Machine Learning Aplicado em Dados Abertos Governamentais para Detecção de Impropriedades na Aplicação de Recursos Públicos

Ramon Dantas Vaqueiro[1], Tatiana Escovedo[2]

ABSTRACT

The high purchasing power of the State and the need to ensure the correct application of public resources raises great concern among citizens. In light of this reality, the legislator granted a differentiated treatment to public purchases and submitted them to greater levels of transparency, determining that their data would be made available openly. In view of popular demand and the availability of access to this resource, its use by Public Bodies that are empowered is imperative. However, a significant part of this data is textual, which requires diligent and specific pre-processing by those who use it. This article focuses on a database corresponding to materials purchased by the Federal Government throughout the year 2021. The objective was to group the textual descriptions of similar purchases, allowing their comparison. Among the possible applications from the correct grouping, there is the identification of outliers in the prices of acquisitions, signaling an indication of possible impropriety. As a result, a clustering considered satisfactory was obtained in 72% of the cases.

KEY WORDS

Public Purchases; Text Mining; Machine Learning.

RESUMO

O elevado poder de compra do Estado e a necessidade de se zelar pela correta aplicação dos recursos públicos suscita no cidadão grande preocupação. À luz dessa realidade, o legislador conferiu um tratamento diferenciado para as compras públicas e as submeteu a maiores níveis de transparência, determinando que seus dados passassem a ser disponibilizados abertamente. Em face da demanda popular e da disponibilidade de acesso a esse recurso, torna-se imperiosa a sua utilização pelos órgãos competentes. Todavia, parte significativa desses dados são textuais, ou seja, não estruturados, o que requer um diligente e específico pré-processamento por parte de quem os utilize. Este artigo se debruça sobre uma base de dados correspondente aos materiais comprados pelo Governo Federal na modalidade pregão ao longo do ano de 2021. Objetivou-se agrupar as descrições textuais de compras semelhantes, permitindo sua comparação. Dentre as aplicações possíveis a partir do correto agrupamento, encontra-se a identificação de *outliers* nos preços das aquisições, sinalizando um indício de eventual impropriedade. Como resultado, obteve-se uma clusterização considerada satisfatória em 72% dos casos.

PALAVRAS CHAVE

Compras Públicas; Mineração de Texto; Aprendizado de Máquina.

1 INTRODUÇÃO

Muito se discute sobre a importância da transparência no setor público e como promover um melhor controle social de seus gastos. Certamente, a disseminação do acesso à internet, o avanço do governo digital e a criação de legislações acerca do tema da transparência pública, como a Lei de Acesso à Informação [BRASIL, 2011], contribuíram para que qualquer cidadão disponha de uma enorme variedade de dados governamentais. Ademais, em se tratando de um país de dimensões continentais como o Brasil, é de se imaginar que a administração pública possua um grande volume de dados que necessitem ser publicizados em portais da transparência, sejam nas esferas federal, estadual, municipal ou distrital.

A despeito dos relevantes aspectos citados, faz-se mister salientar que a disponibilização de dados públicos em si não basta. Faz-se necessário que tais recursos possam ser efetivamente analisados pelo contribuinte, para que deles possa extrair conhecimento e uma melhor compreensão da destinação dos recursos que lhe foram tributados. Nesse sentido, observa-se uma grande dificuldade nos numerosos casos em que parcela dos dados envolvidos não são estruturados, como nas descrições das compras públicas. No caso em tela, a especificação dos itens adquiridos pelo governo é feita em formato de texto livre, o que, alinhado com o volume massivo de compras, torna impraticável que uma análise mais complexa seja feita por um cidadão desprovido de uma ferramenta computacional que o auxilie.

[1,2] Departamento de Informática

Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Brazil

e-mail: [1] ramonvaqueiro19@gmail.com, [2] tatiana@inf.puc-rio.br

Nesse diapasão, este estudo se propõe a utilizar dados públicos, disponíveis em plataformas governamentais na internet, lançando mão de algoritmos de *Machine Learning* para realizar agrupamentos de itens adquiridos pela administração pública. Ao se rotular a compra realizada e agrupá-la com suas semelhantes, facilita-se uma série de análises, haja vista que o dado passa a ser tratado como estruturado. Facilita-se, dessa maneira, a detecção de impropriedades na aplicação de recursos públicos, a comparação da eficiência nas aquisições de órgãos distintos, o entendimento global de quais compras estão sendo realizadas, dentre outras aplicações.

Com a finalidade de lograr êxito no que se propõe no parágrafo anterior, este trabalho está organizado da seguinte forma: na Seção 2 discorreremos brevemente sobre a fundamentação teórica deste artigo, abordando, dentre outros assuntos, aspectos basilares da teoria de *Machine Learning* e apresentaremos alguns trabalhos relacionados a sua aplicação em dados governamentais; na Seção 3 apresentaremos a solução proposta e todos os passos que serão utilizados nos experimentos, desde a coleta e tratamento dos dados, aos algoritmos a serem utilizados nas etapas de mineração e clusterização; na Seção 4 apresentaremos os resultados dos experimentos, avaliando a qualidade dos agrupamentos e exploraremos uma das possibilidades de aplicação; e ,por fim, na Seção 5 apresentaremos as conclusões e propostas de pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA E REVISÃO DA LITERATURA

2.1. Compras Governamentais

A Constituição Federal [BRASIL, 1988], em seu inciso XXI do artigo 37, obrigou que, em regra, as compras públicas fossem precedidas de processo licitatório, ou seja, um procedimento mais complexo que a mera aquisição direta no mercado e que visa a assegurar a contratação da proposta mais vantajosa para a Administração Pública. Tal obrigatoriedade foi regulamentada por uma ampla legislação, sendo as normas mais céleres e abrangentes: lei nº 8.666/1993 [BRASIL, 1993], Lei Geral de Licitações; lei nº 10.520/2002 [BRASIL, 2002], que institui a modalidade pregão, para a aquisição de bens e serviços comuns; e a Nova Lei de Licitações, lei nº 14.133/2021 [BRASIL, 2021]. Ademais, toda compra pública deve possuir amparo em crédito orçamentário para sua realização, ou seja, deverá possuir autorização legislativa por meio da Lei Orçamentária Anual. Dessa forma, via de regra, após a realização da licitação e do posterior contrato, a Administração Pública ainda deve emitir uma Nota de Empenho em favor do credor, por meio da qual se compromete a realizar o pagamento desde que as condições previstas no empenho sejam cumpridas, sendo vedada a realização de despesa sem prévio empenho [BRASIL, 1964].

De maneira a dar maior publicidade a seus dados, dentre eles as compras citadas, foi editada pelo Governo Federal a lei complementar nº 131/2009 [BRASIL, 2009], conhecida como Lei da Transparência, que determinou a disponibilização de dados da execução orçamentária e financeira dos Entes Federados. Nesse sentido, são mantidos sites e APIs pelo Poder Público, sendo o Portal da Transparência[4], o Portal de Compras[5] e a API de Compras Governamentais[6] os mais conhecidos e utilizados. Nas plataformas citadas, pode-se obter dados afetos às descrições dos itens adquiridos pela União, tanto aquelas constantes das licitações, quanto das notas de empenho, suas quantidades, unidades de fornecimento, valores envolvidos, dentre outras informações. Dessa forma, além de fomentar o controle social originalmente previsto, provê-se aos Órgãos de Controle instituídos pela Carta Magna [BRASIL, 1988] subsídios para que possam desempenhar suas atividades.

2.2. Mineração de Textos

Mineração de texto, também conhecida como Processamento de Linguagem Natural (*Natural Language Processing*, ou NLP), pode ser entendida como um processo de extração de informações e conhecimentos úteis de coleções de documentos. Os fluxos de tal processo podem se assemelhar bastante com mineração de dados, porém, ao se trabalhar com informações textuais, uma série de desafios e especificidades ocorrem [FELDMAN *et al*, 2007]. Ao realizar mineração de texto, estamos manipulando uma grande quantidade de dados não estruturados ou semiestruturados, sendo necessário realizar um extenso e cuidadoso processo de pré-processamento para a extração de características e, a partir disso, utilizá-las em outros algoritmos para a descoberta de padrões e conhecimentos. Além disso, ao se manipular tais dados normalmente é necessário lidar com uma alta dimensionalidade de atributos e dados muito esparsos. Esses aspectos serão detalhados nas subseções seguir.

2.2.1 Pré-processamento

Devido à falta de informações estruturadas em dados textuais, o pré- processamento é uma etapa fundamental, pois é nela que extrairemos características do texto e geraremos uma informação estruturada. Por isso, essa é a normalmente a etapa em que se concentra a maior parte do tempo, para se realizar a limpeza necessária nos dados e a extração de informações que serão úteis nas próximas etapas.

Dependendo do tipo de aplicação, as etapas podem sofrer algumas alterações, não sendo necessário seguirem uma ordem fixa e podem ocorrer mais de uma vez durante o processo. De modo geral são aplicados os seguintes passos [SARKAR, 2016]:

- **Padronização do texto:** nesta primeira etapa é realizada uma primeira limpeza do texto, removendo caracteres especiais, pontuações (dependendo da aplicação), removendo acentuação e transformando todos os caracteres para minúsculo;
- **Tokenização (atomização):** neste momento é realizada a quebra do texto em blocos básicos de informação, a depender da aplicação essa divisão pode ser feita em parágrafos, frases ou palavras. O caso mais comum é cada token representar uma palavra;

[4] Disponível no endereço: https://www.portaltransparencia.gov.br/

5 Disponível no endereço: https://www.gov.br/compras/pt-br

6 Disponível no endereço: http://compras.dados.gov.br/docs/home.html

- **Remoção de *stop words*:** este processo consiste na remoção de palavras que possuem pouca ou nenhuma informação ou relevância no contexto. Os casos básicos de *stop words* são pronomes, conjunções preposições e artigos, porém é necessário avaliar também palavras que no domínio da aplicação não representam informação relevante e removê- las também;
- **Lematização (*lemmatization*):** esta etapa visa reduzir uma palavra a sua base (lema) de forma a agrupar as diversas variações de uma palavra. A lematização normalmente utiliza informações do contexto para resolver problemas de ambiguidade. Um ponto importante é que este processo sempre retorna uma palavra existente no dicionário. Por exemplo: livro, livrinho e livreto seriam transformados em livro;
- **Radicalização (*Stemming*):** este processo é semelhante a lematização, buscando agrupar as variações de uma palavra em seu radical. Porém normalmente o resultado desse processo não retorna uma palavra. Por exemplo: livro, livrinho e livreto seriam transformados em "livr".

Após a realização destas etapas, é realizada a vetorização do documento, o que consiste em transformar os dados tratados em um vetor de característica com informações numéricas que poderá ser utilizado posteriormente em algoritmos de aprendizado de máquina [WEISS e INDURKHYA, 2015]. A estratégia mais frequente é a chamada *bag of words*: nela, é criado um dicionário com todos os *tokens* existentes, no qual cada um representará uma característica a que será atribuída um valor que pode representar tanto a quantidade de ocorrências do *token* como uma medida de importância da palavra, como por exemplo o chamado TF-IDF [BENGFORT *et al*, 2018].

2.2.2 TF-IDF

O TF-IDF é uma medida estatística que busca representar a importância de uma palavra em um documento, mas levando em consideração também a sua relevância no conjunto de documentos analisados. [BENGFORT *et al*, 2018] Sua sigla é formada pela junção das duas médias utilizadas para seu cálculo:

- **Frequência do termo** (*term frequency* – TF): contabiliza a quantidade de vezes que um determinado termo ocorre em um documento. Portanto quanto maior for a frequência no documento, maior será a importância do termo;
- **Frequência inversa do documento** (*inverse document frequency* – IDF): contabiliza o inverso da ocorrência do termo no conjunto dos documentos. O objetivo desta medida é reduzir a importância de termos muito frequentes (que não são bons candidatos para diferenciar os elementos) e priorizar termos mais raros (que ajudam a distinguir melhor os documentos). Assim, quanto maior for a frequência do termo nos documentos, menor será a importância.

Portanto a combinação dessas medidas permite compensar o peso que determinado termo terá considerando a frequência que ele costuma aparecer em um conjunto de documentos. Levando a um caso extremo, caso um termo ocorra em todos os documentos de um conjunto o cálculo do IDF será 0 o que resultará um TF-IDF também 0 não importando o valor de TF, mostrando que um termo que ocorre em todos os documentos não possui influência na discriminação entre documentos [BENGFORT *et al*, 2018].

2.2.3 Redução da dimensionalidade

Após a realização do pré-processamento dos dados textuais e sua vetorização, deparamo-nos com mais um desafio inerente a este tipo de dado: a alta dimensionalidade dos atributos e dados muito esparsos. A depender do domínio de aplicação da mineração de textos, o número de atributos (palavras encontradas no conjunto de dados) alcança o patamar de 25.000. Ao mesmo tempo, uma pequena porcentagem dessas características aparece em um determinado documento, em muitos casos a matriz de atributos dos domínios pode ter menos de 1% dos seus valores não nulos [FELDMAN *et al*, 2007]. Essas duas questões influenciam muito o desempenho e a qualidade dos resultados de diversos algoritmos de M*achine Learning*, portanto a redução da dimensionalidade do vetor característica pode melhorar a extração de conhecimento.

Uma das formas de se realizar essa redução da quantidade que características é a utilização do algoritmo denominado decomposição em valores singulares (*Singular Value Decomposition* - SVD). De maneira sintética, o SVD realizará um agrupamento das características que normalmente ocorrem juntas e formará a partir delas um novo conjunto de atributos (que pode ser entendido como uma combinação dos atributos originais).

Em mineração de textos, a utilização do SVD é estudada na área de modelagem de tópicos e sua aplicação é denominada como análise semântica latente (*Latent Semantic Analysis* – LSA). Neste contexto, podemos interpretar os novos atributos gerados pela aplicação do SVD como a criação de tópicos, ou seja, o agrupamento de termos que normalmente ocorrem em conjunto representando uma mesma ideia ou tema [BENGFORT *et al*, 2018].

Com a extração das características mais relevantes dos textos, o próximo passo na mineração de texto é a utilização dessas informações para a extração de algum tipo de conhecimento dos dados. Dependendo do problema a ver avaliado podemos realizar tarefas como classificação, sumarização, agrupamento, dentre outras atividades.

2.3. Clusterização

Métodos de clusterização consistem em algoritmos de *Machine Learning* utilizados em tarefas de aprendizado não supervisionado como forma de identificação de padrões e agrupamento de dados. De forma geral, estes algoritmos podem ser vistos como problemas de otimização, cujo objetivo é maximizar a similaridade *intracluster* e minimizar a similaridade *interclusters*. Para isso, precisam utilizar características numéricas para que seja possível calcular a distância entre dois elementos e definir, assim, se são similares ou não [ESCOVEDO e KOSHIYAMA, 2020]. Dentre os diversos métodos diferentes de agrupamento, podemos citar o *K-means* e o *DBSCAN*.

2.3.1 K-means

O algoritmo de clusterização mais conhecido é o *K-means* e seu funcionamento consiste na divisão do espaço de características em um número definido de agrupamentos [IGUAL e SEGUÍ, 2017]. Dados os parâmetros de número de *clusters* e a forma de distância a ser utiliza, o processo de agrupamento ocorre de forma iterativa: inicialmente, os centros dos *clusters* são definidos aleatoriamente; a seguir, os elementos são associados ao centro mais próximo e é calculado um novo centro dos agrupamentos de acordo com os elementos a ele associados; e o processo se repete até que haja uma convergência (os centros parem de se mover) ou se atinja um determinado número de iterações. Isto posto, é importante ressaltar que este algoritmo é sensível a inicialização dos centros e que cada elemento sempre será associado a um cluster.

2.3.2 DBSCAN

O algoritmo *DBSCAN*, abreviação de Clusterização Espacial Baseada em Densidade de Aplicações com Ruído (*Density Based Spatial Clustering of Application with Noise*) [ESTER *et al*, 1996], realiza agrupamentos baseado na densidade da distribuição dos elementos no espaço de característica. Sua principal vantagem é não ser necessário definir previamente a quantidade de agrupamentos, porém é necessário informar um limiar de densidade, que será utilizado para identificar se um ponto pertence ou não a um *cluster*. Uma característica importante do *DBSCAN* é que, ao final do processo, pode ocorrer de nem todos os elementos possuírem um *cluster* associado.

2.3.3 Distância de cossenos

Dado que a forma de se calcular a distância entre dois elementos é um parâmetro importante em algoritmos de clusterização, a literatura especializada em mineração de texto ressalta que uma das medidas que melhor performa para dados textuais é a distância de cossenos [WEISS e INDURKHYA, 2015] e [SARKAR, 2016]. Tal medida é calculada a partir do ângulo formado entre os vetores características de dois elementos, seu valor varia de 0 a 1 e quanto mais próximo de 1 mais similares são os itens.

2.3.4 Formas de avaliação

Um dos desafios para a avaliação da aplicação de algoritmos não supervisionado é a frequente ausência de categorias anotadas. Usualmente a validação de agrupados gerados por algoritmos de clusterização é feita manualmente por um especialista na área de negócio, que irá avaliar se a divisão proposta faz sentido, ou apresenta algum ganho de informação.

Porém além da análise manual existe uma medida numérica que pode auxiliar na avaliação da qualidade dos agrupamentos. O coeficiente de *Silhouette* é calculado para cada elemento e reflete se um elemento está mais parecido com o cluster associado ou com outro cluster a seu redor [BENGFORT *et al*, 2018]. A medida varia de -1 a 1, de forma que quanto mais próximo de 1 mais semelhante ao centroide do próprio cluster, valores próximos a 0 representam elementos muito próximo da fronteira de decisão entre *clusters* e valores negativos indicam elementos associados erroneamente. Com o coeficiente de cada elemento podemos avaliar também a média por agrupamento como forma de visualizar se estes estão com elementos muito dispersos.

2.4. Trabalhos Relacionados

Tendo em vista o grande volume e variedade de dados gerados pelo setor público, diversos autores destacam a importância de se realizar processos de descoberta de conhecimento em base de dados como forma de se viabilizar análises efetivas.

Neste contexto, Balaniuk (2010) enfatiza a importância da utilização de ferramentas de mineração de dados como forma de agilizar e priorizar o trabalho de auditores. Além disso, deixa claro que não se deve criar a expectativa de que o processo de mineração irá conseguir encontrar todos os casos desejados e que inúmeros desafios ocorrerão durante o processo, como a má qualidade dos dados, dificuldade de acesso a dados e baixa integração de sistemas, mas isso não deve ser utilizado com empecilho para o investimento neste tipo de projeto. Diante do volume e complexidade dos dados,
o autor ressalta a importância de especialistas na área para propor tipologias, ou seja, indícios que podem caracterizar ilícitos, e a partir delas realizar a exploração dos dados. O artigo propõe uma adaptação do CRISP DM [SHEARER, 2000] ao contexto de trabalho do Tribunal de Contas da União (TCU), enfatizando como o processo auxilia no cálculo de métricas de risco e relevância para priorizar auditorias.

Além do volume de dados a ser analisado, uma grande parte deles se encontra em campos de informação textual que normalmente é inserida em formato livre, carecendo assim de qualquer estrutura que possa facilitar a extração de informações. Nesta conjuntura, diversas pesquisas têm abordado técnicas de mineração de textos para suprir esta demanda. Pode-se ressaltar no âmbito de auditoria de compras públicas [CARVALHO *et al*, 2014], [CARVALHO, 2015], [PAIVA, 2017], [ALMEIDA *et al*, 2018], [AMARAL e RODRIGUES, 2020] e [FONTES, 2022], propondo a
análise de descrições de produtos em diferentes instrumentos como notas ficais ou notas de empenho.

Carvalho (2015) cita que uma das grandes responsabilidades da Controladoria-Geral da União (CGU) é identificar as compras do governo com valores diferentes dos praticados pelo mercado e que essa seria uma tarefa muito dificultosa, haja vista que os dados a ela afetos não seriam suficientemente estruturados, assim como seu volume e diversidade. Explica que, no âmbito da CGU, foi desenvolvida uma solução que consistiu na criação e manutenção de um banco de dados com preços de referência por produto. Sua identificação e agrupamento seria possível a partir do campo descrição do item de empenho, que apresenta informação de maneira desestruturada, mas que possui um certo padrão em seu preenchimento, selecionando-se os que possuem apenas código de material e descartando-se os empenhos de serviço. Nesse estágio são filtrados apenas os itens de interesse previamente estabelecidos pelos especialistas, por meio dos códigos de material pré-definidos. Em seguida, combina-se o código citado com um conjunto de palavras-chave extraídas do mesmo campo para caracterizar um produto em questão e, por fim, calcula-se o preço de referência. Por meio da metodologia citada, a CGU teria analisado 51 produtos diferentes. Adicionalmente ressalta-se um trabalho anterior do mesmo autor [CARVALHO *et al*, 2014] em que foi proposta uma metodologia semelhante, porém em que era utilizado o preço dos itens para se realizar sua clusterização.

O trabalho de Paiva (2017) tem como objetivo identificar de forma automática produtos nas descrições textuais de notas de empenho. O autor utiliza uma vertente denominada *bag of phrases* e seleciona para cada produto o conjunto de palavras sequenciais que melhor o caracteriza. A metodologia é definida de maneira que ocorra a menor interação humana possível, podendo ser incluída uma etapa opcional com um especialista para refinamento das regras geradas. O artigo foca em analisar o conjunto dos produtos mais comprados, de forma a possuir amostras suficientes para testar a metodologia. É proposto um algoritmo próprio para gerar frases candidatas e filtrar a melhor alternativa. De forma a avaliar a qualidade dos resultados, utilizou o algoritmo de clusterização DBSCAN em conjunto com outras informações da compra para verificar se as regras geraram agrupamentos homogêneos. Além disso, foi utilizado um algoritmo de detecção de *outlier* para a identificação de possíveis produtos classificados de forma errônea. Para finalizar, o artigo apresentou um conjunto de aplicações que poderia utilizar a identificação de produtos proposta no trabalho. Realizou os experimentos para 25 produtos e em 16 deles obteve bons resultados.

Já Almeida *et al.* (20118) buscam padronizar as unidades de medida de um mesmo produto de forma a facilitar sua comparação. Em todo o desenvolvimento do trabalho, foi utilizada a aplicação KNIME *Analytics*. Após a realização da mineração do texto, foi realizado o agrupamento com base na frequência de ocorrência. Foram utilizados cerca de setenta mil registros de itens e avaliados 24 produtos nos experimentos. Os autores relataram uma redução de 78% das unidades de medidas nos produtos analisados, conseguindo agrupar a maior parte dos produtos em unidades de medidas padronizadas.

Amaral e Rodrigues (2020) enfatizam que grande parte dos dados disponíveis para auditoria se encontra em formato textual, o que demandaria técnicas de mineração de texto por parte dos órgãos de controle. Nesse sentido, desenvolvem um estudo de caso, por meio da aplicação de algoritmo de modelagem de tópicos, que utilizou a técnica de mineração de tópicos latentes (*Latent Dirichlet Allocation* – LDA) em 65 mil registros de itens de compras públicas entre os anos de 2008 e 2017. Dessa forma, disponibilizam informações úteis às ações de controle, identificando as características semelhantes nas descrições daqueles registros, clusterizando-os. Todavia, a pequena extensão nessas descrições prejudicou o desempenho do algoritmo, tendo sido elencada como um desafio a ser superado em trabalhos futuros. Dentre as métricas de avaliação, utilizou-se também da análise humana.

Fontes (2022) propôs a utilização de mineração de texto em notas fiscais para a criação de um classificador voltado para identificar medicamentos e combustíveis, de forma a auxiliar auditores fazendários na identificação de indícios de evasão fiscal e ou de irregularidade nas compras. O trabalho se divide em dois estudos, um para medicamentos e outro para combustíveis, mas em ambos os casos foi necessário a criação de uma base anotada para realizar o treinamento do modelo. No experimento referente aos medicamentos foi criado um classificador hierárquico de forma a extrair algumas características específicas dos itens em cada etapa, por exemplo, princípio ativo, concentração e formato. Para este estudo foi relatado uma acurácia de 99%. No experimento referente a combustíveis, foram utilizados classificadores Bayesianos para identificar se as descrições eram referentes a combustíveis ou produtos afins e depois os identificar, sendo relatada uma acurácia de 100%.

3 SOLUÇÃO PROPOSTA

De modo a realizar a descoberta de conhecimento na base de dados deste trabalho, lançou-se mão da metodologia do Processo Padrão Interindústrias para Mineração de Dados (CRISP-DM) [SHEARER, 2000], haja vista sua popularidade e completude. Dessa forma, a solução proposta será abordada de acordo com as seis etapas cíclicas conforme ilustrado na Figura 1.

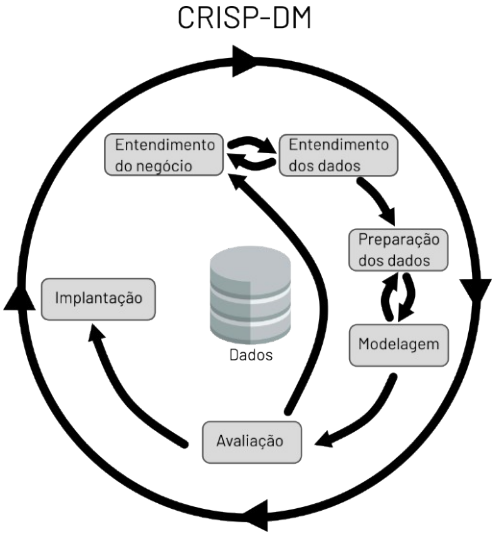

Figura 1. Ciclo do CRISP-DM. Fonte: baseada em [SHEARER, 2000].

3.1. Entendimento do Negócio

Como elucidado na introdução e fundamentação teórica deste trabalho, a constituição e o arcabouço legislativo brasileiro definem como são realizadas as aquisições públicas e como seus dados devem ser publicizados de forma a permitir um controle efetivo sobre os gastos governamentais. Além disso, asseveram que o processo licitatório deve preceder as compras estatais como regra e que por meio desse deve ser garantida a contratação da proposta mais vantajosa. Todavia, o entendimento de que a despesa realizada realmente se deu em proveito da melhor das propostas nem sempre se obtém facilmente por meio da mera análise direta do conteúdo dos atos publicados nos portais competentes, haja vista que os dados são oriundos de sistemas voltados à gestão. Dessa forma, nem sempre foram pensados com o intuito original de serem explorados pelo controle social ou aquele realizado pelos órgãos especializados, possuindo dados não estruturados, como descrições textuais livres por exemplo. Por conseguinte, dificulta-se sobremaneira a comparação entre as diferentes licitações e seus itens. Portanto, faz- se necessário lançar mão de técnicas de mineração de texto para se extrair melhores informações desse tipo de dado e a realização posterior de agrupamentos tornando possível a comparação entre as diferentes aquisições.

Em face do exposto, serão utilizadas páginas institucionais que disponibilizam de forma aberta dados detalhados de compras públicas com o intuito de extrair informações afetas ao processo

licitatório, assim como aos itens adquiridos por meio dele, a fim de realizar a descoberta de conhecimento na base citada.

3.2. Entendimento dos Dados

Neste trabalho, foi utilizada a API de Compras Governamentais (disponível em http://compras.dados.gov.br) para se ter acesso aos dados do Sistema Integrado de Administração e Serviços Gerais (SIASG), por meio do qual são operacionalizadas compras do Governo Federal. As pesquisas são feitas por meio dos parâmetros utilizados nas URLs e retornam arquivos XML, JSON ou CSV, também sendo possível uma visualização em HTML. Em face da grande quantidade de consultas demandas e do consequente volume de dados relacionado, desenvolveu-se código próprio para realizar a tarefa, utilizando-se de: linguagem de programação Python, ambiente de desenvolvimento Jupyter Lab e bibliotecas, como Pandas (para manipulação de dados) e Joblib (para o processamento paralelo das requisições).

Por meio das ferramentas supracitadas, foram extraídos 395.073 registros que correspondem aos itens de licitações realizadas pelo Governo Federal ao longo do ano de 2021. A compilação das extrações resulta em uma *Dataframe* constituído de 11 atributos (detalhados na Tabela 1), sendo julgados os mais relevantes, após diligente análise exploratória e à luz dos conhecimentos obtidos na fase de entendimento do negócio, os que se seguem: o número da licitação, a descrição do item, código do material (CATMAT), quantidade licitada, unidade de fornecimento e valor licitado. Foi escolhido o campo descrição detalhada do item para empregar a Mineração de Textos.

Atributo	Descrição
codigo_item_material	Código do material (CATMAT)
criterio_julgamento	Critério usado para a escolha do vencedor da licitação
data_publicacao	Data de publicação da licitação
descricao_detalhada_item	Descrição detalhada do item licitado
fornecedor_vencedor	Fornecedor vencedor da licitação (CNPJ)
numero_licitacao	Identificador da licitação
numero_item_licitacao	Número identificador do item na licitação
valor	Valor vencedor
quantidade	Quantidade licitada do item
uasg	Unidade Administrativa de Serviços Gerais, órgão que realizou a licitação
unidade_fornecimento	Unidade de fornecimento do item

Tabela 1. Atributos do *Dataframe* utilizado.

3.3. Preparação dos Dados

Nesta etapa, os dados passaram por um processo de preparação e limpeza, por meio de tratamentos com a identificação de valores faltantes, realizando-se novas extrações da API para substituí-los ou removendo registros ou até mesmo atributos. No que concerne ao atributo de descrição textual do item, pode-se citar: retirada de pontuações e de caracteres especiais; opção pelo padrão minúsculo; realização da tokenização; importação tanto de *stop words* comuns da língua portuguesa, quanto aquelas julgadas pertinentes em face do domínio do problema; lematização e radicalização; criação de um vetor característica a partir da descrição textual, por meio do cálculo da medida estatística TF-IDF; e redução da dimensionalidade do vetor citado, por meio da LSA. Utilizou-se nesta etapa das seguintes bibliotecas adicionais: *NLTK* e *Spacy* para o processamento do texto; e S*klearn* para a geração do vetor característica e redução de dimensionalidade.

3.4. Modelagem dos Dados

Por ocasião do registro de cada item da compra pública no SIASG, há a necessidade de que seja escolhido um código dentre os itens cadastrados no Catálogo de Material (CATMAT). Tais códigos agrupam itens que, apesar de não possuírem uma granularidade a nível de objeto, representam agrupamentos de artefatos de natureza semelhante. Em face do exposto, objetivou-se clusterizar as descrições complementares para cada código de CATMAT isoladamente, pois entende-se que, em tese, já se constituem de um agrupamento preliminar de itens semelhantes, o que contribuiria para a obtenção de melhores resultados. Observou-se também a sugestão de quantidade de *clusters* obtida por meio dos métodos DBSCAN e *elbow*. A seguir, foram utilizados algoritmos de clusterização para agrupar as descrições, representadas por meio dos vetores características, que correspondam a um mesmo produto, discernindo-as dos demais itens. Para realizar a clusterização por meio do DBSCAN e do Kmeans e utilizar o *elbow method*, lançou- se mão da biblioteca *Sklearn*.

3.5. Avaliação do Modelo

Avaliou-se a inicialmente a qualidade dos agrupamentos dos produtos gerados na etapa anterior por meio do cálculo do coeficiente de *Silhouette* médio para cada *cluster*. Posteriormente, verificou-se que, frequentemente, os itens mal clusterizados eram aqueles que possuíam os menores coeficientes dentro de cada cluster. Em face do exposto, estabeleceu-se um valor mínimo para a medida em tela, de modo que os registros que não obtivessem o patamar fossem eliminados do *cluster* e da análise. A seguir, recalculou-se o coeficiente médio para cada *cluster*, eliminando aqueles que não alcançassem o valor mínimo estabelecido para os agrupamentos. Por fim, selecionou-se aleatoriamente para análise humana 25 CATMAT dentre os 100 mais frequentes. Como ferramentas de apoio para a última análise citada, utilizou-se de nuvens de palavras, obtidas por meio da biblioteca *wordcloud*. Em relação ao coeficiente de *Silhouette*, foram utilizadas as bibliotecas *Sklearn*, para o calcular, e *Plotly*, para gerar seus gráficos.

3.6. Implantação

Após a validação do processo de agrupamento, faz-se possível a comparação entre os diversos itens que encontram dentro de um mesmo cluster. Dentre as possibilidades, elenca-se o cálculo do preço médio e mediano de referência do grupo e a posterior comparação de cada elemento com aquele valor, objetivando-se identificar eventuais indícios de sobrepreço.

4 APLICAÇÃO DA SOLUÇÃO

Para a extração das informações e criação da base de a ser utilizada neste trabalho, foram realizados vários testes em diferentes sites de disponibilização de dados abertos. Inicialmente, tentou-se baixar os dados referentes a licitações e a notas de empenho diretamente do Portal da Transparência ou utilizar a API constante do mesmo site,

porém havia restrições como descrições muito curtas das compras públicas, indisponibilidade de dados julgados essenciais e falta de um dicionário de dados para elucidar o significado de colunas e arquivos, dificultando o cruzamento e a consolidação das informações.

Posteriormente, utilizou-se da API de Compras Governamentais, cujo site, no momento da escrita deste artigo, sinaliza encontrar-se em uma versão Beta, porém por meio dela se pode dispor de uma documentação melhor e uma variedade maior de dados. Para a extração desses, foi criado um código em Python para realizar as requisições, obtendo como resposta arquivos no formato JSON, conforme explicado na Seção anterior. Entretanto, foram encontradas algumas dificuldades durante o processo como: tempo muito variável e, em geral, elevado de resposta para as requisições, resultando diversas vezes na total ausência de retorno do servidor ou erros; e, durante a análise exploratória, percebeu-se que determinados campos possuíam uma alta taxa de valores nulos o que, mesmo após se realizar extrações de locais variados, visando mitigar a ausência, limitou a quantidade de dados utilizáveis neste trabalho ao valor citado na Seção 3.2.

Na etapa de preparação dos dados, houve especial cuidado para que fossem geradas boas *features* ao final do pré-processamento. Com a finalidade de gerar um banco de stop words específicas do domínio, foram analisados diversos registros de forma a identificar palavras que não apresentariam valor semântico para o problema em tela. Para a auxiliar a visualização das palavras mais frequentes, foi utilizado um gráfico de *wordcloud*, que foi julgado como de grande valia no processo

Ao se realizar a análise exploratória no conjunto de dados, pôde-se verificar a variabilidade do campo do código de CATMAT, que possuía cerca de 64 mil valores distintos. Para se reduzir o escopo de onde seria aplicada a solução a um montante exequível, foram avaliados 25 dentre os seus 100 valores mais frequentes, resultando em categorias que possuíam entre 2000 e 180 registros cada.

No que concerne à escolha do parâmetro "k", que representa a quantidade de *clusters* por CATMAT, inicialmente se analisou a formação de "cotovelos" por meio do método de *elbow*. Todavia, não se observou que a utilização do parâmetro na quantidade sugerida pelo método tenha sido proveitosa, quando submetida à crítica humana. Os valores costumavam ser muito elevados e tendiam a aproximar os agrupamentos das descrições textuais individualizadas, formando *clusters* de itens únicos ou algo próximo a isso. Em direção oposta, os valores sugeridos pelo DBSCAN tendiam a ser muito pequenos, agrupando no mesmo cluster itens muito diferentes e não atribuindo um cluster a muitos elementos.

Em face do exposto no último parágrafo, o coeficiente de *Silhouette*, em conjunto da análise humana, passou a ser a principal ferramenta para auxiliar na obtenção dos valores da quantidade de *clusters* para cada código de CATMAT. Nesse sentido, lançou-se mão de gráficos para análise do coeficiente, que foram de grande utilidade para se identificar a quantidade adequada do parâmetro e a interação entre variações no parâmetro "k", os valores do coeficiente e os impactos práticos nos agrupamentos, observáveis pela análise humana. A Figura 2 ilustra um exemplo de gráfico de *Silhouette* utilizado para a avaliação da qualidade dos *clusters* para em determinado *k*.

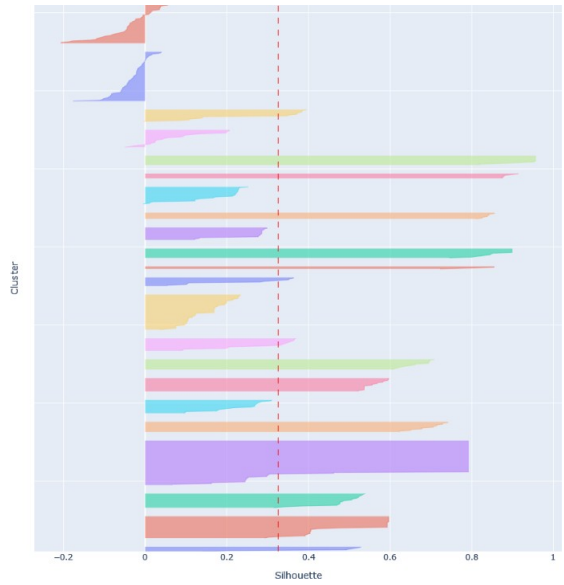

Figura 2. Exemplo de gráfico de *Silhouette* com 22 *clusters*.

Por meio das ferramentas citadas no parágrafo anterior, observou-se que, neste domínio do problema, o coeficiente de *Silhouette* tende a crescer em conjunto do parâmetro "k", até que o último atinja um valor máximo, quando tende a reduzir seu valor ou a manter no patamar independente do aumento da quantidade de agrupamentos. A despeito do exposto, percebeu-se que, por vezes, aumentos do coeficiente em tela significam agrupar em *clusters* diferentes descrições que, apesar de escritas de forma diversas, representam o mesmo item. Dessa forma, na maioria dos experimentos realizados, após a análise humana, verificou-se que o melhor valor para o parâmetro "k" se encontrava entre o fornecido pelo DBSCAN ("valor mínimo") e aquele que maximizava o coeficiente de *Silhouette* ("valor máximo").

Durante o processo para a escolha do melhor número de *clusters* para cada CATMAT, pôde-se observar inúmeros casos de itens que foram registrados no SIASG com o código errado. Na ocorrência de tais casos, buscou-se não subdividir essas anomalias, tentando agrupá-las na menor quantidade de *clusters* possível para posterior remoção. A Tabela 2 apresenta alguns exemplos de itens anotados em categorias erradas.

Item	Descrição do CATMAT
sanduicheira elétrica: material: aço inox [...]	Gás Liquefeito De Petróleo (GLP)
colar elizabetano para cães [...]	Gás Liquefeito De Petróleo (GLP)
leite em pó [...]	Óleo Vegetal Comestível
óleo para motor de popa [...]	Óleo Vegetal Comestível
bisturi oftalmológico [...]	Álcool Etílico
monitor computador, tamanho tela: 23 [...]	Álcool Etílico

Tabela 2: Exemplos de itens associados a códigos de materiais errados

Observou-se que a utilização de um valor mínimo de 0,2 para os coeficientes de *Silhouette* médios de cada *cluster*, assim como para cada elemento individualmente, como regra de corte, foi o

suficiente na maioria dos casos para afastar os itens erroneamente registrados, como também para aperfeiçoar o processo de clusterização em si.

Após a análise humana, em conjunto da utilização do coeficiente de Silhouette, dos 25 CATMAT selecionados aleatoriamente dentre os 100 mais frequentes, observou-se que em 18 deles (72%) foi obtido um bom resultado. Nos 7 outros casos (28%), percebeu-se que o algoritmo de clusterização foi capaz de realizar agrupamentos com valores semânticos coerentes, mas que não seriam capazes de proporcionar uma comparação plena entre os itens, sendo julgados como insatisfatórios.

Debruçando-se mais detidamente sobre os casos julgados insatisfatórios, realizou-se uma análise em que se chegou ao entendimento de que poderia haver três grupos principais de características que levaram ao insucesso daquelas ocorrências. No primeiro deles, a descrição do item em si era muito densa, ou seja, havia um conjunto muito grande de características distintivas em cada uma delas. Pode-se ilustrar o ocorrido por meio da Tabela 3, que representa o agrupamento de itens similares, porém diferentes, distinguindo-se pelo comprimento do parafuso por exemplo, no mesmo *cluster*. No segundo caso, havia uma quantidade substancial de itens registrados originalmente com o código de CATMAT equivocado, o que dificultou sobremaneira seu isolamento daqueles corretos, assim como sua eliminação, mesmo se valendo do valor mínimo do coeficiente de *Silhouette*. A Tabela 4 exemplifica o ocorrido, ao elencar itens diversos que estão registrados no SIASG equivocadamente como "luva de proteção". No terceiro e último grupo, há itens que possuíam descrições muito longas, mas em que apenas poucas palavras realizavam a distinção de fato. A Tabela 5 explana o exemplo em que, a despeito do comprimento da descrição, apenas a fabricante do automóvel e o porte do veículo possuiriam valor para distinguir a qual cluster cada item deveria pertencer.

Item	Cluster
PARAFUSO - Allen 5/16" TIPO: Allen CABEÇA: Cilindrica SEXTAVADO: Interno MATERIAL: Aço Inoxidável (304) ACABAMENTO: Passivado ROSCA: UNC F.P.P: 18 COMPRIMENTO: 1/2"	2
PARAFUSO - Allen 1/2" TIPO: Allen CABEÇA: Cilindrica SEXTAVADO: Interno MATERIAL: Aço Inoxidável (304) ACABAMENTO: Passivado ROSCA: UNC F.P.P: 13 COMPRIMENTO: 1.1/2"	2
PARAFUSO - Allen 1/2" TIPO: Allen CABEÇA: Cilindrica SEXTAVADO: Interno MATERIAL: Aço Inoxidável (304) ACABAMENTO: Passivado ROSCA: UNC F.P.P: 13 COMPRIMENTO: 1"	2

Tabela 3. Exemplos de itens com descrições densas, dificultando a divisão completa em itens distintos.

Item	Descrição do CATMAT
Legume in natura, tipo: pimentão verde	Luva de proteção
Verdura in natura, tipo: alface americana	Luva de proteção
Pegador para gelo em aço inox.	Luva de proteção
Água mineral natural, tipo: sem gás [...]	Luva de proteção
Molho de mesa, tipo: catchup [...]	Luva de proteção

Tabela 4. Exemplos de itens associados a códigos de materiais errados que não foram cortados pelo coeficiente *Silhouette*.

Item	Cluster
Peças para a manutenção corretiva e preventiva de veículos automotores para os diversos tipos de viaturas da marca FORD.	7
Peças para a manutenção corretiva e preventiva de veículos automotores para os diversos tipos de viaturas da marca NISSAN.	7
Fornecimento de peças mecânicas, elétricas e de funilaria para veículos PESADOS da marca VOLKSWAGEN, com as mesmas características e especificações técnicas, de peças de produção original (ABNT NR 15296), do fabricante, com o maior desconto sobre a tabela da montadora/ fabricante.	19
Fornecimento de peças mecânicas, elétricas e de funilaria para veículos LEVES da marca FIAT, com as mesmas características e especificações técnicas, de peças de produção original (ABNT NR 15296), do fabricante, com o maior desconto sobre a Tabela da Montadora/ Fabricante.	19

Tabela 5. Exemplos de itens com descrições muito longas e com poucas palavras de diferenciação.

Após a avaliação da qualidade dos agrupamentos, é possível os utilizar para realizar diversas análises, dentre elas, identificar *outliers* nos preços de aquisição dos itens. De forma a exemplificar esta aplicação, utilizou-se de um dos *clusters* gerados em conjunto com a informação da unidade de fornecimento do item para se calcular o preço médio e mediano.

Na Figura 3 é possível ver um gráfico com os itens avaliados sob a ótica do preço licitado e da quantidade licitada, sendo traçadas duas linhas representando o valor médio (linha vermelha) e o valor mediano (linha verde). Devido à grande variação na magnitude dos valores, ambos os eixos estão em escala logarítmica. No gráfico é possível perceber alguns *outliers* com valores muito acima ou muito abaixo do esperado, auxiliando dessa forma na identificação de indícios de eventuais impropriedades na aplicação do recurso público, assim como na de possíveis boas práticas de gestão.

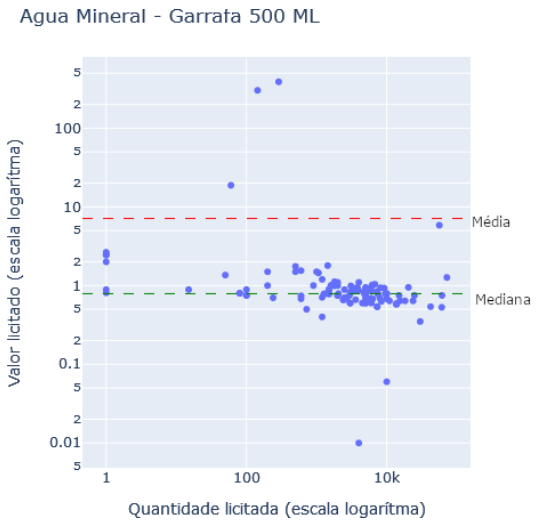

Figura 3. Gráfico valor pago x quantidade adquirida do item água mineral – garrada 500ml

Conforme discorrido ao longo deste trabalho, o emprego de Mineração de Textos possui grande potencial de aplicação em dados públicos. Nesta Seção, foi demonstrado ser possível utilizar tais técnicas e aplicar seus resultados em análises de forma a identificar indícios de má aplicação do recurso público. Ademais, seu emprego nos dados de compras governamentais é de grande valia para processar de forma mais automática o volume de dados disponível nas bases governamentais. Entretanto devido ao grande espectro de itens adquiridos pelo setor público (foram analisados por este trabalho itens tão diversos quanto ração animal, combustíveis e material de escritório, por exemplo) há a necessidade de envolver especialistas das áreas de negócio para se refinar os resultados e realizar mais análises.

5 CONCLUSÃO

Este artigo utilizou algoritmos de *Machine Learning* e mineração de textos em descrições textuais de compras públicas com o objetivo de extrair informações relevantes de dados não estruturados, agrupando itens semelhantes e possibilitando a realização de análises mais aprofundadas, auxiliando na detecção de indícios de impropriedades na aplicação de recursos públicos.

Após a extração dos dados de portais públicos, foi necessário realizar um intenso processo para o tratamento do texto e possibilitar sua utilização em algoritmos de agrupamento. As etapas de extração e tratamento dos dados foram as que demandaram mais tempo e atenção durante todo o processo. Na aplicação da clusterização foram encontrados alguns desafios que podem ser melhor explorados futuramente para um aperfeiçoamento dos resultados. Conforme descrito na Seção 4, ao longo da execução dos experimentos observou-se, via análise humana e após a utilização do coeficiente de Silhouette, que 72% dos casos avaliados resultaram em agrupamentos com valores semânticos coerentes o suficiente para realizar comparações entre os preços licitados dentro de cada *cluster*.

Ademais destacam-se algumas possibilidades de melhorias para trabalhos futuros, de forma a aperfeiçoar os resultados: melhorar a obtenção dos dados, testando a extração das informações de outras fontes mais estáveis; avaliar formas de tratamento para itens com CATMAT designados de forma errônea; e investigar melhorias no pré-processamento das descrições, de maneira a melhorar o desempenho da metodologia em descrições muito grandes ou muito complexas.

REFERÊNCIAS

[AMARAL e RODRIGUES, 2020] AMARAL, João Alberto; RODRIGUES, Jairson Barbosa. Alocação de Tópicos Latentes—Um Modelo para Segmentação de Dados de Auditoria do Governo de PE. Revista de Engenharia e Pesquisa Aplicada, 2020, 5.1: 40-49.

[ALMEIDA *et al*, 2018] ALMEIDA, Gustavo *et al*. Improvement of transparency through mining techniques for reclassification of texts: the case of brazilian transparency portal. In: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age. 2018. p. 1-9.

[BRASIL, 1964] BRASIL. LEI Nº 4.320, DE 17 DE MARÇO DE 1964. Estatui Normas Gerais de Direito Financeiro para elaboração e contrôle dos orçamentos e balanços da União, dos Estados, dos Municípios e do Distrito Federal. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l4320.htm

[BRASIL, 1988] BRASIL. Constituição da República Federativa do Brasil. Brasília, DF: Senado Federal: Centro Gráfico, 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm

[BRASIL, 1993] BRASIL. LEI Nº 8.666, DE 21 DE JUNHO DE 1993.Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/LEIS/L8666cons.htm

[BRASIL, 2002] BRASIL. LEI Nº 10.520, DE 17 DE JULHO DE 2002. Institui, no âmbito da União, Estados, Distrito Federal e Municípios, nos termos do art. 37, inciso XXI, da Constituição Federal, modalidade de licitação denominada pregão, para aquisição de bens e serviços comuns, e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/2002/L10520.htm

[BRASIL, 2009] BRASIL. LEI COMPLEMENTAR Nº 131, DE 27 DE MAIO DE 2009. Acrescenta dispositivos à Lei Complementar no 101, de 4 de maio de 2000, que estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp131.htm

[BRASIL, 2011] BRASIL. Lei nº 12.527, de 18 de novembro de 2011. Lei de Acesso à Informação. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

[BRASIL, 2021] BRASIL. LEI Nº 14.133, DE 1º DE ABRIL DE 2021. Lei de Licitações e Contratos Administrativos. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm

[BALANIUK, 2010] BALANIUK, Remis. A Mineração de Dados como apoio ao Controle Externo. Revista do TCU, n. 117, p. 79-86, 2010.

[BENGFORT *et al*, 2018] BENGFORT, Benjamin *et al*. Applied text analysis with python: Enabling language-aware data products with machine learning. " O'Reilly Media, Inc.", 2018.

[CARVALHO *et al*, 2014] CARVALHO, Rommel *et al*. Using Clustering and Text Mining to Create a Reference Price Database. Learning and NonLinear Models, v. 12, n. 2014, p. 38-52, 2014.

[CARVALHO, 2015] CARVALHO, Rommel Novaes. Categoria Profissionais 2° Lugar: Uso de mineração de dados e textos para cálculo de preços de referência em compras do governo brasileiro. 2015.

[ESCOVEDO e KOSHIYAMA, 2020] ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise. Casa do Código, 2020.

[ESTER *et al*, 1996] ESTER, Martin *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd. 1996. p. 226-231.

[FELDMAN *et al*, 2007] FELDMAN, Ronen *et al*. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press, 2007.

[FONTES, 2022] FONTES, Raphael Silva. Avaliação experimental de um classificador para apoiar a detecção de fraudes em compras públicas. 2022.

[IGUAL e SEGUÍ, 2017] IGUAL, Laura; SEGUÍ, Santi. Introduction to data science. In: Introduction to data science. Springer, Cham, 2017. p. 1-4.

[PAIVA, 2017] PAIVA, Eduardo Soares de. Geração de regras de identificação de produtos em descrições textuais de compras apresentadas em portais de transparência pública. 2017. Dissertação de Mestrado.

[SARKAR, 2016] SARKAR, Dipanjan. Text analytics with python. New York, NY, USA:: Apress, 2016.

[SHEARER, 2000] SHEARER, Colin. The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing, v. 5, n. 4, p. 13-22, 2000.

[WEISS e INDURKHYA, 2015] WEISS, Sholom M.; INDURKHYA, Nitin; ZHANG, Tong. Fundamentals of predictive text mining. Springer, 2015