Predição de Relevância em Licitações Públicas com Modelos Híbridos Baseados em Texto e Metadados

Beatriz Pinheiro de Lemos Lopes , Ivo de Medeiros , Adaílton Ferreira de Araújo

[1] Instituto de Informática
Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

beatrizpinheiro@discente.ufg.br, ivopaixao@ufg.br, adailton@ufg.br

Abstract. *Manual screening of public tenders is costly due to the large document volume and the strong imbalance between relevant and irrelevant opportunities. This work proposes a hybrid approach for relevance prediction using Natural Language Processing and machine learning models applied to metadata and document text. The system uses BERTimbau when structured textual information is available and an alternative pipeline based on document embeddings and Random Forest otherwise. Experiments on real historical datasets containing thousands of annotated public tenders evaluated the system from a document ranking perspective. On average, the top 20% ranked documents contained approximately 75% of the relevant tenders, while the top 50 documents achieved 72% recall. These results demonstrate the potential of the proposed approach to support automatic screening in public procurement.*

Resumo. *A triagem manual de editais públicos é um processo custoso devido ao grande volume de documentos e ao forte desbalanceamento entre oportunidades relevantes e irrelevantes. Este trabalho propõe uma abordagem híbrida para predição de relevância utilizando técnicas de Processamento de Linguagem Natural e aprendizado de máquina aplicadas a metadados e ao texto dos documentos. O sistema utiliza o BERTimbau quando informações textuais estruturadas estão disponíveis e, caso contrário, um fluxo alternativo baseado em embeddings documentais e Random Forest. Experimentos realizados em bases históricas reais contendo milhares de editais públicos anotados avaliaram o sistema sob a perspectiva de ranqueamento de documentos. Em média, os 20% documentos mais bem posicionados concentraram aproximadamente 75% dos editais relevantes, enquanto os 50 primeiros documentos do ranking alcançaram 72% de recall. Esses resultados demonstram o potencial da abordagem proposta para apoiar a triagem automática em compras públicas.*

1. Introdução

O mercado de compras públicas brasileiro movimenta anualmente centenas de bilhões de reais por meio de processos licitatórios realizados por órgãos federais, estaduais e municipais [Ministério da Economia 2026]. A digitalização desses processos ampliou significativamente o número de editais publicados diariamente em plataformas governamentais e agregadores especializados, exigindo que empresas monitorem continuamente um

grande volume de documentos para identificar oportunidades alinhadas ao seu portfólio e estratégia comercial. Apesar desse contexto, a análise de editais permanece fortemente dependente de triagem manual, realizada por equipes que precisam ler dezenas ou centenas de documentos por dia em um cenário de elevada heterogeneidade textual, ausência de padronização documental e complexidade técnica, jurídica e administrativa, frequentemente distribuída em centenas de páginas e múltiplos anexos.

Nesse ambiente, apenas uma pequena parcela dos editais publicados costuma ser realmente relevante para uma empresa específica, o que torna a priorização automática das oportunidades tão importante quanto a própria identificação de relevância. O problema deixa de ser apenas decidir se um edital é potencialmente interessante e passa a incluir a ordenação dos documentos de forma a concentrar, nas primeiras posições, aqueles com maior probabilidade de aderência técnica e comercial. Abordagens baseadas em ranqueamento tornam-se, assim, particularmente adequadas para direcionar o esforço operacional ao subconjunto de editais mais promissor, reduzindo o volume de documentos que demandam leitura detalhada pelas equipes.

Do ponto de vista computacional, a automação desse processo impõe desafios relevantes às áreas de Processamento de Linguagem Natural (PLN), Recuperação de Informação e Aprendizado de Máquina. Documentos licitatórios apresentam elevada variabilidade semântica, inconsistências estruturais, redundância textual e uso recorrente de terminologias ambíguas ou semanticamente equivalentes para descrever o mesmo objeto, o que limita abordagens puramente lexicais. Além disso, informações relevantes podem estar dispersas ao longo de diferentes seções e anexos, exigindo modelos capazes de capturar contexto semântico em documentos extensos e heterogêneos.

Nos últimos anos, modelos baseados em transformers têm obtido resultados expressivos em tarefas de classificação e representação semântica, em especial variantes pré-treinadas para o português, como o BERTimbau [Souza et al. 2020], amplamente utilizadas em domínios jurídico-administrativos. Paralelamente, métodos clássicos de aprendizado de máquina continuam úteis em cenários com bases fortemente desbalanceadas, como o presente estudo, no qual apenas aproximadamente 3,34% dos editais foram classificados como relevantes. Além disso, restrições computacionais e ausência de metadados estruturados tornam arquiteturas híbridas combinando embeddings semânticos, modelos neurais e algoritmos tradicionais alternativas promissoras para triagem documental em larga escala. Inserido nesse contexto, este trabalho propõe uma abordagem híbrida para predição de score contínuo de relevância de editais públicos, baseada em modelos de PLN e aprendizado de máquina aplicados sobre metadados e conteúdo textual, cuja arquitetura de processamento em lote e fluxos de inferência é detalhada na próxima seção.

2. Metodologia

Nesta seção, descrevemos a arquitetura de processamento em lote adotada e os modelos utilizados para o cálculo do score de relevância dos editais a partir de múltiplas fontes de dados, vide Figura 1, em linha com a abordagem híbrida apresentada na seção 1.

2.1. Arquitetura Geral

Os documentos e seus metadados são inicialmente ingeridos dos fornecedores RHS [RHS Licitações 2026] e Conlicitação [Conlicitação 2026] e armazenados em um serviço

de armazenamento em nuvem, que atua como repositório central do pipeline. Em seguida, um serviço gerenciado de processamento em batch é acionado assim que novos documentos se tornam disponíveis nesse repositório, executando a solução de cálculo de relevância. Essa solução aciona, em primeira instância, um modelo principal baseado em metadados e, em fluxo alternativo, um modelo baseado em embeddings documentais, responsáveis por estimar a probabilidade de relevância de cada edital, posteriormente convertida em um score contínuo na escala de 1 a 100 consumido pelos serviços de ranking e priorização operacional.

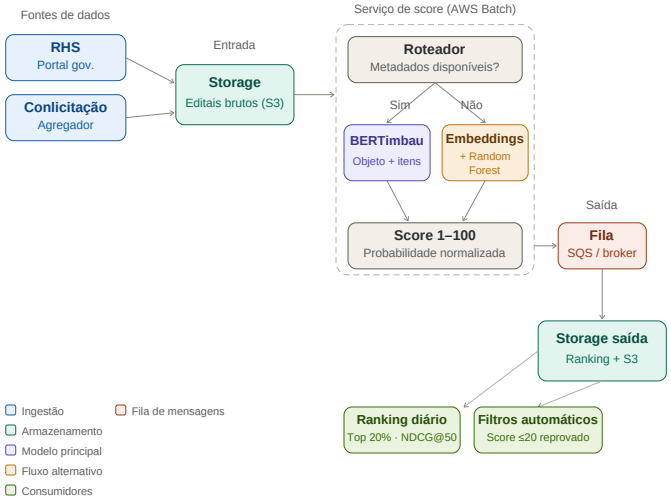

Figura 1. Pipeline híbrido de predição de relevância de editais públicos.

2.2. Modelo principal baseado em metadados

No fluxo principal de inferência, o sistema utiliza metadados textuais provenientes dos fornecedores RHS [RHS Licitações 2026] e Conlicitação [Conlicitação 2026] como entrada para o modelo de relevância. Entre os principais campos utilizados estão o objeto do edital e a lista de itens licitados, que descrevem textualmente os produtos, serviços ou soluções demandadas no processo licitatório. O campo objeto geralmente contém uma descrição resumida da contratação, enquanto o campo itens apresenta detalhamentos individuais dos produtos ou serviços requisitados. Os textos passam por uma etapa de pré-processamento envolvendo remoção de caracteres especiais, normalização de espaços, remoção de acentuação, transformação para minúsculas e remoção de stopwords, resultando em uma sequência textual única para cada edital. Essa sequência é utilizada no fine-tuning do modelo BERTimbau [Souza et al. 2020], formulado como uma tarefa de classificação binária entre editais relevantes e irrelevantes.

Devido ao forte desbalanceamento entre classes, a função de perda adotada foi a *Weighted Cross-Entropy Loss*, na qual os pesos das classes são definidos inversamente proporcionais à frequência de ocorrência de cada classe no conjunto de treinamento, utilizando a estratégia balanced da biblioteca *Scikit-Learn* [Pedregosa et al. 2011]. A função de perda ponderada busca reduzir o viés do modelo em direção à classe majoritária, penalizando mais fortemente erros sobre a classe minoritária. A utilização de pesos na função

de perda busca reduzir o viés do modelo em direção à classe majoritária, penalizando mais fortemente erros cometidos sobre a classe minoritária.

2.3. Fluxo alternativo com embeddings documentais

Em cenários nos quais os metadados textuais estão ausentes ou insuficientes, o sistema utiliza um mecanismo alternativo baseado no conteúdo textual dos documentos. Nesse fluxo, os documentos passam inicialmente por um classificador de páginas similar ao usado no trabalho de [Silva et al. 2024] responsável por identificar páginas potencialmente relevantes. As páginas classificadas como relevantes são convertidas em embeddings semânticas utilizadas como representação vetorial do edital. A partir dessas representações, modelos clássicos de aprendizado de máquina são utilizados para inferir o score de relevância do documento. Essa abordagem permite maior robustez em cenários nos quais os metadados estruturados não estão disponíveis ou apresentam baixa qualidade.

2.4. Estratégia de ranqueamento e saída operacional

O sistema utiliza um score contínuo de relevância para ordenar automaticamente os editais em rankings diários consumidos pelos serviços de priorização e filtragem apresentados na Figura 1. As probabilidades previstas para a classe positiva pelos modelos do fluxo principal e do fluxo alternativo são convertidas linearmente para uma escala de 1 a 100, em que valores mais altos indicam maior probabilidade de o edital ser considerado relevante pelas equipes operacionais. Essa normalização facilita a interpretação do score por diferentes sistemas consumidores e padroniza os critérios de ordenação, permitindo que o pipeline híbrido produza rankings consistentes a partir de fontes e modelos distintos.

3. Experimentos

Esta seção apresenta os experimentos realizados para avaliar o sistema proposto para predição de relevância de editais públicos. A avaliação foi dividida em duas etapas complementares:

- Avaliação inicial dos modelos de classificação de relevância;
- Avaliação operacional posterior focada especificamente no desempenho do sistema de ranqueamento.

O objetivo principal dos experimentos foi analisar a capacidade do sistema em priorizar automaticamente editais relevantes em cenários reais de triagem documental.

3.1. Base de Dados

O conjunto principal utilizado no treinamento do modelo baseado em metadados foi composto por 6.525 editais coletados entre agosto e outubro de 2024. Os documentos foram previamente anotados utilizando informações de aprovação e reprovação presentes no fluxo operacional da empresa. A distribuição das classes apresentou forte desbalanceamento, contendo 218 editais relevantes e 6.307 editais irrelevantes. Os campos utilizados como entrada do modelo principal foram objeto do edital e itens licitados.

Para o fluxo alternativo baseado em embeddings documentais, foi construída uma segunda base contendo 3.226 documentos balanceados entre editais aprovados e reprovados. Nesse cenário, as representações vetoriais foram geradas a partir de páginas previamente classificadas como relevantes.

Posteriormente, foi realizada uma nova avaliação operacional utilizando dados reais de produção referentes ao período de março de 2026, com foco específico na análise do comportamento do sistema de ranqueamento diário.

3.2. Configuração Experimental

Na primeira etapa experimental foram avaliados o modelo principal baseado em BERTimbau [Souza et al. 2020], o fluxo alternativo baseado em embeddings documentais e diferentes algoritmos clássicos de aprendizado de máquina. Essa avaliação inicial considerou as métricas tradicionais de classificação de precisão, recall e F1-score.

No treinamento do modelo principal baseado em BERTimbau foram utilizados os seguintes hiperparâmetros: taxa de aprendizado (*learning rate*) de 2×10^{-5}, *weight decay* de 0.01, tamanho de lote (*batch size*) igual a 8 e 3 épocas de treinamento. A inferência foi realizada utilizando a estratégia de *argmax*, considerando a classe positiva como editais aprovados.

Para o fluxo alternativo baseado em embeddings documentais foram avaliados os algoritmos Random Forest, Gradient Boosting e Support Vector Machine (SVM), utilizando validação cruzada e diferentes combinações de hiperparâmetros. Entre os modelos avaliados, o Random Forest apresentou o melhor desempenho geral, sendo selecionado como modelo final do fluxo alternativo.

Já a segunda etapa experimental concentrou-se no desempenho do sistema de ranqueamento em ambiente operacional real. Nesse cenário, os editais foram agrupados por dia de captura e ordenados de acordo com o score de relevância produzido pelo sistema. Foram calculadas métricas de recuperação considerando diferentes cortes do ranking diário, incluindo Recall@Top-20%, Recall@50, Precision@50 e NDCG@50, com foco na capacidade do sistema em priorizar automaticamente editais relevantes para análise manual.

4. Resultados

4.1. Resultados do Modelo de Relevância

Na tabela 1 apresentamos os resultados para os experimentos com o modelo baseado em BERTimbau (modelo principal) e para o modelo alternativo, que usa Random Forest.

A Figura 2 evidencia que os editais reprovados apresentam forte concentração nas faixas de *score* mais baixas: 85,2% já estão acumulados na primeira decila (01–09), com apenas 10,6% atingindo a faixa mais alta (90–99). Os editais aprovados exibem comportamento oposto: 70,6% concentram-se na decila superior (90–99), enquanto apenas 29,4% distribuem-se pelas faixas inferiores (01–89). Esse comportamento divergente entre as duas curvas indica que o modelo de *scoring* apresenta elevada capacidade discriminativa, capturando de forma eficaz os editais reprovados nas faixas baixas e os aprovados nas faixas altas.

4.2. Resultados de Ranqueamento

Além da avaliação de classificação, foi realizada uma análise do desempenho do sistema sob a perspectiva de ranqueamento operacional. Os editais foram agrupados diariamente e ordenados conforme o score contínuo de relevância produzido pelo pipeline híbrido.

Tabela 1. Métricas por modelo.

Modelo	Revocação	Precisão	F1-Macro
Modelo Principal	1,00	0,20	0,63
Modelo Alternativo	0,816	0,875	0,849

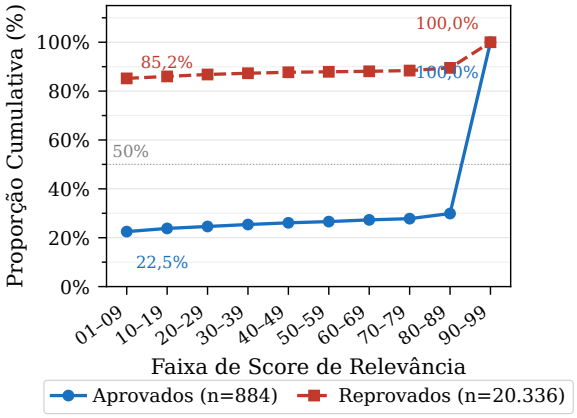

Figura 2. Proporção Cumulativa Editais Aprovado e Reprovados.

A Tabela 2 apresenta os resultados médios obtidos com dados reais de produção referentes a março de 2026. Em média, os 20% primeiros documentos do ranking concentraram 75,47% dos editais relevantes do dia. Considerando apenas os 50 primeiros documentos classificados, o sistema atingiu 71,61% de recall e NDCG@50 igual a 0,4038, indicando boa capacidade de priorização de documentos relevantes nas primeiras posições do ranking.

Tabela 2. Resultados médios das métricas de ranqueamento.

Recall@Top-20%	Recall@50	Precision@50	NDCG@50
0,7547	0,7161	0,1409	0,4038

5. Conclusão

Este trabalho apresentou uma abordagem híbrida para predição de score de relevância em editais públicos utilizando modelos de PLN e aprendizado de máquina aplicados sobre metadados e conteúdo textual dos documentos. Os resultados demonstraram que o modelo baseado em BERTimbau obteve elevado recall na identificação de editais relevantes, sendo responsável pela análise da maior parte dos casos operacionais (aproximadamente 85,8% dos editais). Já o fluxo alternativo baseado em embeddings documentais e Random Forest, utilizado em cerca de 14,2% dos casos, apresentou desempenho satisfatório em cenários sem metadados estruturados.

A avaliação sob a perspectiva de ranqueamento mostrou que o sistema foi capaz de concentrar a maior parte dos editais relevantes nas primeiras posições do ranking diário, reduzindo significativamente o volume de documentos necessários para análise manual. Além disso, a distribuição dos scores evidenciou boa capacidade discriminativa entre editais aprovados e reprovados.

Os resultados obtidos demonstram o potencial de arquiteturas híbridas para auxiliar processos automáticos de triagem documental em compras públicas. Como trabalhos futuros, pretende-se investigar técnicas de aprendizado de ranking e abordagens multimodais para exploração de informações estruturais presentes nos documentos licitatórios.

Referências

Conlicitação (2026). Conlicitação. `https://conlicitacao.com.br/`. Acessado em: 11 maio 2026.

Ministério da Economia (2026). Painel de compras do governo federal - relatório anual. Portal de Transparência. Acesso em: 08 maio 2026.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

RHS Licitações (2026). Rhs licitações. `https://licitacao.com.br/`. Acessado em: 11 maio 2026.

Silva, E., Medeiros, I., Menezes, M., and Kamikawachi, D. (2024). Segmentation and summarization for extracting information about information technology equipment from government procurement notice. In *Anais do XII Symposium on Knowledge Discovery, Mining and Learning*, pages 145–152, Porto Alegre, RS, Brasil. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.