# Classificação semântica de pedidos de acesso à informação[1]

*Clasificación semántica de solicitudes de acceso a la información*

*Semantic classification of requests to information*

*Flávia Lemos Sampaio Xavier, Ricardo Brigato Scheicher e Roberta Akemi Sinoara*

https://doi.org/10.36428/revistadacgu.v14i27.537

**Resumo:** Desde o início da implementação da Lei de Acesso à Informação no Brasil até 2020, existiu uma demanda crescente de pedidos de acesso à informação no âmbito da Controladoria-Geral da União (CGU) e de todo Poder Executivo federal. A busca por um modelo de automação de processos utilizando Inteligência Artificial visa levar à redução de custos para a administração pública e à melhoria das condições de trabalho, bem como auxilia na eficiência das respostas à sociedade. Neste trabalho, foi realizada a aplicação do método de classificação semanticamente enriquecida por expressões do domínio com uma análise comparativa dos resultados de classificação dos pedidos de acesso à informação usando como base algoritmos com diferentes níveis de explicabilidade e transparência para o processo. A melhor acurácia foi obtida pelo modelo do algoritmo *Support Vector Machine*, com valor de 91,1% e Medida-F1 *Weighted* de 91,7%, enriquecido pela representação de textos gBoED. Outros destaques também podem ser observados para algoritmos que oferecem maior explicabilidade. Os resultados apresentaram grande potencial quanto ao uso deste modelo para classificação dos pedidos de acesso à informação não apenas na CGU, mas em todo o setor público.

**Palavras-chave:** Direito de Acesso à Informação, Mineração de Textos, Classificação Semântica; Transparência.

**Resumen:** Desde el principio de la aplicación de la Ley de Acceso a la Información hasta 2020 hubo una creciente demanda de solicitudes de acceso a la información en el ámbito de la Contraloría General de la Unión (CGU) y de todo el poder ejecutivo federal. La investigación de un proceso de clasificación de pedidos más automatizado, con el uso de Inteligencia Artificial, tiene como objetivo la reducción de costes para la administración pública, la mejora de las condiciones laborales de los servidores que realizan esta tarea y apoya la elaboración de respuestas más rápidas para la sociedad. En este trabajo se realizó la aplicación del método de clasificación semánticamente enriquecido por expresiones de dominio con un análisis comparativo de los resultados de clasificación de las solicitudes de acceso a la información utilizando algoritmos con diferentes niveles de explicabilidad y transparencia para el proceso. La mejor acurácia obtenida fue por el modelo de algoritmo *Support Vector Machine*, con un valor del 91,1% y una Medida-F1 *Weighted* del 91,7%, enriquecido con la representación de textos gBoED. Otros aspectos destacados los resultados del modelo generado por algoritmos que ofrecem una mayor explicabilidad. Los resultados mostraron un gran potencial en cuanto al

1. Artigo submetido em 17/07/2022 e aceito em 09/01/2023.

uso de este modelo para clasificar las solicitudes de acceso a la información no solo en la CGU sino en todo el sector público.

**Palabras clave:** Derecho de Acceso a la Información, Minería de Textos, Clasificación Semántica; Transparencia.

**Abstract:** Since the Freedom of Information Act implementation until 2020, there was a growing demand for requests to information, within the scope of the Office of the Comptroller General (CGU) and the entire federal executive branch. The search for a process automation model using Artificial Intelligence aims to achieve cost reduction for the public administration and improvement of the working conditions as well as supports the response efficiency to society. This work applied a method for classification improvement using semantically enriched information derived from domain expressions and carried out a comparative analysis of the classification results of requests to information using algorithms with different explainability and transparency levels. The best accuracy obtained by the Support Vector Machine algorithm model was 91,1% and Weighted score-F1 of 91,7%, enriched by the gBoED text representation model. Also noteworthy are the results of the model generated by algorithms that offers greater. The results indicate great potential regarding the use of this model to classify requests to information not only at the CGU but also across the public sector.

**Keywords:** Right of Access to Information, Text Mining, Semantic Classification; Transparency.

## INTRODUÇÃO

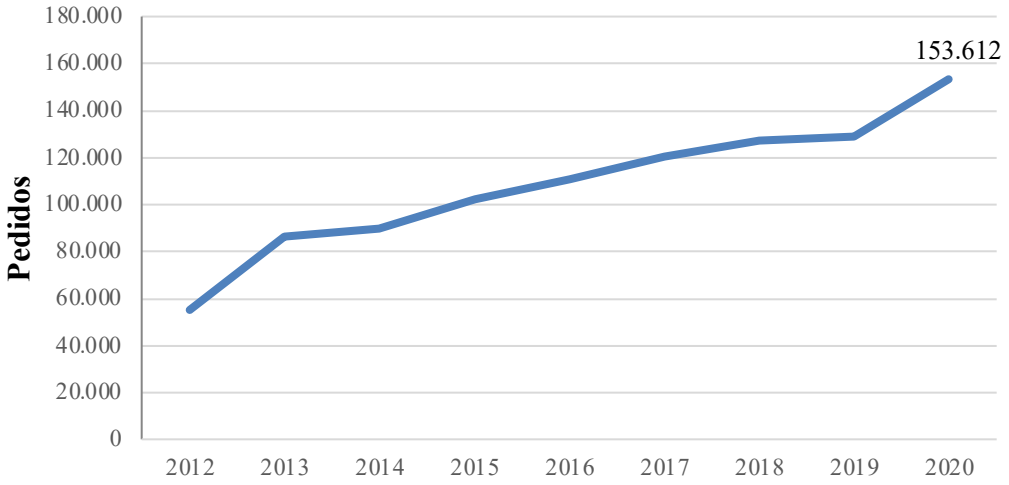
O direito básico de acesso à informação é garantido pela Constituição brasileira de 1988 em seu artigo 5º, inciso XXXIII. A Lei de Acesso à Informação (LAI), Lei nº 12.527 (2011) regulamenta este direito constitucional, para estabelecer que os órgãos e entidades públicas devam garantir um processo transparente de gestão da informação, por amplo acesso e divulgação; disponibilidade, autenticidade e integridade; proteção de informações confidenciais e informações pessoais e, eventualmente, restrição de acesso à informação, nos casos em que a publicidade de tais informações pode colocar em risco a segurança da sociedade ou do Estado.

De acordo com o Painel da Lei de Acesso à Informação[2], desenvolvido pela Controladoria-Geral da União (CGU), existe uma demanda crescente pela implementação efetiva de meios que garantam o direito de acesso à informação no Brasil. Segundo dados deste Painel, somente os órgãos e entidades federais receberam cerca de um milhão de pedidos de acesso à informação desde maio de 2012, início da vigência da LAI no Brasil. Em 2012, primeiro ano da implementação, foram registrados 55.025 pedidos e, no ano de 2020, em que houve a maior demanda absoluta da série histórica até este ano, alcançou um total de 153.612 pedidos, correspondendo a um crescimento de 180%. No âmbito da CGU, no primeiro ano foram registrados 1.192 pedidos e, em 2020, 1.484 pedidos, portanto, a tendência de crescimento neste órgão foi de 25%. Nas Figuras 1 e 2 são apresentados gráficos da evolução anual dos pedidos de acesso à informação ao Poder Executivo Federal e à CGU, respectivamente.

Segundo a LAI os pedidos de informação solicitados pelos cidadãos à administração pública devem ser respondidos no prazo de 20 dias, prorrogáveis por mais 10 dias mediante justificativa. Os pedidos de informação são submetidos a uma análise por parte dos servidores públicos responsáveis e são avaliados pela possibilidade de determinada informação ser Concedida ou Negada. Portanto, esta fase consiste em realizar a triagem dos pedidos, com base em pesquisas dos precedentes jurisprudenciais que contenham semelhanças temáticas e circunstanciais aos novos pedidos de acesso. Ela precede a elaboração de pareceres técnicos realizados pelos servidores da equipe da CGU, que visam fundamentar a decisão final da autoridade, a quem cabe decidir no órgão pelo tipo de resposta ao requerente.

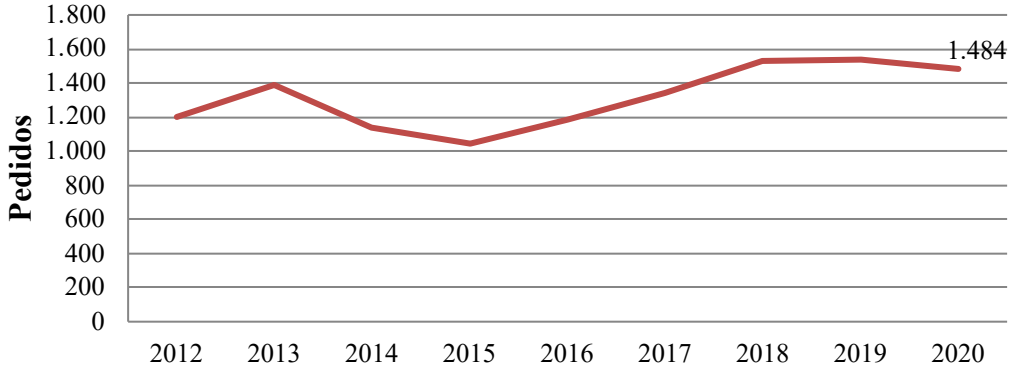
2. Controladoria-Geral da União. (2020). Painel Lei de Acesso à Informação. Disponível em http://paineis.cgu.gov.br/lai/index.htm

**FIGURA 1 - EVOLUÇÃO ANUAL DA QUANTIDADE DE PEDIDOS DE ACESSO À INFORMAÇÃO AO PODER EXECUTIVO FEDERAL DE 2012 A 2020**




*Fonte: Extraída do Painel da Lei de Acesso à Informação.*

**FIGURA 2 - EVOLUÇÃO ANUAL DA QUANTIDADE DE PEDIDOS DE ACESSO À INFORMAÇÃO À CGU DE 2012 A 2020**




*Fonte: Extraída do Painel da Lei de Acesso à Informação.*

Diante da crescente demanda de pedidos, os servidores responsáveis por receber e analisar a viabilidade da concessão das informações solicitadas necessitam de meios cada vez mais confiáveis e eficientes de realizarem esse processo e cumprirem o prazo determinado. Com base nessa necessidade, por meio de técnicas de Inteligência Artificial (IA) e Mineração de Textos (MT), este trabalho visa à aplicação de algoritmos de classificação de textos considerando aspectos semânticos, com o objetivo de tornar o trabalho dos profissionais mais rápido e preciso, contribuindo com a automação e com o alcance de maior eficiência da fase inicial do processo de resposta aos pedidos de informação à CGU.

Portanto, neste trabalho desenvolveu-se um modelo preditivo, transparente e replicável que possibilite ao servidor da CGU contar com a classificação automatizada dos novos pedidos de acesso à informação, no processo de triagem, com base no banco de precedentes dos pedidos à CGU. O método pode ser caracterizado como um incremento aos métodos tradicionais de classificação e validado como um agregador de conhecimento à tomada de decisão no processo de triagem de pedidos de informação na

CGU. A generalização deste modelo tem potencial de ser adaptada às necessidades de outros órgãos do Poder Executivo federal e até mesmo de órgãos públicos estaduais e municipais.

## TRABALHOS RELACIONADOS

Para contextualizar o estudo, é importante esclarecer que a utilização de pedidos e outras demandas dos cidadãos como subsídio para melhorar a qualidade dos serviços oferecidos pelo Estado é uma prática comum no Brasil e no mundo. Com a crescente evolução de técnicas de mineração de texto e capacidade computacional nos últimos anos, observa-se um aumento do seu emprego na automatização de processos relacionados à classificação e análise de grande volume de informações não estruturadas que chegam aos órgãos públicos diariamente.

Esse entendimento de que associar tecnologia, serviços governamentais e participação social é uma tendência, é afirmado em Mehr, Ash e Fellow (2017) e enfatizado em Chun et al. (2010, p.1):

> A revolução nas tecnologias da informação e comunicação (TIC) vem mudando não apenas o cotidiano das pessoas, mas também as interações entre governos e cidadãos. O governo digital ou o governo eletrônico começou como uma nova forma de organização pública que suporta e redefine as informações novas e existentes, as comunicações e as interações relacionadas às transações com as partes interessadas (por exemplo, cidadãos e empresas) por meio das TIC, especialmente por meio de serviços de internet, com o objetivo de melhorar o desempenho e os processos do governo.

Tjandra, Warsito e Sugiono (2015) apresentam um exemplo de automatização no processo de atendimento em serviços oferecidos aos cidadãos da cidade de Surabaia na Indonésia. A ferramenta resolve o direcionamento de denúncias e outras comunicações cidadãs pela combinação de algumas técnicas de pré-processamento de textos e o uso de um algoritmo de classificação de documentos, com vistas apoiar a triagem apontando para qual departamento da cidade a manifestação deverá ser direcionada.

No Brasil, destacam-se alguns modelos de classificação automatizada desenvolvidos por órgãos públicos. Exemplo recente é o "Victor", ferramenta de IA desenvolvida pelo Supremo Tribunal Federal (STF) em conjunto a Universidade de Brasília (UnB) que tem por objetivo apoiar o serviço judiciário. A ferramenta lê todos os Recursos Extraordinários que chegam ao STF e identifica também por meio de um problema de classificação quais desses recursos vinculam aos temas de repercussão geral, conforme o requisito determinado pelo art. 102, § 3º, da Constituição Federal. O objetivo foi padronizar a atividade e aprimorar os resultados de atendimento às demandas sociais pelo tribunal (Maia Filho & Junquilho, 2018).

A CGU e o Tribunal de Contas da União (TCU) desenvolveram modelo chamado "Alice" (Análise de Licitações e Editais). Trata-se de modelo que lê as licitações, contratos e os editais publicados nos Diários Oficiais, trazendo aos auditores o número de processos por estado, assim como o valor dos riscos de cada um. Considerando esses dados, o modelo produz um documento informando se há indícios de fraudes ou não. O modelo, com apoio de outras ferramentas das TICs utilizadas pela CGU, foi avaliado como eficaz na constatação de irregularidades em licitações públicas, promovendo a economicidade dos recursos públicos com detecções de cerca de R$ 11,2 bilhões de reais de irregularidades com recursos para o enfrentamento da Covid-19 (Panis, 2020; Carvalho, 2020).

A CGU desenvolveu outro modelo de IA para apoiar a triagem nas análises das denúncias quanto ao risco existente no âmbito das ouvidorias públicas (Ferramenta de Análise de Risco em Ouvidorias - FARO). O processo de triagem consiste em avaliar se há o mínimo de informações necessárias presente na denúncia e, caso haja, encaminhar a denúncia para que uma área competente realize a apuração dos fatos denunciados. A principal contribuição apresentada pelo modelo foi uma abordagem baseada na extração de entidades nomeadas, o que permitiu que se utilizassem bases de dados oficiais do governo federal para o enriquecimento das informações extraídas diretamente do conteúdo da denúncia e de seus anexos. Tais informações puderam, então, ser utilizadas como variáveis em um modelo de classificação. Os resultados do FARO alcançaram acurácia balanceada de 76% e medida F1 para a classe positiva de 61% e foram considerados adequados para o seu emprego no fluxo de trabalho da CGU. Atualmente o modelo está implantado, com

monitoramento automatizado e apoia a triagem e análise de risco de mais de 300 denúncias recebidas por mês na instituição.

Fundamentação Teórica

Dados obtidos do mundo real são volumosos e, de maneira geral, não possuem qualquer tratamento, podendo ter origem em diferentes fontes e apresentar formatos variados. Segundo os autores García, Luengo e Herrera (2015), três elementos definem a qualidade dos dados: precisão, integridade e consistência. Infelizmente, a maioria dos conjuntos de dados do mundo real apresentam condições opostas e, caso não sejam pré-processados, o conhecimento proveniente desses dados pode não ser útil nem válido.

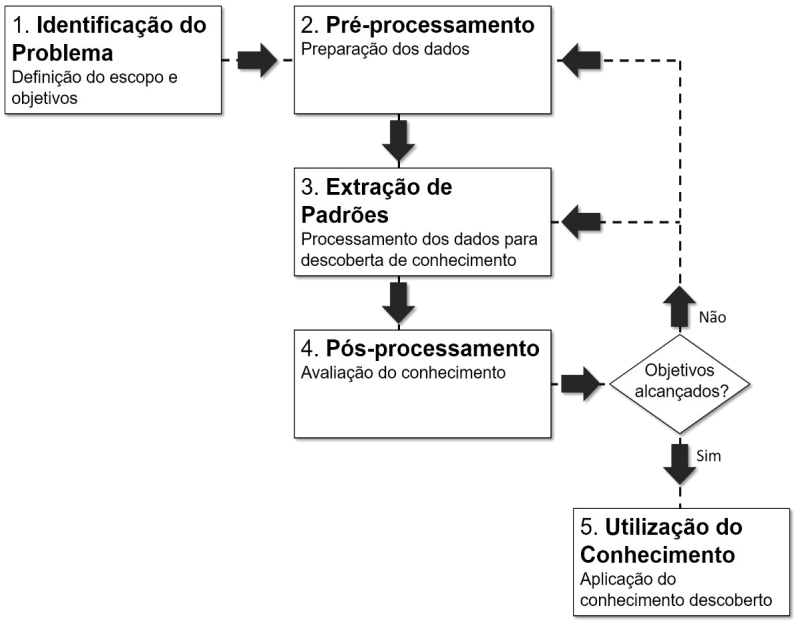
Conforme citado em Rezende, Pugliesi, Melanda e Paula (2003), as bases de dados das grandes empresas e de instituições públicas contêm uma potencial mina de ouro de informações valiosas, porém, de acordo com Rezende et al. (2003), estes dados raramente são obtidos de forma direta, estando presentes de modo semiestruturados ou não estruturados. Existem 3 formas de classificar os dados de acordo com sua estrutura: (i) dados estruturados; (ii) dados semiestruturados; e (iii) dados não estruturados.

Os dados estruturados são organizados em um padrão fixo, podendo ser relacionados com atributos ou variáveis, enquanto os dados não estruturados não seguem uma estrutura rígida. O exemplo mais típico de dados estruturados é um banco de dados organizados conforme a definição de um esquema, que define as tabelas com seus respectivos campos (ou atributos) e tipos (formato). O esquema pode ser pensado como uma metainformação do banco de dados, ou seja, uma descrição sobre a organização dos dados que serão armazenados no banco. Por outro lado, exemplos de dados não estruturados são: e-mails, artigos, pareceres, documentos em PDF, imagens, comentários e postagens em redes sociais, interação entre consumidores, áudios e vídeos. Os dados semiestruturados, por seu turno, ficam entre os extremos: não são estruturados de forma rígida, mas também não são totalmente não estruturados. Apresentam uma representação estrutural heterogênea, como os arquivos nos formatos HTML das páginas da Web.

Atualmente muitos esforços são direcionados a obter dados estruturados a partir de dados não estruturados ou semiestruturados, para que estes possam ser utilizados para extração de conhecimento. No caso da extração de conhecimento de dados textuais, ou seja, textos escritos em língua natural, tem-se o processo de Mineração de Textos, considerado uma extensão da Mineração de Dados (MD), que possui o objetivo de obter conhecimento útil a partir desses dados textuais, para a utilização em tarefas de tomada de decisão. Utiliza-se, portanto, da aplicação de técnicas para analisar esse tipo de dados não estruturados e descobrir padrões que não eram conhecidos previamente (Sinoara, 2018).

O processo de Mineração de Textos é composto por cinco etapas principais, como é possível verificar na Figura 3.

**FIGURA 3 - PROCESSO DE MINERAÇÃO DE TEXTOS**




*Fonte: Sinoara (2018)*

A primeira etapa consiste na identificação do problema e definição do escopo da coleção de textos. Em seguida, na segunda etapa ocorre o pré-processamento ou a preparação dos dados, que inclui a aplicação de técnicas para limpeza, remoção de palavras irrelevantes (denominadas stopwords) e padronização dos textos. A terceira etapa consiste da extração de padrões em busca de obtenção de conhecimento, que compreende a escolha da tarefa de mineração a ser empregada, a escolha do algoritmo e a extração dos padrões em um modelo de aprendizado. Na etapa de pós-processamento, medidas para avaliação do conhecimento são aplicadas de modo a identificar a qualidade do modelo gerado. Estas medidas podem ser divididas entre medidas de desempenho, como precisão, erro, suporte e tempo de aprendizado, e medidas de qualidade. Por último, na quinta etapa, caso os objetivos tenham sido alcançados, ocorre a utilização do conhecimento em aplicações de destino. Caso contrário, um novo ciclo deve ser executado com mudanças nas etapas anteriores, como em técnicas de pré-processamento e na definição de parâmetros para extração de padrões (Sinoara, 2018; Rezende et al., 2003).

O principal desafio do processo de MT está na etapa de pré-processamento, no qual a semântica dos textos deve ser considerada ao representar os textos em um formato adequado para o algoritmo de aprendizado de máquina. As representações baseadas no modelo espaço-vetorial são as mais comuns da área de aprendizado de máquina e são as mais utilizadas na classificação automática de textos (Rossi, 2015). Neste modelo, os documentos correspondem aos vetores e as dimensões correspondem a termos ou atributos da coleção de textos. Entende-se por "termos" as dimensões geradas com base nas palavras de um texto, seja uma palavra simples, sejam sequências ou conjuntos de palavras. Para os objetivos deste estudo, consideram-se as dimensões das representações baseadas no modelo espaço-vetorial relativos aos termos. A representação que utiliza palavras simples como termos da coleção de documentos, gerando uma matriz documento-termo é denominada *Bag-of-Words (BoW)* (Rossi, 2015).

A escolha do algoritmo ou da tarefa na etapa de extração de padrões é feita com base nos dados disponíveis e no tipo de conhecimento que se deseja descobrir, podendo corresponder a atividades preditivas ou descritivas.

As atividades preditivas consistem na generalização de exemplos ou experiências passadas com respostas conhecidas em uma linguagem capaz de reconhecer a classe de um novo exemplo. As duas principais tarefas preditivas são classificação e

regressão. Enquanto a classificação consiste na análise das características dos documentos para atribuição a uma categoria específica previamente definida (valor categórico), a regressão consiste na predição de um valor contínuo, como predizer o lucro.

Já as atividades descritivas consistem na identificação de comportamentos intrínsecos do conjunto de textos, sem uma classe especificada. Algumas tarefas descritivas são o agrupamento (*clustering*), que consiste na tarefa de aproximação ou agrupamento dos registros com base nas similaridades entre eles, e a associação, que consiste na identificação dos relacionamentos dos atributos, no formato condicional de "se...então" (Rezende et al., 2003).

Uma vez eleita a tarefa a ser empregada, existe uma variedade de algoritmos para executá-la. A escolha de algoritmo é realizada de forma subordinada à linguagem de representação dos padrões a serem encontrados. Podem-se utilizar algoritmos dos mais variados para testar a sua acurácia ou outras métricas. Para classificação tem-se, por exemplo, as árvores de decisão, regras de produção, modelos lineares, modelos não lineares (Redes Neurais Artificiais), modelos baseados em exemplos (*K-Nearest Neighbors - KNN*) e modelos de dependência probabilística (Redes Bayesianas).

A classificação é considerada uma tarefa de aprendizado supervisionado, que visa extrair padrões de um conjunto de exemplos de entrada e, a partir dos padrões aprendidos, mapear novos exemplos em um número finito de categorias. Os exemplos consistem em um conjunto de atributos e um atributo-meta discreto. O objetivo de um algoritmo supervisionado é encontrar padrões entre os atributos e uma classe, de modo que o processo de classificação possa usar esses padrões para predizer a classe de um exemplo novo e desconhecido. Assim, a classificação consiste em obter um modelo baseado em um conjunto de exemplos que descrevem uma função não conhecida. Esse modelo é então utilizado para predizer o valor do atributo-meta de novos exemplos (Rezende et al., 2003, p.17-18). Classificar plantas ou animais, identificar e-mails como Spam e definir se uma informação pública deve ter acesso concedido ou se deve ter o acesso negado são exemplos de problemas de classificação.

A classificação de textos pode ser dividida em dois níveis de complexidade: por tópicos e semântica (Sinoara, 2018). O nível de classificação por tópicos consiste em problemas de classificação que dependem basicamente do vocabulário. Nesse problema, cada classe possui termos bastante característicos, e, portanto, o léxico (vocabulário) possui grande relevância para representar o conteúdo dos documentos. Na classificação de nível semântico são necessárias mais informações do que apenas o léxico. Tais problemas requerem uma análise mais profunda, além apenas das palavras, visto que os documentos de classes distintas podem usar o mesmo vocabulário.

Para exemplificar os diferentes níveis de classificação, Sinoara (2018) apresenta sentenças de documentos extraídos de notícias de esportes. D1: "*Guga é o campeão do Tennis Masters Cup. Ele venceu Agassi por três sets a zero no jogo final*" e D2: "*Hamilton larga na pole position e vence o Grande Prêmio do Canadá. Após colisão, Massa abandona a prova.*". Na classificação por tópicos, a principal questão relacionada ao conjunto de sentenças seria "A qual esporte a sentença faz referência?". Nesse nível de classificação, apenas a presença dos termos "Guga", "Tennis Masters Cup", "sets", "Agassi" e "jogo", em D1, já indica que se trata do esporte Tênis e a presença dos termos "Hamilton", "pole position", "Grande Prêmio", "Massa" e "prova", em D2, já indica o esporte Fórmula 1. Logo, cada classe (ou grupo esperado) pode ser determinada em grande parte pelo vocabulário utilizado e, portanto, nesses casos a representação relativa à frequência das palavras (Bag of Words – BoW) é suficiente para trazer bons resultados nesse tipo de classificação. Porém, caso a classificação dos documentos estivessem relacionadas à questão "Esse documento refere-se à vitória de um atleta brasileiro? ", então é necessário saber que os termos "Guga" e "Massa" correspondem a atletas brasileiros e os termos "campeão" e "vence" referem-se ao significado de vitória.

**Expressões do Domínio** são estruturas capazes de representar um conhecimento, carregando consigo determinado nível semântico devido à união de termos importantes dentro de um determinado domínio e termos que identificam cada uma das classes em um processo de classificação de nível semântico. O trabalho de Marques, Matsuno, Sinoara, Rezende e Rozenfeld (2015) introduz a representação de textos *Bag of Expressions of Domain (BoED)*, aplicada especificamente para o domínio da

área de Desenvolvimento de Produtos e Serviços, área específica da Engenharia de Produção. Marques et al. (2015) aplicou a representação BoED na classificação de documentos de artigos científicos que relatam o desenvolvimento teórico de um método ou a aplicação de métodos já existentes.

Scheicher, Sinoara, Koga e Rezende (2016) realizaram a generalização da representação para diferentes domínios. Tal representação passa a ser denominada *generalized Bag of Expressions of Domain (gBoED)*. Nesse trabalho é definido que cada expressão do domínio corresponde à união de um Termo do domínio e um Identificador de classe. Termos do domínio são termos relacionados ao domínio ou área do problema, importantes para aquela coleção de documentos e para a organização ou classificação esperada como resultado do processo de MT. Identificadores de classe são palavras ou expressões que estão particularmente ligadas a uma determinada classe e, assim, são consideradas como termos ou palavras-chaves daquela classe. Para a formação das expressões do domínio que compõem a gBoED é necessária a construção de uma lista de termos do domínio e um conjunto de listas de identificadores de classe, construídas por especialistas do domínio. Tais elementos são descritos e formalizados a seguir.

- Lista de Termos do Domínio = $\{k_1, k_2, \ldots, k_n\}$
- Listas de Identificadores de Classe = $\{\{ck_{11}, ck_{12}, \ldots, ck_{1j}\}, \ldots, \{ck_{m1}, ck_{m2}, \ldots, ck_{ml}\}\}$

A gBoED corresponde a uma matriz do tipo atributo-valor, cujos atributos formados pelas colunas são as expressões do domínio e as linhas correspondem aos documentos. Uma métrica é associada entre uma expressão do domínio e cada documento. Um esquema da representação para uma coleção de n documentos é apresentado na Figura 4.

**FIGURA 4 - ESQUEMA DA REPRESENTAÇÃO DE COLEÇÃO DE DOCUMENTOS GBOED**

| | $k_1\_ck_{11}$ | ... | $k_1\_ck_{1j}$ | ... | $k_i\_ck_{11}$ | ... | $k_i\_ck_{1j}$ | ... | $k_1\_ck_{m1}$ | ... | $k_1\_ck_{ml}$ | ... | $k_i\_ck_{m1}$ | ... | $k_i\_ck_{ml}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $d_1$ | | | | | | | | | | | | | | | |
| $d_2$ | | | | | | | | | | | | | | | |
| ⋮ | | | | | | | | | | | | | | | |
| $d_n$ | | | | | | | | | | | | | | | |

*Fonte: Scheicher et al. (2016)*

No processo de construção das representações baseadas em expressões do domínio, os documentos são segmentados por sentença e as expressões são formadas pela união de todos os termos do domínio e todos identificadores de classe presentes em cada sentença. Não é considerada a remoção das *stopwords*, pois neste caso tais palavras podem ser consideradas como auxiliares na composição semântica das expressões.

Duas diferentes versões da representação gBoED são consideradas de acordo com a métrica associada a elas. A *gBoED_Freq* é a versão cuja métrica corresponde à frequência das expressões em cada documento e a *gBoED_Dist* é a versão da representação cuja métrica é formada pelo inverso da distância entre os termos de cada expressão do domínio. A distância é medida em quantidade de palavras e considerar o inverso desta medida significa que a expressão possui maior "peso" quanto mais próximos estão os termos que a compõem (Scheicher, Sinoara, Felinto, & Rezende, 2019).

Considerando o exemplo anterior, o documento D1: "Guga é o campeão do Tennis Masters Cup. Ele venceu Agassi por três sets a zero no jogo final" é composto por duas sentenças. Considerando, também, as classes "vitória de um atleta brasileiro" e "derrota de um atleta brasileiro", é possível considerar "Guga" e "Ele" como termo do domínio e os termos "campeão" e "venceu" como identificadores da classe "vitória de atleta brasileiro". Nas Figuras 5 e 6 são apresentadas as representações gBoED_Freq e gBoED_Dist para este documento do exemplo D1, com a métrica associada às expressões formadas. O digito "0" presente na expressão corresponde à indicação da classe "vitória de atleta brasileiro".

**FIGURA 5 – EXEMPLO DE gBoED_Freq**

| gBoED_Freq | |
|---|---|
| guga_0_campeão | ele_0_venceu |
| 1 | 1 |

*Fonte: Próprio autor*

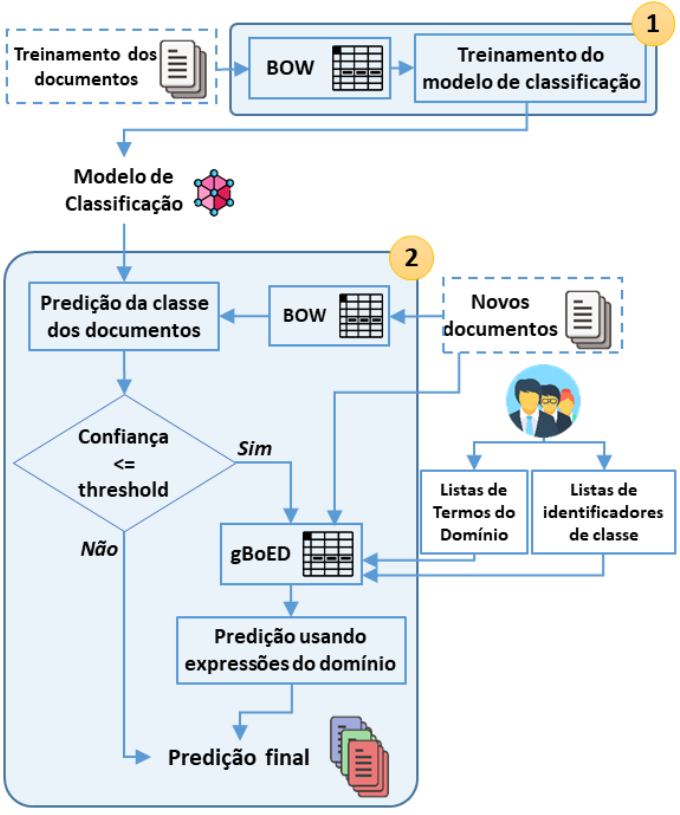
**FIGURA 6 – EXEMPLO DE gBoED_Dist**

| gBoED_Dist | |
|---|---|
| guga_0_campeão | ele_0_venceu |
| 0,5 | 1 |

*Fonte: Próprio autor*

Em Scheicher et al. (2019), os autores apresentam um método de classificação semanticamente enriquecida por expressões do domínio, que utiliza as representações gBoED como informações enriquecidas para melhorar os resultados em cenários de classificação de nível semântico. Na Figura 7 é apresentado um diagrama que corresponde ao fluxo do método composto por duas etapas principais.

**FIGURA 7 - DIAGRAMA DO MÉTODO DE CLASSIFICAÇÃO SEMANTICAMENTE ENRIQUECIDA**




*Fonte: Scheicher et al. (2019)*

A primeira etapa corresponde ao treinamento de um modelo de classificação baseado em uma representação tradicional da coleção de documentos do tipo BoW. Nesta etapa, qualquer algoritmo de classificação pode ser aplicado para obter um modelo de classificação. Na segunda etapa, o modelo gerado na etapa anterior é utilizado para predizer a classe do novo conjunto de documentos. As informações semanticamente enriquecidas da gBoED são aplicadas para melhorar os resultados de classificação para aqueles documentos cuja confiança de predição seja menor ou igual a um limiar global definido.

Neste trabalho, o método de classificação semanticamente enriquecida baseada na gBoED foi aplicado na classificação de pedidos de acesso à informação, conforme apresentado na próxima seção.

## MATERIAL E MÉTODO

No processo de validação dos pedidos de acesso à informação, cada pedido é analisado individualmente e verifica-se se a informação solicitada é passível de ser concedida ou, caso sejam informações sigilosas ou restritas de acesso, os pedidos são negados. De maneira geral, as informações são consideradas como passíveis de serem concedidas. Já nos casos de acesso negado, é possível exemplificar pelos pedidos de informações para acessar processos jurídicos que estão em segredo de justiça, pedidos com dados pessoais cuja privacidade deve ser protegida, como o pedido de acesso ao endereço de um servidor público, ou pedidos de acesso ao código-fonte de softwares internos, como da ferramenta "Análise de Licitações Públicas e Editais Públicos" (ALICE). Cada um desses casos possui fundamentações jurídicas para receberem negativa ao pedido de acesso à informação. Por exemplo, a motivação para a CGU não disponibilizar o código fonte do programa ao público é apoiada pela Lei 9.609 (1998), que prevê a proteção da propriedade intelectual desenvolvida em um programa de computador. Ainda de acordo com o parecer recente da CGU, contido no Processo 00075.000246/2018-4546[1], "tal apoio normativo está em consonância com o disposto no art. 22 da lei brasileira, que fundamenta a negativa do acesso ao ALICE pelo público em geral, uma vez que representa o procedimento executado durante o processo de planejamento de auditoria da CGU e do Tribunal de Contas da União, contendo toda a estratégia do trabalho de fiscalização a ser realizado, bem como as provas que servirão para comprovar ou não os fatos durante o trabalho de campo das equipes".

Neste trabalho, o método de classificação baseada na representação semanticamente enriquecida gBoED foi aplicado a uma base de dados de pedidos de acesso à informação, visando melhores resultados de classificação de nível semântico e a disponibilização de um modelo de classificação para apoio ao trabalho de análise inicial dos pedidos. Vale notar que este domínio de aplicação, a execução da LAI no âmbito da administração pública, em particular da CGU, é bastante diferente do domínio original, utilizado para o desenvolvimento da BoED. A base de dados originária contém os dados semiestruturados do "Relatório de pedidos de acesso à informação e solicitantes" no âmbito do Poder Executivo federal que estão disponíveis na plataforma FalaBr[3], com atualização dinâmica no Sistema Eletrônico de Serviço de Informação ao Cidadão[4] (e-SIC) e atualização periódica no Portal Brasileiro de Dados Abertos. Além disso, o código-fonte, as listas de termos e os resultados deste trabalho estão disponíveis em repositório na *Web*[5].

Inicialmente, a base foi composta por 3.617 solicitações de pedidos realizados à CGU no período de 2016 a 2020. Esse recorte temporal foi escolhido considerando a disponibilidade dos dados que continham o detalhamento dos pedidos, identificados na coluna 'DetalhamentoSolicitacao', que permitiram a realização de processo de MT. Em uma primeira exploração dos dados, a Figura 8 apresenta uma nuvem de palavras referente às palavras contidas na coluna de detalhamento dos pedidos à CGU, cujo tamanho de cada palavra está relacionado a frequência em que aparecem no conjunto de textos.

É possível identificar que as palavras mais frequentes são palavras como "município", "acesso", "informação", "boa tarde" ou "sobre". Tais palavras não permitem extrair muita informação para a construção de um classificador para a triagem dos pedidos.

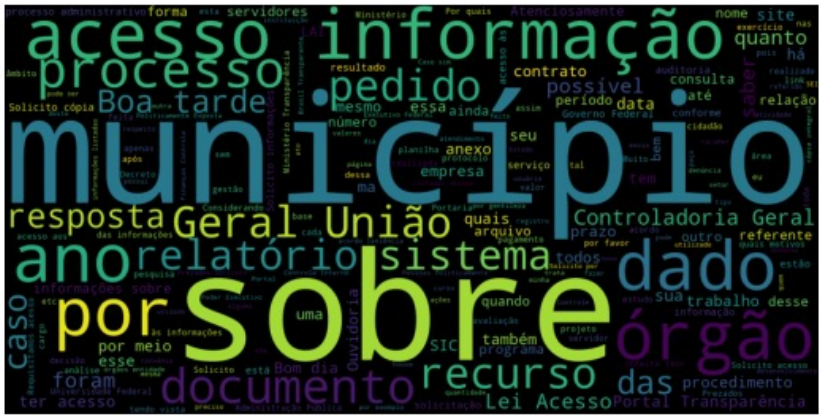
3. Controladoria-Geral da União. (2020). FalaBr: Disponível em: https://falabr.cgu.gov.br/

4. Controladoria-Geral da União. (2020). Sistema Eletrônico do Serviço de Informação ao cidadão. Disponível em: https://falabr.cgu.gov.br/publico/DownloadDados/DownloadDadosLai.aspx

5. O código fonte, as listas de termos e resultados podem ser acessados no endereço: https://github.com/ProjetoLAI/Classificador-dos-Pedidos-de-Acesso-Informa-o-na-CGU

Assim, palavras com frequência muito alta nos textos analisados tornam-se irrelevantes. Esse fato demonstra que a etapa de pré-processamento é muito importante para gerar uma boa representação dos dados.

**FIGURA 8 – NUVEM DE PALAVRAS DO DETALHAMENTO DOS PEDIDOS À CGU, DE 2012 A 2016**




*Fonte: Próprio autor.*

Após a análise dos dados originais, realizou-se um processo de limpeza e preparação da base, de modo a adequá-la tanto para o treinamento dos modelos usando BoW quanto das representações gBoED. Nesse processo foi realizada a remoção de informações irrelevantes para o objetivo deste trabalho, como números de protocolo, solicitações a outros órgãos e instituições públicas, prazos de solicitação. Também foram removidos dados ruidosos, como aqueles gerados devido a falhas durante a coleta ou inserção de dados, dados duplicados, pedidos genéricos, incompreensíveis ou que requeiram tratamento adicional. Após o processo de limpeza, o banco de dados passou a contar com 3.443 pedidos de acesso a informações e as seguintes colunas:

- IdPedido: Identificador único do pedido;
- DetalhamentoSolicitacao: Texto contendo o pedido. Atributo descritivo e não estruturado que foi utilizado como a base textual para construção do processo de classificação;
- TipoResposta: atributo que possui o tipo de resposta dada ao pedido (atributo-meta). É o atributo utilizado como rótulo para o treinamento do classificador. Na etapa de pré-processamento foram removidos da base os pedidos cujas tipologias de respostas foram: "informação inexistente", "não se trata de solicitação de informação", "órgão não tem competência para responder sobre o assunto", "pedido duplicado ou repetido" e "Acesso Parcialmente Concedido". Realizou-se então uma simplificação do modelo preditivo com a adoção das tipologias binárias que de fato importam inicialmente para o estudo de caso (atributos-meta ou classes a serem preditas). Com isso, as classes resultantes na base da pesquisa, após o pré-processamento, são "Acesso Concedido" e "Acesso Negado".

Outra característica da base de dados é a existência de um desbalanceamento entre as classes. Dos 3.443 pedidos resultantes da limpeza, 3.187 pedidos pertencem à classe "Acesso Concedido" e 256 pertencem à classe "Acesso Negado". Constatou-se que a proporção do desbalanceamento dessa classe é historicamente grande, o que pode ser explicado pela regra da transparência máxima prevista na LAI, que neste período estudado alcançou aproximadamente 93% dos acessos concedidos em relação à classe minoritária, com apenas 7% dos pedidos com acesso negado.

Com base no método da Figura 7, na etapa 1 é realizado o treinamento do modelo de classificação utilizando a representação tradicional BoW. Nesta etapa, foram construídos diversos modelos de classificação, a partir de diferentes algoritmos e variação de parâmetros. Foi utilizado também, a separação do conjunto de dados em 80% para treinamento (total

de 2.754 pedidos) e 20% para teste (total de 689 pedidos). Para diminuir o impacto do desbalanceamento foi aplicada a abordagem under-sampling no conjunto de treinamento, eliminando-se aleatoriamente exemplos da classe majoritária. Para o desenvolvimento da representação textual BoW, foi realizado o pré-processamento dos pedidos aplicando as técnicas a seguir:

- Padronização: visa à remoção do ruído para que o modelo possa detectar mais facilmente os padrões nos dados.
- Remoção de *stopwords* em português e de palavras com tamanho atípico, por exemplo, palavras que contêm menos de 2 letras ou mais de 10 letras.
- Radicalização: redução das palavras à sua forma raiz. Por exemplo, as palavras "chuva" e "chover" têm radicais semelhantes "chuv". Essa técnica permite padronizar as diferentes variações de uma palavra, reduzindo a dimensionalidade da representação.

Após o pré-processamento dos pedidos, foi gerada a representação BoW com métrica Tf–idf (*Term frequency–inverse document frequency*) com um total de 1.962 palavras ou características.

No domínio de aplicação deste trabalho, existe uma necessidade de transparência dos algoritmos, de modo a minimizar os riscos de distorção, de imparcialidades embutidas ou de erros dos resultados que porventura possam prejudicar um indivíduo em detrimento de outro. Portanto, a partir da representação BoW foram treinados múltiplos modelos de classificação com variação de diversos algoritmos e parâmetros descritos a seguir:

- **C4.5**: algoritmo de indução de árvores de decisão. Foi utilizado o valor 0,25 para parâmetro *confidence factor* e critérios para escolha do atributo: Entropia e Gini.
- **K-nearest neighbor (KNN)**: Os valores utilizados de k foram 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45, 55. O algoritmo foi executado com duas opções de medida de distância: Distância Euclideana e Distância de Cosseno.
- **Multinomial Naïve Bayes (MNB)**: algoritmo baseado em Naïve Bayes, com parâmetro alpha considerando os valores: $10^{-5}$, $10^{-4}$, $10^{-3}$, $10^{-2}$, 10-1 e 1.
- **Suport Vector Machine (SVM)**: algoritmo *Sequential Minimal Optimization (SMO)*. Nesse algoritmo foram consideramos três tipos de kernel: linear, polinomial (expoente=2) e *RBF (Radial Basis Function)*. Os valores considerados para cada tipo de kernel foram $10^{-4}$, $10^{-3}$, $10^{-2}$, $10^{-1}$, 1, 10, $10^{2}$, $10^{3}$, $10^{4}$.

Para fins de comparação de resultados quantitativos foi realizado também o treinamento de classificadores utilizando Redes Neurais Profundas, como as Redes Neurais Convolucionais – RNCs – que contêm cinco tipos de camadas: camada de entrada, de convolução, de agrupamento, as completamente conectadas e as de saída - e algoritmos de Redes Recorrentes – GRU e LSTM. Para o treinamento dos modelos gerados pelos algoritmos baseados em Redes Neurais, foi considerada a utilização do método de representação *Word Embedding/GloVe: Global Vectors for Word Representation*.

Após todos os modelos gerados, são escolhidos aqueles com melhores resultados de acurácia e medida F1 para serem utilizados na etapa de enriquecimento utilizando as expressões do domínio. A saída dos modelos é uma das classes preditas, "Acesso Concedido" ou "Acesso Negado", e um valor de confiança de predição. Se a confiança for superior a um limiar definido, a classe predita é considerada como a predição final para o pedido, caso contrário, a melhoria de classificação é realizada pelo método gBoED (etapa 2 da Figura 7).

Vale ressaltar a importância da utilização de algoritmos que permitam ao ser humano, seja parte da equipe da CGU, seja o requerente da informação, ou qualquer outro interessado, compreender o conteúdo das bases de dados, das ações e dos critérios adotados pelo classificador. No domínio de aplicação deste trabalho, existe uma necessidade de transparência dos algoritmos, de modo a minimizar os riscos de distorção, de imparcialidades embutidas ou de erros dos resultados que porventura possam prejudicar um indivíduo em detrimento de outro.

Essa análise comparativa do desempenho dos diferentes tipos de classificadores é importante visto que interpretabilidade ou explicabilidade versus desempenho dos modelos é um *trade-off* comum no aprendizado. Em geral, modelos mais complexos (como SVM e redes neurais) tendem a ter desempenho melhor do que os modelos mais transparentes (como as árvores de decisão – C4.5), no en-

tanto o maior desempenho vem acompanhado de menor explicabilidade (Gunning, 2017; Linardatos, Papastefanopoulos, & Kotsiantis, 2020).

Na etapa 2, expressões do domínio compõem a representação enriquecida gBoED. Elas são consideradas informações semanticamente enriquecidas indicadas pelo especialista da área do domínio, extraídas dos pedidos e utilizadas na melhoria dos resultados de classificação cuja confiança foi menor ou igual ao limiar definido. Conforme apresentado na Figura 4, uma expressão do domínio é composta pela união de um termo de domínio e um identificador de classe.

No contexto de pedidos de acesso a informação, os termos de domínio são palavras que indicam uma pergunta ou um pedido. Os identificadores de classe são termos importantes para uma classe específica. Por exemplo, para a classe "Acesso Concedido", tem-se os termos "salário" ou "remuneração" dos "servidores públicos". A divulgação nominal da remuneração no setor público brasileiro já foi objeto de grande controvérsia, envolvendo entendimentos conflitantes manifestados nas mais diversas decisões judiciais dos tribunais brasileiros, com o dilema ganhando contornos mais evidentes com a promulgação da Lei 12.527 (2011), e posterior regulamentação pelo Decreto 7.724 (2012).

Embora não haja referência explícita à divulgação de dados sobre funcionários públicos na lei brasileira, a jurisprudência é clara. Diante da controvérsia histórica, a decisão do Supremo Tribunal Federal sobre o processo 652.777/SP (recurso extraordinário com agravo) reconheceu a repercussão geral e estabeleceu que "é legítimo publicar, inclusive em um site eletrônico mantido pela Administração Pública, os nomes de seus funcionários e o valor de seus salários e as respectivas vantagens pecuniárias" ("ARE/652777", 2015)[6]. Sua decisão está em clara consonância com a regra de máxima transparência prevista na lei brasileira, com vistas à promoção do controle social (Martins, Lopes, & Cademartori, 2017).

Como exemplos de identificadores da classe "Acesso Negado" tem-se os termos "relatório de auditoria", "documentos de fiscalização" e "papel de trabalho", uma vez que essas informações estão, via de regra, relacionadas à previsão legal de sigilo, pois podem comprometer atividades de inteligência de auditoria, bem como investigação ou fiscalização em curso, relacionadas à prevenção e ao combate à corrupção.

Na Tabela 1, são apresentados exemplos de termos que compõem as listas utilizadas para a construção das expressões de domínio da gBoED com seus respectivos sinônimos. Após a seleção das expressões de domínio e construção das representações gBoED_Freq e gBoED_Dist, foi possível submetê-las ao método de classificação semanticamente enriquecida.

6. Supremo Tribunal Federal. (2015). Recurso Extraordinário com Agravo 652.777 São Paulo. Disponível em: https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=TP&docID=8831570

**TABELA 1 - EXEMPLOS DE TERMOS DAS LISTAS**

| TERMOS DE DOMÍNIO | IDENTIFICADORES DA CLASSE: ACESSO CONCEDIDO | IDENTIFICADORES DA CLASSE: ACESSO NEGADO |
|---|---|---|
| preciso de ajuda; preciso; solicito | Programa de Formação Continuada em Ouvidoria; PROFOCO | PAD; PADs; processo administrativo disciplinar; sindicância; investigação; em curso; em andamento |
| agradeceria que; agradeceria se | salário; salários; remuneração | relatório de auditoria; relatório final de auditoria; íntegra de relatório de auditoria; documentos de fiscalização; papéis de trabalho; papel de trabalho |
| disponibilizar | portal de transparência; site da transparência; grau de transparência; Programa Brasil Transparente; Escala Brasil Transparente; escala; ranking | extratos bancários |
| não encontrei | contrato; edital; licitação; gastos; orçamento | arquitetura detalhada do ALICE; código fonte; especificações produzidas para o software |
| desejo saber; desejo; desejo conhecer; desejo receber | Programa de Fortalecimento das Ouvidorias; PROFORT | nome do denunciante; identidade do denunciante; CPF |
| informar; gentileza informar; vistas; dei entrada; informa | Diário Oficial da União; D.O.U.; DOU; Boletim Interno | procedimentos de caráter preparatório; documentos preparatórios; preparatório; preliminares |

*Fonte: Próprio autor.*

Para avaliar os modelos gerados foram utilizadas as métricas de Acurácia e Medida-F1 *Weighted*. A acurácia indica uma performance geral do modelo, ou seja, mede o total de predições corretas do modelo. A Medida-F1 é uma média harmônica entre a precisão e revocação. A pontuação F1 Média Weighted (ou pontuação F1 *Weighted*) é calculada usando a média ponderada baseada no suporte de cada classe, levando em consideração o desbalanceamento das classes. Esse nível de avaliação proporcionado pela medida F1 Média *Weighted* é importante para este estudo, pois a classificação incorreta de um pedido como "Acesso Concedido" ou como "Acesso Negado", ou seja, para ambas as classes, pode implicar igualmente prejuízos para a administração pública e para o solicitante (custos desnecessários, frustração na obtenção do direito, perda da credibilidade institucional e, em última instância, pode inclusive implicar infrações administrativas e responsabilização administrativa e penal para a administração e para os servidores responsáveis). Por isso, o estudo utilizou-se da Medida-F1 *Weighted* e recomenda a revisão humana para reduzir a probabilidade de ocorrência de quaisquer riscos de classificações incorretas e eventuais prejuízos aos envolvidos.

## RESULTADOS E DISCUSSÃO

Considerando o processo de mineração a partir do modelo de classificação semanticamente enriquecida por expressões do domínio, o primeiro resultado obtido está relacionado à quantidade de pedidos de informação representados pelas expressões do domínio. Do total de 3443 pedidos obtidos após o processo de limpeza, foi possível representar 2847 pedidos (82,7% da base de dados) com a representação gBoED_Freq e 3039 pedidos (88,2% da base de dados) com a representação gBoED_Dist.

Outro resultado obtido está relacionado à representatividade e explicabilidade que as representações gBoED_Freq e gBoED_Dist trazem ao conjunto de dados. Para ilustrar este resultado, considere o seguinte pedido de informação e as suas representações BoW, gBoED_Freq e gBoED_Dist:

> ***Pedido**: Gostaria de saber a arquitetura detalhada do ALICE (Análise de Licitações e Editais), de modo a entender quais seus componentes e formas de integração), e pedir o código fonte que foi usado para construir a ALICE, com instruções para replicação.*

Neste pedido, após o pré-processamento do texto, são considerados os seguintes termos extraídos para formação da BoW: gost, sab, arquitet, detalh, alic, analis, edit, mod, entend, form, ped, codig, font, foi, usad, constru, alic. Portanto, na Figura 9 verifica-se a representação BoW para esse pedido.

Já as expressões do domínio podem ser consideradas como informações enriquecidas para a representação em relação ao texto original. A expressão é um resumo das informações principais contidas no pedido. Como se trata de um texto curto, nesse caso, apenas uma expressão já contribui para uma boa explicabilidade e representatividade do conteúdo principal do pedido.

**FIGURA 9 – BOW QUE REPRESENTA O PEDIDO DO EXEMPLO**

| gost | sab | arquitet | detalh | alic | analisis | edit | mod | entend | form | ped | cod | font | usad | constru |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

*Fonte: Próprio autor.*

Nas Figuras 10 e 11 são apresentadas, respectivamente, as representações gBoED_Freq e gBoED_Dist para um trecho do pedido de exemplo. O digito "1" existente no meio da expressão, indica que ela pertence à classe "Acesso Negado". A classe "Acesso Concedido" é representada pelo digito "0". Em gBoED_Dist a métrica 0,33 indica o inverso da distância em palavras entre "gostaria" e "arquitetura".

**FIGURA 10 – REPRESENTAÇÃO gBoED_Freq**

| gBoED_Freq |
|---|
| gostaria_1_arquitetura_detalhada_do_alice |
| 1 |

*Fonte: Próprio autor*

**FIGURA 11 – REPRESENTAÇÃO gBoED_Dist**

| gBoED_Dist |
|---|
| gostaria_1_arquitetura_detalhada_do_alice |
| 0,33 |

*Fonte: Próprio autor*

A Tabela 2 apresenta as melhores acurácias obtidas pelo método de classificação semanticamente enriquecida por expressões do domínio. Na coluna BoW, é apresentada a melhor acurácia obtida pelo melhor modelo de cada algoritmo gerado na etapa 1 do método, seguida pela Medida-F1 *Weighted* entre parênteses. Nas colunas gBoED_Freq e gBoED_Dist, são apresentadas as melhores acurácias obtidas após o enriquecimento semântico da etapa 2, seguidos pelas Medidas-F1 *Weighted* entre parênteses. O melhor resultado de cada linha é apresentado em negrito. De modo geral, a tabela apresenta os algoritmos de acordo com um nível decrescente de explicabilidade. Como é possível observar, dos algoritmos com menor explicabilidade o SVM obteve melhor desempenho, assim, decidiu-se que o método de classificação semanticamente enriquecida seria aplicado a este modelo e, também, aos modelos com maior explicabilidade.

**TABELA 2 - MELHORES ACURÁCIAS E MEDIDA-F1 *WEIGHTED* DOS CLASSIFICADORES**

| ALGORITMOS | BOW | GBOED_FREQ | GBOED_DIST |
|---|---|---|---|
| C4.5-entropy | 0,689 (0,763) | 0,708 (0,780) | 0,695 (0,770) |
| C4.5-gini | 0,709 (0,778) | 0,709 (0,780) | 0,709 (0,778) |
| KNN-cosine | 0,719 (0,786) | 0,775 (0,832) | 0,782 (0,835) |
| KNN-euclidean | 0,714 (0,781) | 0,765 (0,825) | 0,769 (0,826) |
| MNB | 0,715 (0,782) | 0,782 (0,836) | 0,782 (0,833) |
| SVM-linear | 0,804 (0,845) | 0,813 (0,857) | 0,815 (0,857) |
| SVM-rbf | 0,818 (0,856) | 0,833 (0,870) | 0,834 (0,870) |
| **SVM-poly** | **0,903 (0,907)** | **0,911 (0,917)** | **0,911 (0,917)** |
| GRU | 0,431 | - | - |
| LSTM | 0,431 | - | - |
| RNC | 0,705 | - | - |
| **MELHORES RESULTADOS** | **0,903 (0,907)** | **0,911 (0,917)** | **0,911 (0,917)** |

*Fonte: Próprio autor.*

Observando a Tabela 2, é possível verificar que a melhor acurácia obtida a partir do modelo gerado pela BoW foi utilizando o algoritmo SVM (os parâmetros utilizados no melhor caso foram kernel Polinomial e *Gamma*=10). Nesse experimento, obteve-se acurácia de 90,3% e Medida-F1 *Weighted* de 90,7%, o que significa equilíbrio nas métricas de precisão e revocação para as classes. É possível verificar também que houve melhoria dos resultados tanto utilizando a representação gBoED_Freq quanto gBoED_Dist. A acurácia para o enriquecimento com ambas as representações gBoED aumentou para 91,1% e Medida-F1 *Weighted* de 91,7%. O valor de corte referente à confiança do classificador foi de 60%, com 81 exemplos enviados para reclassificação. gBoED_Freq reclassificou 43 exemplos e gBoED_Dist 37 exemplos.

Como destaque, os resultados obtidos pelo modelo gerado a partir algoritmo SVM, com o kernel Polinomial e *Gamma*=10, e representação BoW superou inclusive os resultados globais das redes neurais profundas, com melhor acurácia para Rede Neural Convolucional (RNC), que atingiu 70,5% e Medida-F1 de 73%.

Outro destaque é o resultado do modelo gerado pelo algoritmo KNN, com o índice Cosine, que possui maior explicabilidade. Nele a acurácia passou de 71,9% usando BoW para 77,5%, usando gBoED_Freq, e chegou a 78,2 usando a gBoED_Dist. Nos outros algoritmos pode-se notar uma melhoria na acurácia e Medida-F1 *Weighted*, mostrando que as expressões do domínio podem contribuir com classificações de nível semântico. Tais resultados possuem aplicabilidade em relação aos dados reais, o que traz grande contribuição à sociedade e ao cumprimento dos direitos fundamentais relacionados ao acesso à informação.

## CONCLUSÕES

Em busca de uma solução para a crescente demanda de registros de pedidos de acesso à informação junto à CGU, este trabalho apresentou um conjunto de resultados aplicáveis para apoiar o processo de classificação dos pedidos de acesso à informação com apoio do método de classificação semanticamente enriquecida por expressões do domínio para melhoria dos resultados, da qualidade da representação e explicabilidade.

Na análise comparativa dos resultados de classificação dos pedidos de informação, o algoritmo SVM, com o kernel Polinomial e *Gamma*=10, alcançou a melhor acurácia no valor de 91,1% e Medida-F1 *Weighted* de 91,7% com o método enriquecimento pelas representações gBoED_Freq e gBoED_Dist.

Estes resultados, obtidos com enriquecimento semântico, superam os resultados de outros modelos que também foram desenvolvidos por mineração de textos e que depois de implantados na CGU já alcan-

çaram impactos positivos aos objetivos estratégicos institucionais, como o FARO, conforme inicialmente apresentado neste estudo. Recomenda-se a implantação do modelo não somente pelas métricas alcançadas, mas também porque os poucos casos que, porventura, o modelo vier a classificar incorretamente (cerca de 8,9%) poderão ser identificados por revisão humana após a triagem, além de existiram quatro instâncias recursais previstas no Poder Executivo Federal no Brasil.

Avalia-se, portanto, que os resultados demonstram grande potencial quanto ao uso deste modelo para classificação dos pedidos e dos recursos de acesso à informação na CGU, bem como em toda a administração pública em todos os Poderes, Executivo, Legislativo e Judiciário, sendo necessário adaptá-lo às bases de dados e ao pré-processamento com a representação das expressões de novos domínios. Como trabalhos futuros, entende-se a possibilidade de incorporar informações complementares existentes em outras bases de dados do órgão, por exemplo, informações contidas em sistemas de gestão de documentos e processos eletrônicos, nos quais é possível consultar se um processo é sigiloso ou se o processo apresenta alguma restrição de acesso. Outra possibilidade é integrar as informações da plataforma Fala.Br com o objetivo de contribuir ainda mais para a melhoria do modelo, tornando a implementação mais eficiente quando se trata do direito de acesso à informação no Brasil.

## REFERÊNCIAS

Carvalho, S. T. N. (2020). **Impacto da inteligência artificial na atividade de auditoria: equacionando gargalos nos repasses da união para entes subnacionais**. Dissertação (mestrado) – Escola Brasileira de Administração Pública e de Empresas, Centro de Formação Acadêmica e Pesquisa. 2020. 114 f.

**Constituição da República Federativa do Brasil de 1988**. (1998). Brasília.

Chun, S.; Shulman, S.; Sandoval, R.; Hovy (2010), **E. Government 2.0: Making connections between citizens, data and government**. Information Polity, IOS Press, v. 15, n. 1, 2, p. 1–9.

**Decreto 7.724** (2012). Regulamenta a Lei nº 12.527, de 18 de novembro de 2011, que dispõe sobre o acesso a informações previsto no inciso XXXIII do caput do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição.

García, S., Luengo, J., Herrera, F. (2015) **Data Preprocessing in Data Mining**. Intelligent Systems Reference Library, Vol. 72. Springer, Cham.

Gunning, D. (2017). **Explainable artificial intelligence (XAI)**. Tech. rep., Defense Advanced Research Projects Agency (DARPA).

**Lei n. 9.609, de 19 de fevereiro de 1998** (1998). Dispõe sobre a proteção da propriedade intelectual de programa de computador, sua comercialização no País, e dá outras providências.

**Lei n. 12.527, de 18 de novembro de 2011**. (2011). Lei de Acesso à Informação. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei n. 8.112, de 11 de dezembro de 1990; revoga a Lei n. 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. **Entropy**, Vol. 23, No. 1, p. 18.

Maia Filho, M. S., & Junquilho, T. A. (2018). Projeto Victor: perspectivas de aplicação da inteligência artificial ao direito. Revista De Direitos E Garantias Fundamentais, 19(3), 218-237. https://doi.org/10.18759/rdgf.v19i3.1587

Marques, C. A. N., Matsuno, I. P., Sinoara, R. A., Rezende, S. O. & Rozenfeld, H. (2015). An exploratory study to evaluate the practical application of pss methods and tools based on text mining. In: **Proceedings of the 20th International Conference on Engineering Design**.

Martins, A. C. M., Lopes, O. A., & Cademartori, S. U. (2017). **O STF e a divulgação nominalmente individualizada da remuneração dos servidores púbicos: uma análise do Recurso Extraordinário 652.777-SP**. Dissertação de Mestrado em Direito - Universidade de Brasília.

Mehr, H.; Ash, H.; Fellow, D. (2017) ***Artificial intelligence for citizen services and government***. Ash Cent. Democr. Gov. Innov. Harvard Kennedy Sch., no. August, p. 1–12.

Panis, A. da C. (2020). **Inovação em compras públicas: estudo de caso do robô Alice da Controladoria Geral da União (CGU)**. Universidade de Brasília, Dissertação de Mestrado, 116f.

Rezende, S. O., J. B. Pugliesi, E. A. Melanda, & M. F. Paula (2003). Mineração de dados. In S. O. Rezende (Ed.), **Sistemas Inteligentes – Fundamentos e Aplicações**. pp. 307–335. Editora Manole.

Rossi, R. G. (2015). **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. Tese de Doutorado em Ciências da Computação e Matemática Computacional - Instituto de Ciências Matemáticas e de Computação, Fundação de Amparo à Pesquisa do Estado de São Paulo.

Scheicher, R. B., Sinorara, R. A., Koga, N. J., & Rezende, S. O. (2016). Uso de expressões do domínio na classificação automática de documentos. In: **Anais do XIII Encontro Nacional de Inteligência Artificial e Computacional**, Vol. 1.

Scheicher, R.; Sinoara, R., Felinto, J., & Rezende, S. (2019). Sentiment classification improvement using semantically enriched information. In: **Proceedings of the ACM Symposium on Document Engineering 2019**.

Sinoara, R. A. (2018). **Aspectos semânticos na representação de textos para classificação automática**. Tese de Doutorado - Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo (USP). São Carlos, Brasil.

Tjandra, S.; Warsito, A. A. P.; Sugiono, J. P. (2015) Determining citizen complaints to the appropriate government departments using knn algorithm. In: **2015 13th International Conference on ICT and Knowledge Engineering (ICT Knowledge Engineering 2015)**. [S.l.: s.n.], p. 1–4. ISSN 2157-099X.

**Flávia Lemos Sampaio Xavier**
flavia.lemos.assessoria@gmail.com
ORCID: https://orcid.org/0000-0001-8575-1062

Controladoria-Geral da União (CGU).

Mestre em Ciência Política pelo Instituto Universitário de Pesquisas do Rio de Janeiro (Iuperj – 2015), especialista em Ciência de Dados, pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP - 2021), com certificação profissional em Ciência de Dados pela Universidade de Harvard (Harvard, 2020). Bacharel em Ciência Política, pela Universidade de Brasília (UnB, 2008). Desde 2017, pesquisa os temas de transparência, democracia, promoção da participação social e atua na Controladoria-Geral da União (CGU), no Observatório Social de Brasília (OSB) e no coletivo Pyladies.

**Ricardo Brigato Scheicher**
ricardoxem@gmail.com
ORCID: https://orcid.org/0000-0002-0588-0831
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (ICMC-USP).
Doutor em Ciências da Computação e Matemática Computacional pela Universidade de São Paulo (ICMC-USP). Atualmente é Líder Técnico e Cientista de Dados Sênior, especialista em linguística computacional na empresa Vitta Tecnologia em Saúde, empresa ligada ao grupo Stone Seguros. Mestre em Ciências da Computação pela Universidade Federal de São Carlos (UFSCar). Atua há dez anos com pesquisa e desenvolvimento de aplicações em Inteligência Artificial, Aprendizado de Máquina, Mineração de dados e textos.

**Roberta Akemi Sinoara**
roberta.sinoara@ifsp.edu.br
ORCID: https://orcid.org/0000-0001-8572-2747
Instituto Federal de São Paulo (IFSP), Campus Boituva
Doutora em Ciências da Computação e Matemática Computacional pela Universidade de São Paulo (ICMC-USP), na área de Inteligência Artificial, com estágio na Università degli Studi di Roma - La Sapienza, Roma, Itália. Atualmente é docente em regime de dedicação exclusiva no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Campus Boituva. Possui mais de quinze anos de experiência em pesquisas em Inteligência Artificial, com trabalhos em Ciência de Dados, Aprendizado de Máquina, e Mineração de Dados e Textos, atuando também em colaborações com grupos de pesquisa em diferentes áreas do conhecimento.