de representação de

[2]NLTK: `https://www.nltk.org/howto/portuguese_en.html#stopwords`
[3]spaCy: `https://spacy.io/models/pt`

texto em que palavras com significado parecido possuem representações vetoriais similares. Portanto, palavras usadas no mesmo contexto acabam por ficar próximas no espaço vetorial. A escolha do modelo de *word embedding* pode, portanto, ter um impacto significativo na precisão da tarefa de classificação.

Neste estudo, três modelos de *word embedding* são avaliados: GloVe, Word2Vec e Wang2Vec. Tais modelos objetivam gerar representações vetoriais de palavras que capturem suas relações semânticas e sintáticas. O GloVe [Pennington et al. 2014] é um modelo que utiliza matrizes de coocorrência de palavras para gerar representações vetoriais, enquanto o Word2Vec e o Wang2Vec são modelos que utilizam abordagens baseadas em redes neurais para gerar representações vetoriais. Em particular, o modelo Word2Vec [Church 2017] é baseado em uma rede neural que aprende a prever o contexto em que uma palavra aparece e gera *embeddings* de palavras com base nos pesos aprendidos. Já o modelo Wang2Vec [Poetsch et al. 2019] é um modelo mais recente que utiliza uma arquitetura semelhante ao Word2Vec, mas com um mecanismo de compartilhamento de peso que melhora a eficiência e escalabilidade do processo de treinamento.

3.4. Classificação de Documentos de Texto

Para classificar documentos de licitação, existem diferentes algoritmos de classificação, desde os clássicos, como Naive Bayes e Árvores de Decisão, até abordagens mais sofisticadas, como modelos baseados em redes neurais [Coelho et al. 2022, Noguti et al. 2020]. No entanto, classificadores clássicos tendem a ser mais simples e menos flexíveis em comparação com abordagens baseadas em redes neurais, o que pode limitar sua capacidade de capturar nuances complexas e informações contextuais em dados textuais. Por sua vez, as redes neurais, como as redes LSTM, têm a capacidade de lidar com tarefas mais complexas, como a classificação de documentos de licitação, devido à sua habilidade de modelar sequências de dados e capturar relações de dependência de longo prazo.

Em particular, as redes LSTM são uma classe de redes neurais recorrentes que têm se mostrado eficazes em tarefas de processamento de linguagem natural devido à sua capacidade de capturar informações contextuais de longo prazo. Como em qualquer outra rede neural, a LSTM pode ter várias camadas ocultas e, à medida que passa por todas as camadas, as informações relevantes são mantidas e todas as informações irrelevantes são descartadas em cada célula. Assim, LSTMs memorizam as informações relevantes que são importantes, descartando o que não importa.

No entanto, a eficácia da abordagem depende de técnicas de pré-processamento e representação de texto adequadas, uma vez que as redes neurais são sensíveis à qualidade dos dados de entrada. Portanto, é fundamental avaliar diferentes abordagens de pré-processamento e representação de texto ao usar redes LSTM para classificar documentos de licitação. Essa avaliação ajuda a determinar qual técnica é mais eficaz para a tarefa em questão e a maximizar o desempenho da classificação.

4. Configuração Experimental

Esta seção apresenta os detalhes técnicos sobre a configuração e avaliação experimental. Além das diferentes abordagens de pré-processamento e representação de texto, duas divisões de treino-teste estratificada foram avaliadas: (i) estratificação por classe; e (ii) estratificação por classe e cidade. Essas divisões de treino-teste foram escolhidas para

avaliar como a estratificação pode impactar no desempenho da classificação de documentos de licitação. Para garantir uma avaliação robusta do desempenho das diferentes combinações avaliadas, a validação cruzada foi aplicada em ambas as divisões.

1. **Estratificação por classe.** Considera que as classes de documentos de licitação podem ter diferentes frequências na base de dados, e a divisão é feita de forma que a proporção de cada classe seja mantida em cada fold;
2. **Estratificação por classe e cidade.** Considera que documentos de licitação de uma mesma cidade podem apresentar características semelhantes, e, portanto, a divisão é feita de forma que a proporção de cada classe e cidade seja mantida em cada fold.

Portanto, são realizados 24 experimentos, um para cada combinação das duas configurações experimentais, quatro abordagens de pré-processamento e três modelos de *word embeddings*. Cada experimento foi realizado com validação cruzada de *5-folds*, não foram utilizadas mais *folds* devido ao elevado tempo de processamento.

Modelos *word embedding*. Os três modelos *word embedding* utilizados foram obtidos do NILC-Embeddings,[4] que é um repositório destinado ao armazenamento e compartilhamento de *word embeddings* para a Língua Portuguesa. O repositório traz vetores gerados a partir de um grande corpus do português do Brasil e português europeu, de fontes e gêneros variados. O treinamento dos vetores ocorreu em quatro modelos diferentes, incluindo o GloVe, Word2Vec e Wang2Vec. Para cada modelo, foram disponibilizados vetores de palavras gerados em várias dimensões. Neste estudo, foram considerados os modelos GloVe, Word2Vec e Wang2Vec com 600 dimensões e a abordagem SKIP-GRAM.

Configuração da LSTM. Para a construção do classificador, foi escolhida uma arquitetura de rede LSTM com três camadas de recorrência. Além disso, foi adicionada uma camada de *dropout* com probabilidade de 20% para evitar overfitting. O modelo foi treinado utilizando a técnica de otimização de Adam, com uma taxa de aprendizado inicial de 0,001 e uma taxa de decaimento de 1e-6. O número de épocas de treinamento foi definido em 8, e o tamanho do lote de treinamento foi ajustado em 64.

Métricas de Avaliação. Para a avaliação dos experimentos, foram consideradas duas métricas: F1-Macro e F1-Weighted. A F1-Macro é a média harmônica das pontuações F1 para cada classe e é útil para avaliar a capacidade do modelo de lidar com classes desbalanceadas, enquanto a F1-Weighted é a média harmônica das pontuações F1 ponderadas pelo número de amostras em cada classe e é mais adequada para avaliar a precisão geral do modelo em relação a todas as classes. Essas métricas foram calculadas para cada modelo avaliado no conjunto de dados de licitações públicas.

5. Resultados Experimentais

Esta seção apresenta os resultados obtidos para as métricas F1-Macro e F1-Weighted para os 24 experimentos realizados com a LSTM, utilizando a validação cruzada de *5-folds*. Em particular, a Tabela 1 mostra que os resultados são bastante similares para as diferentes combinações experimentais avaliadas. O melhor resultado para a F1-Macro (0,971) e F1-Weighted (0,989) foi para a configuração experimental com estratificação por classe, pré-processamento base+ e a representação usando o modelo GloVe.

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

Tabela 1. Comparação das 24 configurações experimentais na classificação dos documentos de licitação, utilizando a LSTM. O melhor resultado para cada estratificação e métrica é sublinhado.

		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	$\underline{0.971 \pm 0.012}$	$\underline{0.989 \pm 0.004}$	$\underline{0.969 \pm 0.005}$	$\underline{0.989 \pm 0.003}$
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

Para avaliar se existe diferença significativa entre as configurações experimentais, as Figuras 3a e 3b apresentam, respectivamente, a F1-Macro e F1-Weighted com teste de Kruskal-Wallis e Wilcoxon pareado para cada configuração experimental. Ambos os testes são não paramétricos e são utilizados para comparar amostras independentes. O teste de Wilcoxon pareado permite comparar apenas duas amostras, enquanto o Kruskal-Wallis permite a comparação de três ou mais amostras [Kim 2014]. A análise dos *p-value* (p) do teste de Kruskal-Wallis nas Figuras 3a e 3b revela que não é possível rejeitar a hipótese nula de que as medianas da F1-Macro e F1-Weighted de cada experimento são as mesmas, pois o *p-value* é maior que 0,05 (ou seja, probabilidade maior que 5%). Portanto, há indícios de que a diferença observada entre os experimentos pode ser devido ao acaso e, portanto, não há diferença significativa entre os experimentos.

Uma exceção é a comparação entre os pré-processamentos utilizando o modelo Wang2Vec com a estratificação por classe para a métrica de avaliação F1-Macro, pois o *p-value* é igual a 0,033, ou seja, pouco menor que 0,05. Entretanto, para essa mesma configuração experimental para a métrica F1-Weighted, o *p-value* possui um valor mais alto, igual a 0,09. Por isso, também foi considerado que não há diferença significativa entre os experimentos com diferentes abordagens de pré-processamento na configuração experimental de ambas estratificações.

Em relação aos resultados do teste de Wilcoxon pareado, observa-se que todos os *p-values* são superiores a 0,05, o que indica a ausência de diferença significativa entre as configurações experimentais quando comparadas duas a duas. Esse resultado reforça o obtido para o teste de Kruskal-Wallis, sugerindo que as diferentes combinações de pré-processamento e representação de texto avaliadas não afetaram de forma significativa o desempenho da rede LSTM na tarefa de classificação de documentos de licitação.

Por meio do nosso conhecimento do problema, é possível que os resultados obtidos possam ser explicados por três razões, são elas: característica dos documentos de licitação, que são textos longos, com pouca padronização e em português; alto número de classes que podem não estar bem representadas pelos documentos coletados; e o modelo utilizado ser a LSTM que, apesar de guardar a informação de sequências de textos longos,

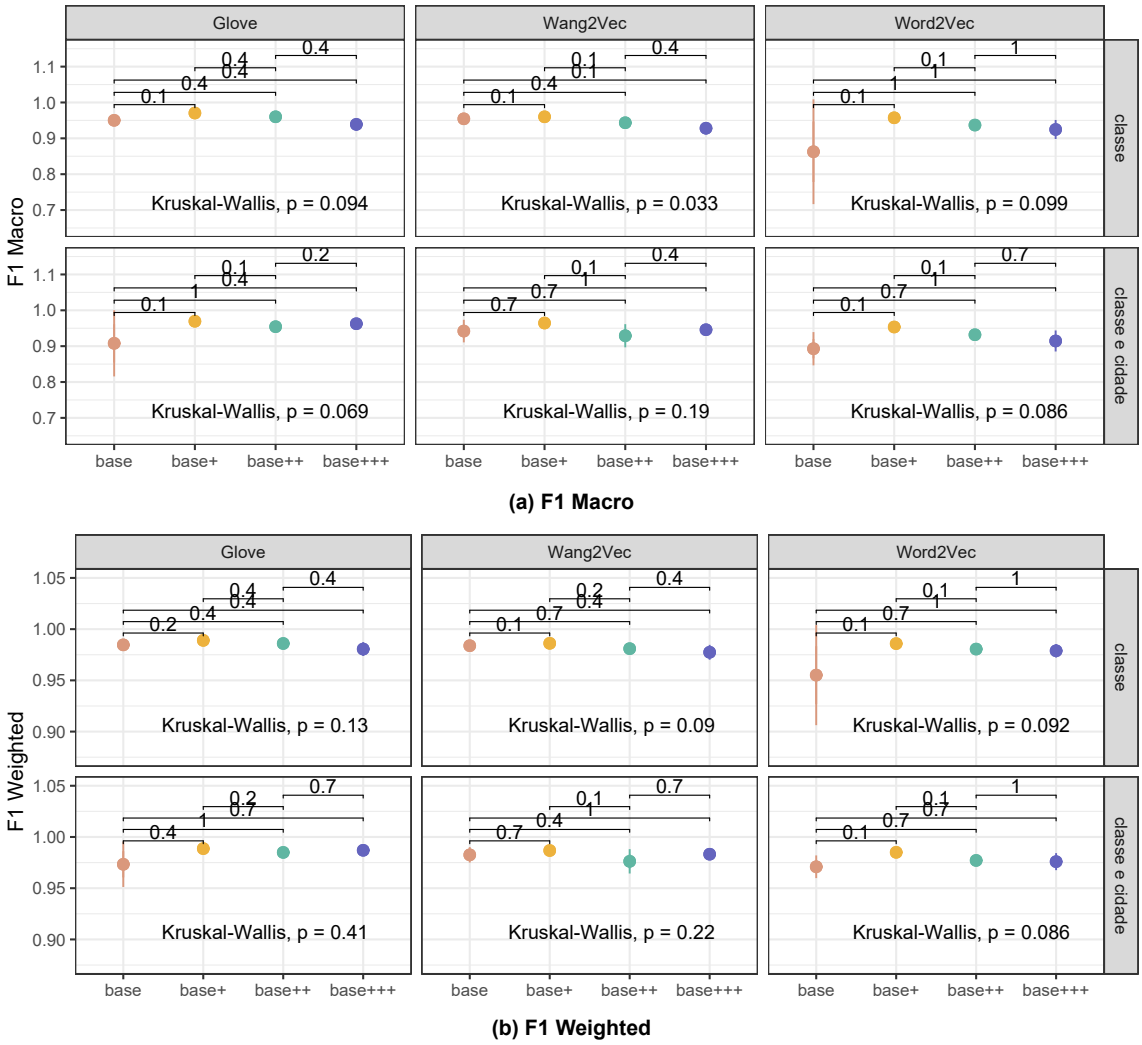
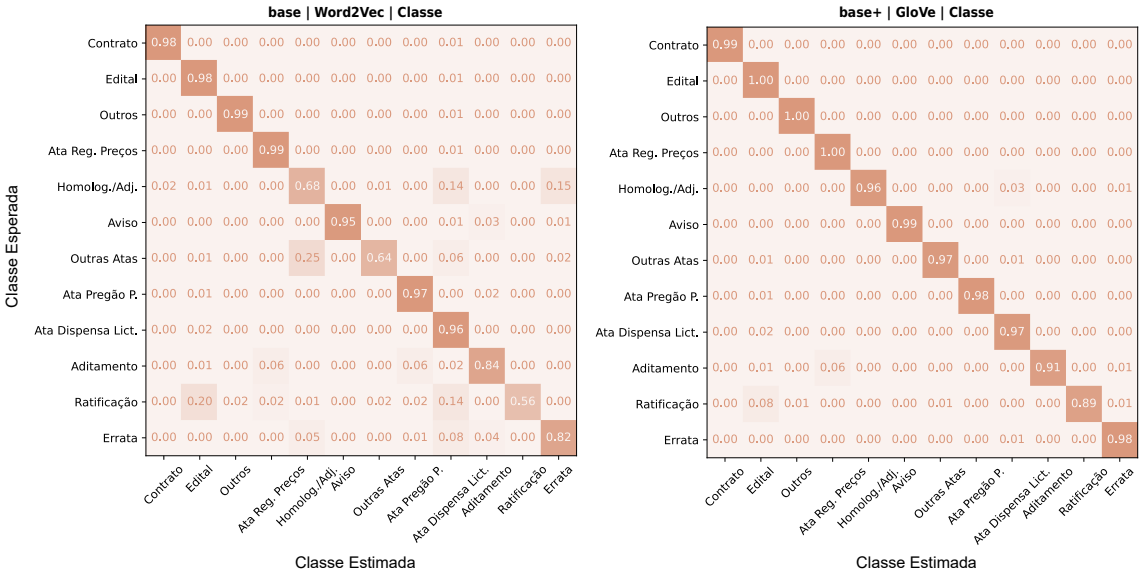


Figura 3. Resultado da classificação das configurações experimentais conforme (a) F1 Macro e (b) F1 Weighted. Testes estatísticos (Kruskal-Wallis e Wilcoxon pareado) foram usados para calcular a diferença significativa entre as abordagens de pré-processamento. Os valores entre os pontos médios representam o p-valor resultante do teste Wilcoxon pareado.

possui limitações quanto ao tamanho da entrada que a rede neural utiliza para predizer a próxima saída [Zhang et al. 2018].

Com o intuito de ajudar no entendimento dos resultados, a Figura 4 apresenta duas matrizes de confusão considerando as 12 classes definidas para a classificação dos documentos de licitação. A Figura 4a refere-se ao experimento com o menor resultado para a métrica F1-Macro, ou seja, é o experimento com a configuração experimental de estratificação apenas por classe, pré-processamento do tipo base e representação textual com modelo Word2Vec. Os resultados mostram que a classe mais difícil de classificar foi a *ratificação*, onde 20% dos documentos dessa classe são preditos como *edital* e 14% como *ata pregão presencial*. Ao comparar com a matriz de confusão da Figura 4b, resultante do experimento com o maior resultado para a métrica F1-Macro (configuração experimental com estratificação por classe, pré-processamento do tipo base+ e representação textual com modelo GloVe), a *ratificação* continua sendo a classe com mais erros,



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

Figura 4. Matriz de confusão da pior e melhor configuração experimental.

sendo confundida principalmente com *edital* (8%).

Finalmente, a matriz de correlação indica que há um problema na representação das classes *ratificação* e *edital* por ser onde o modelo de classificação mais erra nos diferentes experimentos. Não apresentamos as 24 matrizes de confusão por limitação de espaço e por serem bastante similares. Assim, é importante melhor analisar como os textos estão representados nas diferentes classes de forma que elas tenham uma melhor representatividade dos documentos de licitação, o que está de acordo com o resultado obtido por [Souza Júnior et al. 2022].

6. Conclusão

Neste artigo, foi avaliado o impacto das etapas de pré-processamento e representação textual na eficácia da classificação de documentos de licitação. Especificamente, a partir de um conjunto de documentos de licitações, foram avaliadas quatro abordagens de pré-processamento e três modelos de representação textual em um classificador baseado em uma rede neural LSTM. Os resultados mostraram que não há diferença estatística para as métricas F1-Macro e F1-Weighted obtidas para as 24 configurações experimentais avaliadas. Ou seja, a performance do classificador pode estar associada à natureza dos documentos, à quantidade de classes definidas e/ou a limitações do modelo LSTM.

Limitações e Trabalhos Futuros. As configurações experimentais apresentadas neste trabalho não permitem avaliar se o modelo de classificação proposto consegue generalizar para novos documentos de licitação. Por isso, como trabalhos futuros, planeja-se realizar uma configuração experimental para melhor avaliar essa generalização. Além disso, o estado-da-arte na classificação de documentos de texto é o modelo BERT, assim, planeja-se avaliar o impacto nos resultados ao alterar o modelo da LSTM para o BERT.

Agradecimentos. Este trabalho foi financiado pelo Ministério Público de Minas Gerais, pelo Programa de Capacidades Analíticas, e pelo CNPq, CAPES e FAPEMIG.

Referências

[Albalawi et al. 2021] Albalawi, Y., Buckley, J., and Nikolov, N. S. (2021). Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting arabic health information on social media. *J. Big Data*, 8(1):95.

[Bambroo and Awasthi 2021] Bambroo, P. and Awasthi, A. (2021). Legaldb: long distilbert for legal document classification. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–4. IEEE.

[Belém et al. 2022] Belém, F. M., Ganem, M., França, C., Carvalho, M., Laender, A. H. F., and Gonçalves, M. A. (2022). Reforço e delimitação contextual para reconhecimento de entidades e relações em documentos oficiais. In *SBBD*, pages 292–303. SBC.

[Church 2017] Church, K. W. (2017). Word2vec. *Natural Language Engineering*, 23(1):155–162.

[Coelho et al. 2022] Coelho, G. M., Ramos, A. C., de Sousa, J., Cavaliere, M., de Lima, M. J., Mangeth, A., Frajhof, I. Z., Cury, C., and Casanova, M. A. (2022). Text classification in the brazilian legal domain. In *ICEIS (1)*, pages 355–363.

[de Araujo et al. 2023] de Araujo, P. H. L., de Almeida, A. P. G. S., Braz, F. A., da Silva, N. C., de Barros Vidal, F., and de Campos, T. E. (2023). Sequence-aware multimodal page classification of brazilian legal documents. *Int. J. Document Anal. Recognit.*, 26(1):33–49.

[Kim 2014] Kim, H.-Y. (2014). Statistical notes for clinical researchers: Nonparametric statistical methods: 2. nonparametric methods for comparing three or more groups and repeated measures. *Restorative Dentistry & Endodontics*, 39(4):329–332.

[Lima et al. 2020] Lima, M., Silva, R., Lopes de Souza Mendes, F., R. de Carvalho, L., Araujo, A., and de Barros Vidal, F. (2020). 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*, pages 1580–1588, Online. Association for Computational Linguistics.

[Luz de Araujo et al. 2020] Luz de Araujo, P. H., de Campos, T. E., Ataides Braz, F., and Correia da Silva, N. (2020). VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France. European Language Resources Association.

[Noguti et al. 2020] Noguti, M. Y., Vellasques, E., and Oliveira, L. S. (2020). 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.

[Oliveira et al. 2022] Oliveira, G. P., Reis, A. P. G., Mendes, B. M. A., Bacha, C. A., Costa, L. L., Canguçu, G. L., Silva, M. O., Caetano, V., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2022). Ferramentas open-source de qualidade de dados para licitações públicas: Uma análise comparativa. In *SBBD*, pages 116–127. SBC.

[Pennington et al. 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL.

[Poetsch et al. 2019] Poetsch, M., Correa, U. B., and de Freitas, L. A. (2019). A word embedding analysis towards ontology enrichment. *Res. Comput. Sci.*, 148(11):153–164.

[Silva et al. 2022] Silva, M. O., Paula, A. F., Oliveira, G. P., Vaz, I. A. D., Hott, H., Gomide, L. D., Reis, A. P. G., Mendes, B. M. A., Bacha, C. A., Costa, L. L., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2022). LiPSet: Um conjunto de Dados com Documentos Rotulados de Licitações Públicas. In *SBBD DSW*, pages 13–24, Porto Alegre, RS, Brasil. SBC.

[Souza Júnior et al. 2022] Souza Júnior, A. P., Cecilio, P., Viegas, F., Cunha, W., de Albergaria, E. T., and da Rocha, L. C. D. (2022). Evaluating topic modeling pre-processing pipelines for portuguese texts. In *WebMedia*, pages 191–201. ACM.

[Zhang et al. 2018] Zhang, J., Li, Y., Tian, J., and Li, T. (2018). Lstm-cnn hybrid model for text classification. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1675–1680. IEEE.