PEER REVIEW, Vol. 6, Nº 15, 2024
DOI: 10.53660/PRW-2539-4555
ISSN: 1541-1389

Identificação de Duplicidades de Materiais em Processos de Compras Públicas Usando PLN: um estudo de caso em uma Universidade Brasileira

Identifying Duplicates Materials in Public Procurement Processes Using NLP: a Case Study in a Brazilian University

Fernanda Forbici
ORCID: https://orcid.org/0000-0001-8784-6233
Universidade Federal de Santa Catarina, Brasil
E-mail: fernandaforbici@gmail.com
Joelias Silva Pinto Junior
ORCID: https://orcid.org/0000-0001-6810-5878
Universidade Federal de Santa Catarina, Brasil
E-mail: joeliasjunior@gmail.com
Vinícius Faria Culmant Ramos
ORCID: https://orcid.org/0000-0002-8319-743X
Universidade Federal de Santa Catarina, Brasil
E-mail: v.ramos@ufsc.br
João Artur de Souza
ORCID: https://orcid.org/0000-0002-7133-8944
Universidade Federal de Santa Catarina, Brasil
E-mail: joao.artur@ufsc.br

RESUMO

Instituições públicas brasileiras necessitam atender a diversas especificidades para realizarem processos de compras idôneos e eficientes, como não comprar produtos e serviços em duplicidade. No entanto, esse é um problema que acontece de forma recorrente nas aglutinações de demandas. Para mitigar esse problema, este trabalho apresenta técnicas de Processamento de Linguagem Natural (PLN) desenvolvidas para identificar duplicidades nas demandas de compras, aplicadas a um estudo de caso com uma universidade pública brasileira. Utilizando técnicas de embeddings estáticos e dinâmicos, a ferramenta desenvolvida foi capaz de detectar termos similares em descrições de materiais, melhorando a eficiência do processo de aquisição. Os resultados mostraram que modelos de terceira geração de embeddings apresentaram maior precisão na recuperação de termos. O agrupamento de itens, baseado nos *embeddings* das descrições apresentou boas métricas, mas para novos itens a proporção de correspondências corretas foi mediana. A pesquisa permitiu identificar os avanços dos word embeddings dinâmicos, em um contexto e idioma específico, em relação aos modelos anteriores com pouco esforço demandado. Este trabalho contribui para a otimização dos processos licitatórios, reduzindo redundâncias e promovendo maior transparência.

Palavras-chave: Análise de Similaridade; Processamento de Linguagem Natural; Compras Públicas;

Recebido: 01/08/2024 | Aceito: 01/09/2024 | Publicado: 04/09/2024

ABSTRACT

Brazilian public institutions need to meet several specific requirements to carry out appropriate and efficient purchasing processes, such as not purchasing duplicate products and services. However, this is a problem that occurs frequently in the clustering of demands. To mitigate this problem, this paper presents Natural Language Processing (NLP) techniques developed to identify duplicates in purchasing demands, applied to a case study with a Brazilian public university. Using static and dynamic embedding techniques, the developed tool was able to detect similar terms in material descriptions, improving the efficiency of the procurement process. The results showed that third-generation embedding models presented greater accuracy in term retrieval. The grouping of items, based on the embeddings of the descriptions, presented good metrics, but for new items the proportion of correct matches was average. The research allowed us to identify the advances of dynamic word embeddings, in a specific context and language, in relation to previous models with little effort required. This work contributes to the optimization of bidding processes, reducing redundancies and promoting greater transparency.

Keywords: Similarity Analysis; Natural Language Processing; Public Procurement

INTRODUÇÃO

A burocracia estatal vem dando lugar a novos paradigmas de gestão e avanços tecnológicos. A emergência do conceito de "Governo 4.0" representa um salto qualitativo nesse trajeto, simbolizando a integração de tecnologias avançadas e práticas inovadoras na gestão pública (PINHEIRO, 2022). Com o intuito de evitar falhas nas compras públicas, este estudo tem o objetivo de auxiliar na identificação de similaridade de materiais demandados em um mesmo processo licitatório de bens e serviços em um órgão público.

A conformidade com a legislação que veta preços diferentes de um mesmo objeto, salvo as exceções especificadas na Lei 14133/21, é mandatória nas compras públicas, e contribui diretamente para manutenção da legalidade, impessoalidade, moralidade, publicidade, eficiência, do interesse público, da probidade administrativa, da igualdade, do planejamento, da transparência, da eficácia, vinculação a edital, segurança jurídica, razoabilidade, competitividade, proporcionalidade, e da celeridade, da economicidade e do desenvolvimento nacional sustentável (BRASIL, 2021b).

Ainda em relação ao *compliance* das organizações públicas, o Decreto 10947/22 define que as demandas devem ser consolidadas, agregando os objetos de mesma natureza com vistas à racionalização dos esforços de contratação (BRASIL, 2022c). Neste contexto, a identificação de objetos duplicados na etapa de aglutinação e a classificação

destes itens para a consolidação do processo de compra representa um avanço significativo na otimização e transparência no processo de licitação.

Selecionar a proposta mais vantajosa, contudo, nem sempre é um procedimento fácil. No caso das universidades, a padronização na nomenclatura de materiais por parte dos setores demandantes e a variedade de insumos inerentes à natureza desse tipo de instituição interferem no processo de compras. Isso ocorre devido ao fato de existirem materiais utilizados nas atividades de ensino, pesquisa e extensão com especificidades técnicas que somente o solicitante da compra tem o conhecimento, não dispondo, o departamento de compras, de especialistas de todas as áreas para analisar todas as especificações.

Sob esta perspectiva, esta pesquisa investigou técnicas de Processamento de Linguagem Natural (PLN) para identificar duplicidades nas demandas de compras, visando contribuir para a eficiência e transparência dos processos licitatórios.

A pesquisa foi aplicada a uma universidade pública do estado de Santa Catarina, no Brasil, onde foi desenvolvido um sistema utilizando técnicas avançadas de inteligência artificial, nomeadamente *embeddings*, uma representação vetorial densa que representa palavras através das relações semânticas e contextuais (SENO et al., 2023a), para melhorias na busca semântica entre os suprimentos a serem comprados por meio de sinônimos. A sistemática de identificação de duplicidades foi implementada por meio da similaridade semântica, ou seja, cálculo da semelhança entre palavras ou sentenças em relação ao seu significado (RODRIGUES, 2023). Isto se deu por meio do cálculo da distância do cosseno, comparando o cosseno do ângulo entre os dois vetores, onde quanto menor o ângulo, maior a similaridade (CORTES et al., 2024).

Para identificar a capacidade de recuperar termos similares ao termo de busca, foram gerados vetores indexados das descrições dos materiais, por meio dos *corpora*, que são conjunto de dados linguísticos organizados (FREITAS, 2023) pré-treinados em português ou multilingual. Diversos termos apresentaram boas acurácias na recuperação, sendo que os modelos de terceira geração de *embeddings*, ou *embeddings* dinâmicos, gerados por *Transformers*, apresentaram um percentual de similaridade maior. O agrupamento de itens, baseado nos *embeddings* das descrições apresentou boas métricas, mas para novos itens a proporção de correspondências corretas foi mediana. Esta pesquisa contribuiu para identificar os avanços dos *word embeddings* dinâmicos, em um contexto e idioma específico, em relação aos modelos anteriores (de segunda geração ou

embeddings estáticos) com pouco esforço demandado. Foi possível notar em testes com alguns termos técnicos ou específicos que há oportunidade de continuidade da pesquisa para adaptações com técnicas que possam melhorar a acurácia da similaridade semântica da recuperação deste tipo de termos.

ESTUDOS RELACIONADOS

Diversos pesquisadores têm utilizado estratégias tecnológicas para aprimorar produtos e serviços ou analisar contextos de instituições públicas. Modrušan, Rabuzin e Mršić (2020) aplicaram mineração de texto avançada para melhorar a eficiência no processo de aquisição pública e detectar fraudes, utilizando algoritmos como LR (*Logistic Regression*), SVM (*Support Vector Machine*) e Naive Bayes. Siciliani et al. (2023) desenvolveram um sistema de suporte à decisão para compras públicas com uso de IA (Inteligência Artificial*)* e PLN (Processamento de Linguagem Natural), incluindo um *dashboard* com indicador de risco de conluio. Bruckner e Vencovský (2020) usaram mineração de texto para extrair preços de contratos, mas os resultados não atingiram as expectativas, pois ainda demandava validação manual. Torres-Berru, Lopez-Batista e Zhingre (2022) consideraram o potencial de análise textual para identificação de corrupção e viés no processo de compras.

Fantoni et al. (2021) enfrentaram desafios ao converter termos de contratos em especificações técnicas, e Artamonov et al. (2022) construíram uma base de conhecimento para domínios específicos, como ferroviário e nuclear. Lee et al. (2023) propuseram uma estrutura para padronizar itens de obras em contratos, enquanto Souza et al. (2023) utilizaram técnicas de PNL para traduzir dados de contratos de construção.

A inteligência artificial em compras públicas é aplicada em várias etapas do processo, desde a padronização de dados (Lee et al., 2023; Silva et al., 2023; Jaques de Souza et al., 2023), análise de fraudes (Lima et al., 2023; Modrušan et al., 2020), gestão de contratos (Fantoni et al., 2021), até a extração de preços (Bruckner e Vencovský, 2020). A maioria dos estudos utiliza dados abertos governamentais, alguns utilizando ferramentas de raspagem de dados (*scraping)* (Lima et al., 2020; Souza et al., 2023; Artamonov et al., 2022; Bifulco et al., 2021), enfrentando desafios de padronização na estrutura de arquivos e descrição dos itens licitados (Lee et al., 2023).

Lima et al. (2023) destacaram a competitividade do BERT (*Bidirectional Encoder Representation from Transformers*) com o modelo Bertimbau na extração de indicadores de fraudes. Alvarez-Rodrigues et al. (2014) focaram na aplicação de web semântica para otimizar o acesso a informações em e-Procurement na União Europeia. Silva et al. (2023) identificaram sobrepreços em licitações públicas através de PLN. Feitosa e Pinheiro (2017) propuseram a Avaliação de Similaridade Semântica e Inferência Textual (ASSIN) para o português. Pinheiro et al. (2017) e Oliveira et al. (2017) desenvolveram métricas para calcular similaridade entre sentenças em português.

MÉTODOS

O estudo está caracterizado como uma pesquisa aplicada de abordagem qualitativa (VERGARA, 2012). A investigação adotou uma perspectiva pragmática, com enfoque na compreensão da realidade e na interpretação dos dados, buscando entender a natureza da realidade como algo dinâmico e multifacetado, onde o conhecimento é construído por meio da interação entre o pesquisador e o contexto que o estudo é realizado (CRESWELL, 2010).

A investigação também é identificada, em relação aos seus objetivos, como uma pesquisa exploratória e de natureza prescritiva, pela qual espera-se como produto final o conhecimento aprofundado sobre determinado tema (GIL, 2019).

O locus do estudo de caso foi o departamento de compras de uma universidade pública do estado de Santa Catarina, com 58 anos de existência e atuando nas áreas de ensino, pesquisa e extensão. Como técnicas de pesquisa foram utilizadas a entrevista e a observação participante. No caso das entrevistas foram selecionados como sujeitos a Coordenadoria do Setor de Compras de um campus da universidade e dois servidores do mesmo departamento, responsáveis pela execução das compras públicas, sendo a coordenadoria do sexo feminino, técnica universitária de desenvolvimento há 12 anos, e os outros servidores, do sexo masculino, técnicos universitários, há mais de 15 anos na instituição. O objetivo da entrevista foi avaliar as opiniões dos servidores da área de compras e dos setores demandantes sobre como se desenvolve o processo de compras, desde a solicitação dos demandantes até a finalização do processo, como são formalizadas e registradas as demandas em cada fase, o acompanhamento por parte dos requerentes, identificando as dificuldades operacionais, entre outros.

Para o desenvolvimento do artefato tecnológico foram utilizadas técnicas de PLN para identificar duplicidades nas demandas de compras. Os métodos envolveram o uso de *word embeddings* estáticos e dinâmicos para análise de similaridade semântica. Os dados, demonstrados no Quadro 1, foram coletados a partir das descrições de materiais registrados em um sistema de gestão de compras e processados utilizando *corpora* pré-treinados em português ou multilinguais.

Quadro 1 - Registro do material "pisseta"

CAMPO	DESCRIÇÃO
ID	154329
CIASC	05538-7-161
Tipo	30 - CONS
Grupo	61 – LABORATORIO, EQUIPAMENTOS E INSTRUMENTAÇÃO
Classe	12 – VIDRARIA E/OU SIMILARES DE LABORATÓRIO
Natureza Despesa	33903035 – MATERIAL LABORATORIAL
Nome Genérico	FRASCO PARA LABORATORIO, CAPACIDADE 500 ML
Nome Específico	FRASCO POLIETILENO, LAVADOR, CAPACIDADE 500ML, BICO CURVO GRADUADO.
Especificação	Frasco Lavador (Pisseta), tampa com bico curvo, confeccionado em Polietileno e Graduado, volume 500 mL
Orçamento	Não

Fonte: Elaborado pelos autores (2024)

As técnicas e ferramentas utilizadas para execução dos processos, que neste estudo representam as etapas de geração dos modelos e avaliação de similaridade, são explicadas abaixo, seguidas das estratégias utilizadas para realização das três atividades deste estudo.

Técnicas e Ferramentas Utilizadas

Os códigos fontes da geração dos modelos, dos processos e da interface de avaliação dos usuários estão disponibilizados em repositório público[1], sob a licença Apache License 2.0.

[1] https://codigos.ufsc.br/fernanda.forbici/egc

Utilizou-se a linguagem de programação Python 3 criando *notebooks* na ferramenta Google Colab Pro, onde é possível alterar o ambiente de execução aumentando os recursos computacionais, estando disponível até 12.7 GB de RAM e 96.05 unidades de computação.

Foram utilizadas duas fontes de dados neste estudo para criar os vetores indexados, além dos modelos pré-treinados conforme o modelo de PLN utilizado (SBERT, BERT, Word2Vec, FastText). A primeira fonte de dados foi obtida por um *web scraping* que coletou a descrição de materiais de sites de comércio eletrônico, gerando 13137 vocábulos. A segunda fonte de dados foi extraída do banco de dados MySQL do Sistema de Compras, com 135580 registros e convertidos para o formato CSV (*Comma-Separated Values*).

Para realizar o segundo processo, houve a necessidade de coletar dados das aglutinações dos materiais através do sistema de compras, extraindo CSV destes dados do banco de dados do MySQL com a estrutura de número do item e descrição do item. Para a avaliação extrínseca do processo 2 foi disponibilizada uma interface de consulta, para os usuários consultarem os termos e avaliarem os resultados das similaridades através de opções baseadas na escala *Likert*.

Para o processo 3 foi gerado um arquivo CSV com os últimos pedidos feitos pelos requisitantes que ainda não tinham sido aglutinados, de forma a avaliar a separação desses materiais por áreas conforme orienta o Decreto n° 10.947.

No pré-processamento, os dados foram convertidos para caracteres minúsculos e, usando os recursos da biblioteca NLTK (*Natural Language ToolKit*), foi executada a *tokenização*, remoção de *stopwords* e *stemming* com o submódulo RSLP, do Python. Por fim, foram removidos caracteres especiais e dígitos numéricos, resultando nos *tokens* representados pela nuvem de palavras da Figura 2.

Figura 2 - Nuvem de palavras os tokens extraídos do *dataset*

Fonte: Elaborado pelos autores (2024)

Os resultados foram medidos utilizando métricas de similaridade e acurácia, com o objetivo de identificar o modelo mais eficaz para o contexto específico da universidade.

No Quadro 2 está representada uma síntese desta pesquisa, especificando as etapas dos três processos:

Quadro 2 - Consolidação dos Processos Executados

Descrição	**Entrada**	**Processamento**	**Saída**
1. Geração dos modelos estáticos 2. Geração dos modelos dinâmicos	1. Dados pré-processados 2. Dados pré-processados e modelos pré-treinados	1. seleciona a técnica (algoritmo) de *embedding* e a arquitetura, através do Gensim e indexa os tokens derivados dos dados pré-processados 2. De acordo com o modelo pré-treinado carregado, utilizando a biblioteca torch e Transformers ou SentenceTransformer, é codificado o token derivado dos dados pré-	1. Vetor indexado de 150 dimensões para cada *token* derivado dos dados pré-processados 2. Vetor indexado de 768 dimensões para cada *token* derivado dos dados pré-processados, no caso do BERT ou um vetor indexado de 768 dimensões para sentenças inteiras derivadas dos dados pré-processados, no caso do SBERT

		processados	
Duplicidade de materiais através Similaridade entre termos	Modelos de embeddings gerados no processo 1 (no caso dos modelos estáticos, os modelos são combinados entre eles) e termos de busca	É gerado o *embedding* do termo de busca e calculado a similaridade do cosseno, pela biblioteca sklearn, entre os *embeddings*	Percentual de similaridade entre os termos (*embeddings*)

Fonte: Elaborado pelos autores (2024)

Variamos a concatenação dos campos do *dataset* para analisar as possibilidades dos vocabulários gerados em cada modelo, além de avaliar o custo computacional e o tamanho destes modelos. Também foram avaliados três modelos de *embeddings* padronizados pelo NILC (Núcleo Interinstitucional de Linguística Computacional), sendo um GloVe, um FastText e um Word2Vec.

Desta forma, abaixo é descrito cada etapa do estudo, na seqüência em que foram executados:

Processo 1

Nesta etapa foi realizado o pré-processamento das descrições dos materiais e executados testes com modelos de ML(*Machine Learning*) para gerar os *embeddings* estáticos e DL (*Deep Learning*) para gerar os *embeddings* dinâmicos.

Para gerar os modelos estáticos, que são representações vetoriais de palavras baseadas no contexto em que aparecem (SENO et al., 2023b; JURAFSKY e MARTIN, 2023), foi utilizada a biblioteca Gensim do Python, e com o parâmetro vector_size = 150; window = 5; min_count = 1; workers=10; sg = 1, epochs = 50. Para gerar o modelo dinâmico BERT, foi usada a biblioteca torch e a biblioteca Transformers do Python, com o modelo pré-treinado neuralmind/bert-base-portuguese-cased, e para o modelo dinâmico SBERT (*Sentences BERT*) foi utilizada a biblioteca SentencesTransformers, e o modelo pré-treinado paraphrase-multilingual-mpnet-base-v2, ambos com 768 dimensões.

A criação dos modelos contextualizados estáticos utilizou técnicas de Word2Vec e FastText, e a arquitetura adotada foi a SkipGram. Para modelo híbrido (*Stacking Embedding*), que é a combinação dos modelos estáticos (*Word2Vec, FastText* e modelos pré-treinados do NILC) gerados individualmente e amplamente reconhecido por sua

eficácia e potencial promissor (SANTOS, 2019), assumiu-se que os dados de processamento podem ser obtidos pelo somatório desses dados dos modelos mencionados anteriormente.

Também foram avaliados resultados obtidos com os modelos contextualizados dinâmicos, BERT e SBERT, ambos baseados na arquitetura de *Transformers*, uma arquitetura de rede neural não recorrente. Eles fornecem representações vetoriais de contextos das palavras em ambas as direções, permitindo uma compreensão mais profunda do significado (PAES et al., 2023). Na tentativa de gerar os *embeddings* da descrição dos materiais do *dataset* no modelo BERT, o alto consumo de recursos computacionais foi um impeditivo, portanto foi feito o processamento do modelo contextualizado em lote. Os modelos da arquitetura baseada em *Transformers* apresentaram um desempenho melhor no tempo de consulta, além de apresentarem melhores percentuais de similaridade entre o termo de busca e o termo alvo, como por exemplo os termos geladeira e refrigerador, que apresentaram um resultado de 97,60% de similaridade no modelo SBERT e 96,40% no modelo BERT.

Processo 2

Nesta etapa, pretendeu-se discutir a eficiência do modelo na tarefa de Recuperação de Informação baseada na similaridade e na identificação da duplicidade de materiais aglutinados através da avaliação intrínseca e extrínseca dos modelos gerados, realizando consultas em cada um desses modelos.

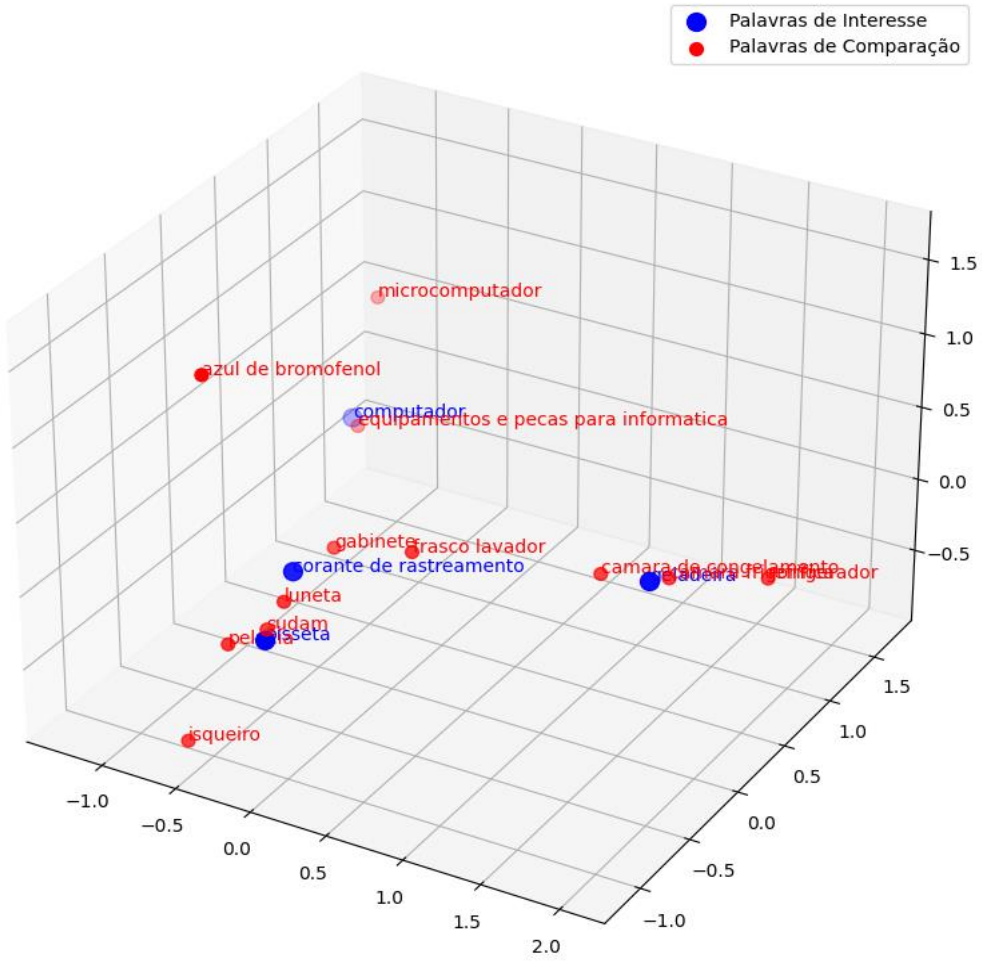
Foi observado que nos modelos individuais, com frequência os termos consultados não eram encontrados no vocabulário (*out-of-vocabulary*, OOV), principalmente para termos técnicos ou específicos. Isto foi superado no modelo híbrido, combinando esses modelos estáticos individuais, mas não alterou significativamente a classificação de similaridade dos termos, e não aumentou o percentual de similaridade, o que talvez não justifique o uso do recurso computacional, de acordo com esses resultados. Verificou-se que a arquitetura *Skip-gram* apresentou um percentual de similaridade de cosseno maior aos termos mais similares em relação a arquitetura CBOW.

Os modelos de *embeddings* baseado em *Transformers* apresentaram um melhor resultado no percentual de similaridade, sendo usado como *feature based*, ou seja, as representações vetoriais geradas pelo modelo pré-treinado são utilizadas como

características (*features*) para alimentar modelos de aprendizado de máquina ou sistemas de análise posteriores, sem ajustar os pesos internos do modelo de *Transformers* durante o treinamento específico da tarefa, mas ainda assim, retornou termos que não foram classificados como similares nas avaliações intrínsecas. Seria necessário a utilização de outras técnicas combinadas, como aprendizagem com utilização de banco de dados vetoriais ou *fine tuning*, onde um modelo pré-treinado é treinado novamente com novos dados.

A similaridade entre o termo de busca e os cinco termos mais similares, calculados através da similaridade por cosseno, avaliados de forma intrínseca pelo percentual de similaridade entre os termos e de forma extrínseca, através de uma interface de avaliação feita pelos usuários do setor de compras, utilizando o modelo híbrido de *word embeddings*, não se mostrou eficaz para o uso de termos técnicos, mas satisfatórios para termos de busca generalistas. A Figura 3 representa a distribuição vetorial dos vocábulos no modelo SBERT, demonstrando que existem outros termos mais próximos do termo de busca que o termo alvo.

Figura 3 - Representação vetorial de vocábulos de exemplo



Fonte: Elaborado pelos autores (2023)

Desta forma, adotou-se a estratégia de identificar termos similares na fase de aglutinação, onde os termos consolidados em uma mesma aglutinação eram apontados quando apresentassem uma similaridade maior que 95% com outro termos presente na mesma aglutinação, trazendo uma efetividade maior no processo de compra e satisfação dos usuários nos testes apresentados.

RESULTADOS

Para atingir o resultado esperado foi desenvolvido um artefato com a implementação de modelos de PLN treinados, para realizar simulações de buscas e recuperação de informações, e uma interface onde o usuário pudesse avaliar a precisão dos resultados. Com o foco de resolver o problema da etapa inicial do processo de

compras da instituição, procurou-se identificar formas eficazes de padronizar a entrada dos dados, como uma parte de soluções mais completas, como sistemas de demandas, orçamentos, entre outros, que em uma etapa posterior auxiliará na otimização do processo de aquisição como um todo. O artefato resultado deste estudo é uma interface para a recuperação da informação para avaliar a similaridade dos termos de busca em um contexto de insumos demandados de uma instituição de ensino superior.

Os principais resultados dos processos conduzidos são apresentados a seguir:

Processo 1:

Os modelos de embeddings estáticos do modelo híbrido, combinando os de contexto com os pré-treinados, apresentaram pouca ocorrência de OOV, assim como os modelos baseados na arquitetura *Transformers*. O Quadro 2 apresenta o custo computacional e o tamanho dos modelos gerados no processo 1, com a seguinte estrutura de dados para a contextualização: Nome genérico, grupo, subgrupo, nome específico e especificação com Word2Vec e FastText com 150 dimensões combinado com os modelos pré-treinados do NILC Word2Vec e FastText, com 100 dimensões. (Modelo 1); Modelo BERT utilizando neuralmind/bert-base-portuguese-cased (Modelo 2); Modelo SBERT utilizando paraphrase-multilingual-mpnet-base-v2 (Modelo 3). No Modelo 1, assumiu-se como custo computacional a soma dos valores resultados para gerar os modelos individualmente. As dimensões e vocabulário não foram somados porque existem tokens repetidos entre os modelos, porém com valores para cada um variando a depender do contexto.

Modelo	CPU Times	Wall Time	Tempo de Treinamento	Dimensões do Modelo	Tamanho do Vocabulário
Word2Vec FastText NILC Word2Vec pré-treinada **Híbrido**	3 µs 2 µs 4 µs **10 µs**	5.96 µs 5.48 µs 6.68 µs **18.84 µs**	42.80 s 191.86 s 50.72 s **145,71 s**	150 150 100	20134 20134 929606
Modelo 2	5 µs	22.6 µs	567.13 s	768	29794
Modelo 3	3 µs	5.72 µs	256.90 s	768	250002

Quadro 2 - Custo computacional para geração dos modelos Fonte: Elaborado pelos autores (2023)

Apesar do tempo decorrido para geração dos modelos de *embeddings* dinâmicos ficarem acima dos *embeddings* estáticos, o percentual de similaridade entre o termo de busca e do termo alvo foram superiores, tanto para termos genéricos quanto para termos mais especializados, e serão demonstrados no Processo 2 decorrentes dos modelos gerados no Processo 1.

Processo 2:

Os modelos de *embeddings* dinâmicos, como BERT e SBERT, mostraram uma precisão significativamente maior na recuperação de termos similares. Os *embeddings* dinâmicos foram capazes de capturar nuances semânticas mais complexas, resultando em uma identificação mais precisa de duplicidades nas descrições dos materiais.

Os termos utilizados, com suas respectivas correspondências, estão relacionados no quadro abaixo (Quadro 3) :

Quadro 3 - Comparação entre Termo de Busca e Termo Alvo por Modelo de Embedding

Termo de Busca	Termo correspondente	Sim. Modelo Híbrido (%)	Sim. SBERT (%)	Sim. BERT (%)
APLICADOR DE AGULHA	APALPADOR DE PRESSÃO	9.10	61.74	93.74
BALDE DE POLIETILENO	BALDE DE PLASTICO	10.43	79.80	96.44
CANETA HIDROCOR	CANETA HIDROGRÁFICA	07.71	83.91	96.00
COMPUTADOR	MICROCOMPUTADOR	10.37	67.25	93.75
CORANTE DE RASTREAMENTO	AZUL DE BROMOFENOL	5.91	38.29	86.05
DISPOSITIVO DE OZONIOTERAPIA	UNIDADE FILTRANTE ESTÉRIL PARA OZONIOTERAPIA	6.33	87.04	93.53
GASOLINA COMUM	ETANOL	5.82	53.68	74.89
GELADEIRA	REFRIGERADOR	10.39	97.60	96.40
HASTE EM PLÁSTICO	SWAB	6.79	41.39	92.82
LUMINARIA DE EMERGENCIA	LAMPADA DE EMERGENCIA	8.73	97.91	98.41
PISSETA	FRASCO LAVADOR	4.33	48.34	92.81

Fonte: Elaborado pelos autores (2023)

A superioridade do percentual de similaridade entre os modelos baseados em *Transformers* pode ter sido influenciado pela quantidade de dimensões destes modelos.

O objetivo nesta etapa, era que o usuário pesquisasse itens similares quando recebesse uma demanda de algum material desconhecido, identificasse se o mesmo item já não havia sido demandado por outro usuário com uma descrição diferente. Na avaliação intrínseca, utilizando apenas o termo de busca, verificou-se que se o termo alvo era um termo técnico, este não foi relacionado entre os termos mais similares na maioria dos modelos (por exemplo “CORANTE DE RASTREAMENTO” foi relacionado com “RASTREADOR GPS”, “ELASTICO NA COR BRANCA” no modelo híbrido, “BOTA”, “PARA-LAMA” no modelo SBERT, “REVELADOR PARA FILME” e “DISTINTIVO PARA QUEPE” no modelo BERT) , mas para termos mais usuais a ocorrência do termo alvo, ou termos da mesma área foi mais frequente (destacados em cinza nas tabelas do apêndice B) (por exemplo “GELADEIRA” retornou “GELADEIRA DOMÉSTICA” e “REFRIGERADOR” nos três modelos), pois havia outros termos com

similaridade maior, mas que na prática não se referenciam ao mesmo produto. No modelo SBERT verificou-se que a recuperação de termos da mesma área do termo de busca foi mais frequente.

Optou-se por investigar a possibilidade de validar a similaridade após essas demandas serem aglutinadas, na etapa posterior à solicitação de demanda no fluxo do processo de compras. Desta forma, verificamos a similaridade entre os itens já aglutinados. Neste processo os resultados foram satisfatórios utilizando o modelo SBERT e BERT, apresentando percentual de similaridade alto onde possivelmente havia pedido de itens duplicados. Nas similaridades um pouco mais baixas (entre 90% a 98%) foram identificados itens duplicados. É importante ressaltar que nesta etapa de avaliação, notou-se que as similaridades muito altas (em média 99%) se referiam a um item em comum, mas com tamanho, capacidade ou algum detalhe diferente, e nestes casos não foram considerados duplicados. No modelo híbrido os percentuais de similaridade entre os itens aglutinados apresentaram valores menores e apontamentos onde os itens não são similares, como "ACIDO ACIDO CLORIDRICO PA ACIDO CLORIDRICO SOL.0,01N (M) - 1L" e "ACIDO ACIDO PERCLORICO P.A. 70%", com percentual de similaridade de 60.68%.

Obtivemos avaliação de 3 usuários. As respostas dos usuários ficaram gravadas em um arquivo CSV, onde a aplicação foi hospedada. Baseada nas respostas, a avaliação extrínseca da recuperação de informação do termo de busca foi baixa, pois recuperava muitos termos que não estavam correlacionados.

A avaliação da identificação da duplicidade de itens nas aglutinações foi feita por e-mail com o *feedback* dos usuários. A assertividade na recuperação dos itens com possibilidade de estarem repetidos indevidamente nas aglutinações foi satisfatória.

DISCUSSÃO

As atividades foram conduzidos para avaliar a similaridade semântica utilizando a similaridade de cosseno, em três modelos de *embeddings*, para avaliar a tarefa alvo entre o termo de busca e o conjunto de dados de treinamento do cadastro dos materiais, em Português Brasileiro, combinado com os modelos pré-treinado Word2Vec Skipgram e FastText Skipgram, ambos com 100 dimensões, disponibilizado pelo NILC para os

modelos densos estáticos e o paraphrase-multilingual-mpnet-base-v2 para os modelos densos dinâmicos.

Verificou-se uma mudança de paradigma na representação vetorial clássica, como *one-hot encoding*, o *bag of words*, até mesmo TF-IDF devido sua esparsidade e falta de escalabilidade para o aumento de documentos e vocabulários. O estado da arte atualmente é utilizar os textos não estruturados com redes neurais. Neste estudo verificamos a melhoria nos resultados da similaridade dos termos utilizando os *embeddings* de terceira geração, os *Transformers*, em relação aos de segunda geração, apresentando valores superiores e menores ocorrências de OOV.

O cálculo de similaridade nos *embeddings* estáticos foi feito termo a termo, e calculada a média quando o termo apresentava mais de uma palavra. Já nos modelos de *embeddings* dinâmicos, a similaridade é calculada diretamente pelo termo composto, uma vez que este modelo considera o contexto das palavras dentro da sentença.

O objetivo inicial foi listar materiais similares a um termo de busca, e verificou-se que para termos comuns como “geladeira” ou “gasolina comum” foram recuperados os itens alvo, ou similares da mesma área. Para termos técnicos ou específicos como “pisseta” , por exemplo, foram recuperados termos similares fora do contexto do material nos três modelos analisados. Esta diferença na precisão dos termos indicados está relacionada à frequência e ao contexto em que esses termos aparecem nos dados de treinamento. Os termos comuns são frequentemente encontrados e usados em diversos contextos, permitindo que o modelo aprenda com precisão suas relações semânticas com palavras similares. Por outro lado, termos técnicos aparecem com menos frequência e em contextos mais limitados, o que leva a associações menos precisas.

Dentre os modelos avaliados o modelo de representação sentencial baseada em arquitetura de redes siamesas (SBERT) apresentou melhor desempenho em relação aos outros modelos estudados nas avaliações de similaridade entre o termo de busca e o termo alvo, e na identificação de itens duplicados nas aglutinações. Ainda assim, não apresentou respostas satisfatórias às consultas, retornando os cinco termos mais similares ao termo de busca, assim como os outros modelos. Para melhoria desta tarefa alvo, propõe-se a combinação de técnicas de *fine-tunning* no modelo SBERT, utilizando outras fontes de dados além do cadastro de materiais, uma vez que este tipo de fonte não apresenta série temporal e nem crescimento exponencial, necessários para este tipo de processamento, o que exige mais recursos computacionais.

Outras alternativas que podem contribuir com a melhoria nos resultados da tarefa-alvo seria a utilização de dados anotados para os termos mais específicos e técnicos, aplicação de RAG (*Retrieval-Argumented Generation*), que combina a recuperação de informação de *corpus* treinados com modelos de geração de textos, como o GPT por exemplo, ou a implementação de retroalimentação desta base de conhecimento, considerando também a avaliação dos usuários, gerando assim novos conhecimentos e aprimorando continuamente a precisão e eficácia do artefato e evitando o *concept drift*, que é a mudança nos padrões subjacentes dos dados ao longo do tempo.

Os resultados obtidos indicam que os modelos de *embeddings* dinâmicos, como BERT e SBERT, são mais adequados para identificar duplicidades nas demandas de compras em comparação com os modelos estáticos. A capacidade dos *embeddings* dinâmicos de capturar contextos mais complexos e sutis nas descrições dos materiais é um fator decisivo para sua superioridade.

Esses achados estão alinhados com estudos recentes na área de PLN que destacam a eficácia dos modelos baseados em *Transformers* para tarefas de similaridade semântica. No entanto, a implementação prática desses modelos em um ambiente de compras públicas requer considerações adicionais, como custo computacional e facilidade de integração com sistemas existentes.

As implicações práticas deste estudo são significativas, pois a adoção de modelos de *embeddings* dinâmicos pode melhorar substancialmente a eficiência e a transparência dos processos licitatórios, reduzindo redundâncias e otimizando a gestão de contratos e aquisições. Além disso, a pesquisa destaca a importância de continuar investigando técnicas de PLN para aprimorar ainda mais a precisão e a aplicabilidade em contextos específicos.

CONCLUSÃO

O intuito deste trabalho foi analisar a similaridade de sentenças curtas, através do uso de *embeddings* densas estáticas e dinâmicas utilizando como fonte de conhecimento os cadastros de materiais utilizados em compras públicas de uma instituição de ensino superior pública com o intuito de diminuir a ocorrência de licitações de itens em duplicidade com nomenclaturas diferentes.

Para isso, foram treinados modelos de *embeddings* utilizando tecnologias de segunda e terceira geração de PLN, e analisando a recuperação da informação neste contexto específico. Foi feita a coleta de dados de sites de comércio eletrônico e do sistema de gestão de compras para aumentar a quantidade de sentenças na geração dos modelos estáticos, em conjunto com os modelos pré-treinados do NILC, e utilizado o modelo pré-treinado paraphrase-multilingual-mpnet-base-v2 do Hugging Face para o SBERT e o neuralmind/bert-base-portuguese-cased para o BERT.

Como a recuperação dos termos depende do contexto das palavras no vetor indexado, e os dados pesquisados são específicos ou técnicos, é incomum estes termos estarem relacionados em modelos pré-treinado, retornando desta forma termos similares bem diferentes dos termos alvo esperados. Baseado neste resultado, optou-se por atacar o problema na próxima etapa do processo de compras, que consiste na aglutinação dos itens pedidos pelos solicitantes. Nesta fase, os modelos se mostraram eficientes em elencar a duplicidade de materiais aglutinados, sendo possível os tomadores de decisão alterarem ou discutirem com os solicitantes outras medidas para contornar a questão de licitar itens iguais que poderiam receber valores diferentes.

Concluímos que a implementação de modelos usando técnicas de processamento de linguagem natural no contexto de compras públicas, especificamente em português não é uma tarefa trivial. A escassez de *corpus* específicos para domínios mais técnicos em português, e anotação de dados, bem como o poder computacional e quantidade de dados necessários para implementar embeddings pré-treinados dificultam o desenvolvimento de pesquisas específicas para esta área. Por outro lado, os *embeddings* de terceira geração têm se mostrado uma boa alternativa para soluções mais especialistas, mesmo que seu maior potencial seja nas tarefas alvo mais generalistas.

Como pesquisas sobre similaridade na língua portuguesa ainda são escassas, podem ser analisadas a combinação de outros métodos, conforme apresentado por Pinheiro et al. (2017). Várias são as abordagens possíveis, como a construção de uma

Base de Conhecimento Léxico (Lexical Knowledge Base- LKB) específica para o domínio de compras públicas, que pode melhorar os resultados sem a necessidade de treinamento adicional com dados anotados; Mapa de Sinônimos utilizando o *ElasticSearch*; o aperfeiçoamento do modelo apresentado neste estudo implementando algumas das técnicas descritas na discussão de resultados, como RAG, retroalimentação da base de conhecimento, ou *fine tunning,* utilizando como fonte de dados as aglutinações ou pedidos feitos ao longo do tempo que aumentaria a quantidade de dados com informações temporais, o que não foi feito devido a restrição de tempo e recursos computacionais demandados.

Sugere-se a continuidade da pesquisa com foco na adaptação de modelos de embeddings para domínios específicos e na exploração de novas técnicas de PLN que possam aumentar ainda mais a precisão e a eficácia da identificação de duplicidades.

AGRADECIMENTOS

Este trabalho foi apoiado pelo Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC/UFSC) e pela UDESC (Universidade do Estado de Santa Catarina), objeto de estudo desta pesquisa, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - código de financiamento 001.

REFERÊNCIAS

ALVAREZ-RODRÍGUEZ, José María; LABRA-GAYO, José Emílio; RODRÍGUEZ-GONZÁLEZ, Alejandro; DE PABLOS, Patrícia Ordoñez. **Empowering the access to public procurement opportunities by means of linking controlled vocabularies. A case study of Product Scheme Classifications in the European e-Procurement sector**. Computers in Human Behavior, v. 30, p. 674–688, jan. 2014.

ARTAMONOV, Alexey; VASILEV, Michael; TUKUMBETOVA, Rufina; ULIZKO, Michael. **Multiagent System for Monitoring, Analysis and Classification of Data from Procurement Services**. Procedia Computer Science, v. 213, p. 96–100, 1 jan. 2022.

BIFULCO, Ida; CIRILLO, Stefano; ESPOSITO, Christian; GUADAGNI, Roberta; POLESE, Giuseppe. **An intelligent system for focused crawling from Big Data sources**. Expert Systems with Applications, v. 184, p. 115560, dez. 2021

BRASIL. **Lei nº 13.460, de 26 de junho de 2017a**. Dispõe sobre participação, proteção e defesa dos direitos do usuário dos serviços públicos da administração pública. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/lei/l13460.htm. Acesso em: 26 dez. 2023.

BRASIL. **Lei nº 14.133, de 1º de abril de 2021b**. Estabelece normas gerais de licitação e contratação para as Administrações Públicas diretas, autárquicas e fundacionais da União, dos Estados, do Distrito Federal e dos Municípios. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm. Acesso em: 20 jun. 2024.

BRASIL. **Decreto n.º 10.947, de 25 de janeiro de 2022c**. Regulamenta o art. 11 da Lei n.º 12.527, de 18 de novembro de 2011, e altera o Decreto n.º 8.777, de 11 de maio de 2016, que institui a Política de Dados Abertos do Poder Executivo federal. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2022/decreto/D10947.htm

BRUCKNER, Tomáš; VENCOVSKÝ, Filip. **The Process of Unit Price Extraction from Public Sector Contracts**. Acta Informatica Pragensia, v. 9, n. 2, p. 170–183, 31 dez. 2020.

CORTES, Eduardo G; VIEIRA, Renata; BARONE, Dante A. C. Capítulo 16 Perguntas e Respostas. In CASELI, Helena de Medeiros.; NUNES, Maria das Graças Volpe (org.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2 ed. BPLN, 2024. https://brasileiraspln.com/livro-pln/2a-edicao/

CRESWELL, John W. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto**. 3. ed. Porto Alegre, RS: Artmed, 2010.

FANTONI, Gualtiero et al. **Text mining tool for translating terms of contract into technical specifications: Development and application in the railway sector**. Computers in Industry, v. 124, p. 103357, jan. 2021.

FEITOSA, David B.; PINHEIRO, Vládia C. **Análise de Medidas de Similaridade Semântica na Tarefa de Reconhecimento de Implicação Textual**. In: SIMPÓSIO

BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 1. , 2017, Uberlândia/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2017. p. 161-170.

FREITAS, Cláudia., Capítulo 14 Dataset e corpus. In CASELI, Helena de Medeiros.; NUNES, Maria das Graças Volpe (org.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln. Acesso em: 12 dez. 2023

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 7 ed. Rio de Janeiro: Atlas, 2019.

JACQUES DE SOUSA, Luís; POÇAS MARTINS, João; SANHUDO, Luís. **Portuguese public procurement data for construction (2015–2022)**. Data in Brief, v. 48, p. 109063, jun. 2023.

JURAFSKY, Daniel; MARTIN, James H. **Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition**. Draft of January 7, 2023. Nova Jersey: Prentice Hall, 2023. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 10 dez. 2023.

KILIMCI, Zeynep H.; AKYOKUS, Selim. **Deep learning- and word embedding-based heterogeneous classifier ensembles for text classification**. Complexity, 2018. Disponível em: https://doi.org/10.1155/2018/7130146. Acesso em: 10 dez. 2023.

LEE, Gyueun et al. **Automatic Classification of Construction Work Codes in Bill of Quantities of National Roadway Based on Text Analysis**. Journal of Construction Engineering and Management, v. 149, n. 2, fev. 2023

LIMA, Marcos et al. **Inferring about fraudulent collusion risk on Brazilian public Works contracts in oficial texts using a Bi-LSTM approach**. Findings of the Association for Computational Linguistics: EMNLP 2020, November 2020. p. 1580 – 1588, 2020.

LIMA, Weslley et al. **Methodology for automatic extraction of red flags in public procurement**. International Joint Conference on Neural Networks (IJCNN). 2023.

MODRUŠAN, Nicola; RABUZIN, Kornelije; MRŠIĆ, Leo. **Improving Public Sector Efficiency using Advanced Text Mining in the Procurement Process**. Proceedings of the 9th International Conference on Data Science, Technology and Applications, 2020.

OLIVEIRA, Hugo Gonçalo; ALVES, Ana Oliveira; RODRIGUES, Ricardo. Gradually **Improving the Computation of Semantic Textual Similarity in Portuguese**. EPIA, 2017. LNAI 10423, pp. 841–854.

PAES, Aline; VIANNA, Daniela; RODRIGUES, Jessica. Capítulo 15 Modelos de Linguagem. In CASELI, Helena de Medeiros.; NUNES, Maria das Graças Volpe (org.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln. Acesso em: 12 dez. 2023

PINHEIRO, Álvaro Farias. **Automação de Processos através da RPA para Transformação Digital**. Fundação Escola Nacional de Administração Pública, Diretoria de Desenvolvimento Profissional. SAIS - Área 2-A - 70610-900 — Brasília DF. 2022.

PINHEIRO, Anderson et al. **Statistical and Semantic Features to Measure Sentence Similarity in Portuguese**. 2017. Uberlândia, Brasil. p. 342–347.

RODRIGUES, Ana Carolina. **Avaliação de representações embeddings para similaridade sentencial no Português**. 2023. Dissertação (Mestrado) – Universidade de São Paulo, São Carlos, 2023. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-02062023-105741/. Acesso em: 23 jan. 2024

SANTOS, Joaquim et al. **Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition**. 2019.

SENO, Eloize Rossi Marques; DE PAIVA, Valéria; PINHEIRO, Vládia. Capítulo 9 Semântica com Técnicas Simbólicas. In CASELI, Helena de Medeiros.; NUNES, Maria das Graças Volpe (org.)a. **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln. Acesso em: 12 dez. 2023.

SENO, Eloize Rossi Marques; CLARO, Daniela; MOTA, Laila; RODRIGUES, Jéssica. Capítulo 10 Semântica Distribucional. In CASELI, Helena de Medeiros.; NUNES, Maria

das Graças Volpe (org.)b. **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln. Acesso em: 12 dez. 2023.

SICILIANI, Lucia et al. **AI-based decision support system for public procurement**. Information Systems, v. 119, p. 102284, 1 out. 2023.

SILVA, Mariana O. et al. **Análise de Sobrepreço em Itens de Licitações Públicas**. In: WORKSHOP DE COMPUTAÇÃO APLICADA EM GOVERNO ELETRÔNICO (WCGE), 11. , 2023, João Pessoa/PB. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023 . p. 118-129. ISSN 2763-8723. DOI: https://doi.org/10.5753/wcge.2023.230608.

TORRES-BERRU, Yeferson; F. LOPEZ-BATISTA, Vivian; CONDE ZHINGRE, Lorena. **A Data Mining Approach to Detecting Bias and Favoritism in Public Procurement**. Intelligent Automation & Soft Computing, v. 36, n. 3, p. 3501–3516, 2023.

VERGARA, Sylvia Constant. **Métodos de pesquisa em administração**. São Paulo: Atlas, 2012.