ComprasMatch: Um Modelo Baseado em Similaridade Semântica e Grafos para Detecção de Sinergias em Contratações Públicas

Guilherme Benevides[1], Tiago Lobo[1], Gabriel Figueirêdo[2], Matheus Almeida[1], Francisco Pereira[1], Bruno Figueirêdo[1]

[1] Curso de Bacharelado em Ciência da Computação
Universidade Estadual de Roraima (UERR) – Boa Vista, RR – Brasil

[2]Curso de Bacharelado em Ciência de Dados e Inteligência Artificial
Universidade Federal da Paraíba (UFPB) – João Pessoa, PB – Brasil

{guilherme.benevides, matheus.almeida}@alunos.uerr.edu.br,

{tiago.lobo, bruno_cbf, fpcarlos}@uerr.edu.br,

gabriel.figueiredo@academico.ufpb.br

Abstract. *This paper presents ComprasMatch, a semantic similarity and graph-based model designed to identify synergies in public procurement through analysis of textual item descriptions. The architecture integrates PNCP data, embedding generation (Gemma 3), and a similarity graph to support shared procurement grouping. Evaluation employed a controlled synthetic baseline and four metrics—accuracy, precision, recall, and F_1-score—with threshold analysis. The model achieved 96.71% precision and a 90.28% F_1-score at $\tau = 0.7$. Comparison with GPT-5 mini, DeepSeek-Chat, and Llama 3 demonstrated competitive performance in terms of F_1-score. Results indicate technical feasibility and potential application in supporting procurement planning.*

Resumo. *Este artigo apresenta o ComprasMatch, modelo baseado em similaridade semântica e grafos para identificação de sinergias em contratações públicas por meio da análise de descrições textuais. A arquitetura integra dados do PNCP, vetorização (Gemma 3) e grafo de similaridade para formação de compras compartilhadas. A avaliação utilizou baseline sintética e quatro métricas — acurácia, precisão, recall e F_1-score — com análise de limiar. O modelo alcançou precisão de 96,71% e F_1-score de 90,28% no limiar $\tau = 0{,}7$. A comparação com GPT-5 mini, DeepSeek-Chat e Llama 3 evidenciou desempenho competitivo em termos de F_1-score. Os resultados indicam viabilidade técnica e potencial aplicação no apoio ao planejamento governamental.*

1. Introdução

O Estado brasileiro atua como o principal indutor da economia nacional, movimentando um mercado de compras públicas que representa aproximadamente 16% do Produto Interno Bruto (PIB) do país [Ministério da Gestão e da Inovação em Serviços Públicos 2025]. A eficiência dessas aquisições impacta diretamente a alocação de recursos públicos. No entanto, o

país enfrenta um desafio crônico na qualidade do gasto. Estimativas apontam que a ineficiência nas despesas governamentais gera uma perda estrutural conservadora de 3,9% do PIB, o que equivale a um desperdício anual na ordem de US$ 68 bilhões [Banco Interamericano de Desenvolvimento 2018].

Buscando mitigar essas ineficiências, a Lei nº 14.133, de 1º de abril de 2021 (Nova Lei de Licitações e Contratos Administrativos), estabeleceu o compartilhamento de aquisições como diretriz da Administração Pública. A literatura evidencia que a agregação de demandas promove um duplo ganho: gera economia de escala com a redução direta no preço do bem e reduz drasticamente os custos de transação processuais [Góis and Mendonça 2023]. Casos práticos documentados validam essa premissa no âmbito subnacional, a exemplo das compras consorciadas do Conisul, em Alagoas, que registraram quedas de até 90% no valor unitário de medicamentos adquiridos em conjunto [Mendes 2023].

Apesar do robusto amparo legal e dos esforços federais recentes, a operacionalização das compras compartilhadas esbarra em severas barreiras operacionais. O agente público atua frequentemente em um cenário de "solidão", elaborando Planos Anuais de Contratações (PAC) e Estudos Técnicos Preliminares (ETP) de forma descentralizada, sem visibilidade sobre demandas análogas de outros órgãos. Este isolamento é agravado pela precariedade estrutural dos dados abertos governamentais; um levantamento recente do Tribunal de Contas da União (TCU) revelou que 86% dos registros de contratações no Portal Nacional de Contratações Públicas (PNCP) apresentam algum tipo de inconsistência ou falha de preenchimento [Ministério da Gestão e da Inovação em Serviços Públicos 2025]. Diante dessa assimetria informacional e do massivo volume de documentos redigidos sem padronização semântica, torna-se humanamente impossível a identificação proativa de sinergias de compra.

Nesse cenário, técnicas de Inteligência Artificial, especificamente o Processamento de Linguagem Natural (PLN) e os Modelos de Linguagem de Larga Escala (LLMs), despontam como ferramentas promissoras para a superação deste gargalo. O presente artigo propõe o desenvolvimento e a validação do *ComprasMatch*, um sistema de inteligência preditiva focado na fase de planejamento pré-contratual. A aplicação utiliza o estado da arte em IA, baseando-se no cálculo de similaridade semântica do modelo Gemma 3 [Google DeepMind 2025], capaz de interpretar intenções de compra subjacentes a descrições textuais distintas e superar o desafio da taxonomia inconsistente nos planos de contratação.

Ao processar os dados abertos, a ferramenta gera um Grafo de Similaridade permitindo a formação de grupos de compras antes mesmo da publicação dos editais. Dessa forma, o objetivo central desta pesquisa é demonstrar como a adoção de tecnologias de similaridade semântica pode transformar dados burocráticos imprecisos em inteligência acionável para o fomento de compras compartilhadas, garantindo eficiência de escala e a efetiva modernização da máquina pública.

O artigo está estruturado da seguinte forma: a Seção 2 apresenta o referencial teórico; a Seção 3 descreve a arquitetura do ComprasMatch; a Seção 4 detalha a metodologia experimental; a Seção 5 discute os resultados e limitações; e a Seção 6 apresenta as

conclusões e direções futuras.

2. Referencial Teórico

Compras públicas baseadas em sistemas eletrônicos ampliam o volume de dados disponíveis, mas preservam um obstáculo recorrente: órgãos descrevem o objeto contratado em texto livre e com baixa padronização. Variações como abreviações, sinônimos e unidades inconsistentes reduzem a comparabilidade entre registros e dificultam identificar produtos equivalentes. A literatura organiza esse cenário em duas tarefas principais: (i) **classificação** de itens em taxonomias padronizadas (p. ex., CPV/UNSPSC) e (ii) **pareamento/agrupamento** de itens semanticamente equivalentes (*item matching*/*clustering*), que sustenta consolidação de demandas e redução de fragmentação.

Na classificação, estudos recentes tratam um espaço de rótulos amplo, hierárquico e desbalanceado. Trabalhos em CPV exploram modelos profundos com classificação *multilabel* e estratégias *zero-shot* para classes raras ou ausentes [Moiraghi et al. 2024]. Em cenários multilíngues, abordagens com *Transformers* buscam reduzir dependência de dados rotulados e favorecer transferência entre idiomas [Bednar et al. 2025]. Em UNSPSC e classificação de gasto, pesquisas indicam que baselines tradicionais, como TF–IDF com classificadores supervisionados enquanto técnicas que exploram explicitamente a hierarquia melhoram consistência e desempenho [Abdullahi et al. 2024].

No pareamento, a área formaliza a tarefa como *entity matching*: o sistema decide se dois registros descrevem a mesma entidade. Abordagens léxicas baseadas em normalização e similaridade textual (p. ex., TF–IDF + cosseno) servem como *baselines*, mas perdem qualidade quando equivalência depende de sinonímia, reordenação de termos e atributos implícitos [Mudgal et al. 2018]. Modelos baseados em linguagem pré-treinada tendem a capturar similaridade semântica e reduzir dependência de engenharia manual de atributos [Li et al. 2021]. Outros trabalhos aplicados indicam que inconsistências no texto distorcem análises e dificultam detecção de padrões em bases abertas [Vale et al. 2023]. No contexto brasileiro, pesquisas também reportam estratégias para identificar similaridades em descrições de objetos contratados com uso de aprendizado profundo [Lima et al. 2025].

Para viabilizar busca e agrupamento em grande escala, a literatura adota *embeddings* contextualizados e comparação vetorial por similaridade do cosseno, o que permite pré-computar representações e acelerar a geração de candidatos [Reimers and Gurevych 2019]. Quando o objetivo exige maior precisão, a área combina uma arquitetura em duas etapas: o sistema gera candidatos com *embeddings* e aplica um modelo de pares para reclassificar os casos mais promissores [Li et al. 2021].

Para tornar o resultado interpretável e útil à gestão, trabalhos modelam relações como um grafo de similaridade: nós representam itens, arestas representam pares acima de um limiar e grupos emergem como componentes ou comunidades [Brum et al. 2024]. Essa modelagem facilita validação, pois a equipe inspeciona grupos em vez de pares isolados e identifica padrões de demanda em múltiplos órgãos. O limiar controla o *trade-off* entre precisão e *recall*; um limiar baixo aumenta cobertura e falsos positivos, enquanto um limiar alto reduz falsos positivos e eleva falsos negativos [Appel et al. 2022]. Bases grandes intensificam esse desafio, pois geram desbalanceamento extremo entre pares equivalentes e não equivalentes; por isso, a literatura recomenda protocolos repro-

dutíveis, métricas adequadas e validação humana contínua [Brum et al. 2024]. Mudanças em catálogos e padrões de redação alteram a distribuição dos dados ao longo do tempo, exigindo auditoria e governança analítica para sustentar desempenho [Guida et al. 2025].

O presente trabalho adota *embeddings* semânticos para representar descrições de objetos contratados e modela suas relações em um grafo de similaridade, permitindo identificar agrupamentos potenciais de demanda entre diferentes órgãos públicos.

3. O ComprasMatch

O ComprasMatch consiste em uma solução tecnológica para otimizar a gestão de compras públicas por meio da identificação de oportunidades de compras compartilhadas. A ferramenta analisa Planos de Contratações Anuais (PCA) de diferentes órgãos e detecta intenções de aquisição de itens semanticamente similares, subsidiando processos licitatórios conjuntos.

Para isso, o sistema centraliza e cruza dados do Portal Nacional de Contratações Públicas (PNCP), reduzindo a fragmentação de processos e favorecendo ganhos de escala nas aquisições governamentais. A arquitetura do ComprasMatch combina técnicas de processamento de linguagem natural, representações vetoriais semânticas e grafos de similaridade para identificar relações entre descrições textuais de itens.

A Figura 1 apresenta uma visão geral das etapas que compõem o funcionamento do sistema.

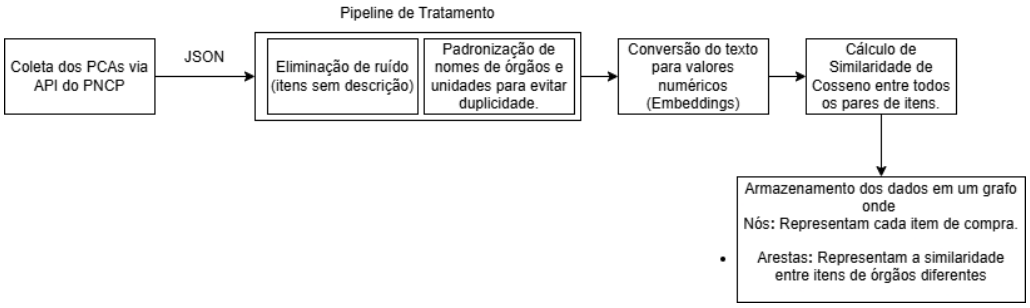

Figura 1. Representação das etapas do ComprasMatch

4. Metodologia

Esta seção descreve os procedimentos metodológicos adotados no desenvolvimento e avaliação do ComprasMatch. Inicialmente, são apresentadas as etapas do pipeline de processamento, incluindo extração, tratamento e vetorização semântica dos dados provenientes do PNCP, bem como a construção do grafo de similaridade utilizado para identificar oportunidades de compras compartilhadas. Em seguida, são descritos os critérios de construção da *baseline*, os procedimentos de comparação entre representações vetoriais e as métricas utilizadas para avaliação do modelo. O objetivo é garantir um protocolo reprodutível, capaz de analisar de forma objetiva a capacidade do sistema em identificar similaridades semanticamente relevantes.

4.1. Pipeline de pré-processamento e vetorização de dados

O pré-processamento garante integridade, comparabilidade e relevância estatística dos dados extraídos do PNCP, conforme o segundo bloco da Figura 1. A estratégia converte

dados brutos obtidos via *Application Programming Interface* (API) em um conjunto estruturado, limpo e semanticamente consistente para as etapas de análise.

4.1.1. Extração, filtragem e consistência de dados

O sistema consome dados em formato JSON diretamente da API do PNCP (Figura 1) em lotes diários. Em seguida, aplica um filtro de existência e validade textual: o pipeline remove automaticamente registros com `descricao_item` ausente, vazia ou contendo ruído textual (*e.g.*, caracteres não-alfanuméricos excessivos). Foram descartados itens com valor nulo ou inválido, e aqueles pertencentes a categorias de serviços de qualquer natureza. Esta última decisão foi tomada visando não criar barreiras iniciais em compras compartilhadas entre estados diferentes.

4.2. Tratamento de atributos e lógica de contextualização

A segunda etapa normaliza metadados para viabilizar comparações e reduzir redundâncias dentro da mesma origem. A normalização inclui a conversão para *lowercase* e a remoção de pontuações e *stopwords* irrelevantes na descrição do item. Durante a iteração, o sistema extrai e mapeia atributos de unidade e procedência: (i) define a unidade a partir de `nomeUnidade` e, na ausência desse campo, utiliza `unidade_fornecimento` como campo alternativo; (ii) registra a origem administrativa associada à descrição do item. Este tratamento é crucial para garantir que as variações gramaticais e de formatação não impactem negativamente a similaridade semântica subsequente.

4.3. Vetorização semântica e geração de *embeddings*

Após o tratamento (Figura 1), o sistema representa cada descrição como vetor denso em um espaço multidimensional por meio de modelos de linguagem pré-treinados [Google DeepMind 2025]. Especificamente, é utilizada a variante Gemma 3.0 Large configurada para gerar vetores com uma dimensionalidade fixa de $d = 768$ (maior dimensionalidade disponível). O codificador transforma a semântica do texto em uma representação numérica, de modo que a proximidade geométrica entre vetores reflita similaridade de significado entre descrições, mesmo diante de variações de escrita e sinonímia.

4.4. Construção do Grafo de Similaridades

Com os itens devidamente vetorizados, o sistema implementa uma arquitetura de grafos de similaridade para mapear as conexões entre itens de diferentes órgãos, conforme ilustrado no quinto bloco da Figura 1. Nesse modelo, o sistema define um grafo $G = (V, E)$, onde cada item de compra corresponde a um nó $v \in V$, associado a um vetor $\mathbf{x}_v \in R^d$ (*embedding*) e a um identificador de órgão $o(v)$.

As arestas entre esses nós são estabelecidas segundo dois critérios fundamentais:

- **Alteridade de Origem:** somente são criadas conexões entre itens pertencentes a órgãos distintos, concentrando-se exclusivamente na colaboração interinstitucional. Formalmente, para dois nós $u, v \in V$, o sistema exige $o(u) \neq o(v)$.
- **Métrica de Cosseno:** a conexão é determinada pelo cálculo da similaridade de cosseno entre os vetores. O sistema calcula

$$\mathrm{s}(u, v) = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \tag{1}$$

e cria uma aresta $(u, v) \in E$ apenas quando $s(u, v) \geq \tau$, com $\tau = 0{,}70$. Assim, o sistema define:

$$(u, v) \in E \iff \big(o(u) \neq o(v)\big) \ \wedge \ \big(s(u, v) \geq \tau\big). \tag{2}$$

4.5. Descrição da *Baseline*

A baseline foi construída com o objetivo de fornecer um conjunto de referência controlado para avaliação dos modelos de similaridade [Cai and Wang 2018]. Para isso, foram gerados dados sintéticos a partir da combinação explícita de produtos e atributos, assegurando que a relação semântica entre os itens fosse conhecida *a priori*.

O conjunto é composto por 2.700 itens. Cada instância textual corresponde a uma variação descritiva de um mesmo produto base, preservando sua identidade semântica essencial. As variações incluem alterações de ordem dos termos, uso de sinônimos e pequenas modificações estruturais, mantendo o núcleo do objeto contratado.

Com base nessa estrutura, definem-se como **pares positivos** aqueles itens que derivam do mesmo produto base e, portanto, representam o mesmo objeto sob descrições distintas. Por outro lado, **pares negativos** correspondem a itens originados de produtos diferentes, caracterizando entidades semanticamente distintas.

A partir dessas definições, constrói-se uma matriz binária de similaridade que serve como referência para comparação com os modelos baseados em *embeddings*. A tarefa consiste em verificar a capacidade do modelo contínuo de similaridade em recuperar as correspondências previamente estabelecidas.

Ressalta-se que essa *baseline* possui caráter controlado e não busca reproduzir integralmente a complexidade dos dados reais de compras públicas. Seu objetivo é fornecer um ambiente reprodutível com verdade de referência conhecida, permitindo avaliar a capacidade do modelo em capturar similaridade semântica. Assim, os resultados devem ser interpretados como evidência de viabilidade técnica, sendo necessária validação complementar em bases reais.

4.6. Justificativa Metodológica: Seleção do Gemma 3

A eficácia do *framework* ComprasMatch depende da capacidade do sistema de mapear descrições heterogêneas de produtos para um espaço semântico unificado. Nesse contexto, a seleção do modelo Gemma 3 como motor para geração de *embeddings* se justifica por sua capacidade de produzir representações semânticas densas e contextuais, adequadas para capturar similaridades entre descrições textuais não padronizadas.

4.6.1. Superação da lacuna vocabular em compras públicas

Sistemas tradicionais de compras frequentemente enfrentam a *lacuna vocabular* (*vocabulary gap*), em que produtos equivalentes são descritos com terminologias distintas, abreviações ou variações estruturais. Métodos baseados em correspondência lexical, como TF–IDF, tendem a capturar apenas similaridade superficial entre termos, sem representar adequadamente a semântica subjacente [Deerwester et al. 1990].

Ao empregar o Gemma 3 para geração de *embeddings*, o ComprasMatch passa a representar descrições textuais em um espaço vetorial contínuo, no qual a proximidade

entre vetores reflete similaridade semântica. Essa abordagem permite capturar relações de sinonímia, variações de escrita e equivalência funcional entre itens, reduzindo a dependência de padronização textual.

4.6.2. Adequação ao problema de similaridade semântica

Diferentes abordagens podem ser utilizadas para tarefas de similaridade semântica, variando desde métodos lexicais até modelos baseados em inferência direta por *prompting*. A Tabela 1 apresenta uma comparação entre essas abordagens no contexto de representação textual.

Tabela 1. Comparação entre abordagens para similaridade semântica.

Categoria	Exemplo	Representação	Características
Lexical	TF–IDF / BM25	Vetor esparso	Baseado em frequência de termos; sensível a variações lexicais e estrutura textual
Embeddings	BERT / Gemma	Vetor denso	Captura similaridade semântica; robusto a sinonímia e variações de escrita
LLMs (*prompting*)	GPT / Llama	Texto direto	Decisão baseada em inferência contextual; maior custo computacional e menor escalabilidade para comparação em larga escala

No contexto deste trabalho, a abordagem baseada em *embeddings* se mostra mais adequada, pois permite pré-computar representações vetoriais e realizar comparações eficientes entre grandes volumes de dados. Essa característica é particularmente relevante em cenários de compras públicas, nos quais a identificação de similaridade exige a análise de múltiplos pares de itens provenientes de diferentes órgãos.

Assim, a escolha do Gemma 3 está alinhada ao objetivo central do ComprasMatch: fornecer uma representação semântica robusta e escalável das descrições de itens, viabilizando a construção de um grafo de similaridade para identificação de oportunidades de compras compartilhadas.

5. Discussão dos Resultados

Esta seção apresenta e discute os resultados obtidos com o ComprasMatch, considerando o desempenho do modelo em função do limiar de decisão e sua comparação com abordagens baseadas em modelos de linguagem de grande porte. Analisa-se o efeito do ajuste do *threshold* sobre as métricas de desempenho e confrontam-se esses resultados com aqueles produzidos por modelos baseados exclusivamente em *prompting*. Por fim, são discutidas as limitações computacionais da abordagem proposta.

5.1. Métricas de Avaliação

A avaliação do ComprasMatch exige métricas capazes de refletir seu comportamento em um cenário estruturalmente desbalanceado, no qual a maioria dos pares possíveis não representa correspondência válida. Nesse contexto, a acurácia isolada não descreve adequadamente o desempenho do modelo, pois pode permanecer elevada mesmo quando a detecção de pares semanticamente equivalentes se mostra limitada.

Por essa razão, o estudo adota conjuntamente as métricas de *acurácia*, *precisão*, *recall* e F_1-score, com ênfase na relação entre precisão e recall. A precisão assume

papel central porque falsos positivos implicam sugerir consolidações indevidas entre órgãos distintos, o que pode gerar impactos administrativos. O *recall* indica a capacidade do modelo de recuperar oportunidades reais de compras compartilhadas, evitando perda de ganho de escala. O F_1-score sintetiza esse equilíbrio e constitui a métrica de maior relevância analítica neste trabalho, especialmente em bases desbalanceadas [Sokolova and Lapalme 2012].

As métricas derivam da matriz de confusão apresentada na Tabela 2.

Tabela 2. Matriz de confusão para classificação de pares de itens

	Identificação correta	Identificação incorreta
Equivalente	TP	FN
Distinto	TN	FP

No contexto do ComprasMatch, TP representa pares corretamente identificados como semanticamente equivalentes; FP indica sugestões indevidas de equivalência; FN corresponde a oportunidades reais não detectadas; e TN representa distinções corretamente preservadas. Esse enquadramento permite interpretar diretamente o impacto institucional dos erros do modelo, alinhando avaliação estatística e aplicabilidade prática.

5.2. Análise dos Resultados em Função do Limiar de Decisão

O sistema converte a similaridade entre *embeddings* em decisão binária por meio de um limiar (*threshold*). A escolha desse parâmetro determina diretamente o equilíbrio entre falsos positivos e falsos negativos. Para avaliar esse efeito, as métricas de precisão, *recall*, F_1-score e acurácia foram analisadas em função de diferentes valores de limiar (Figura 2).

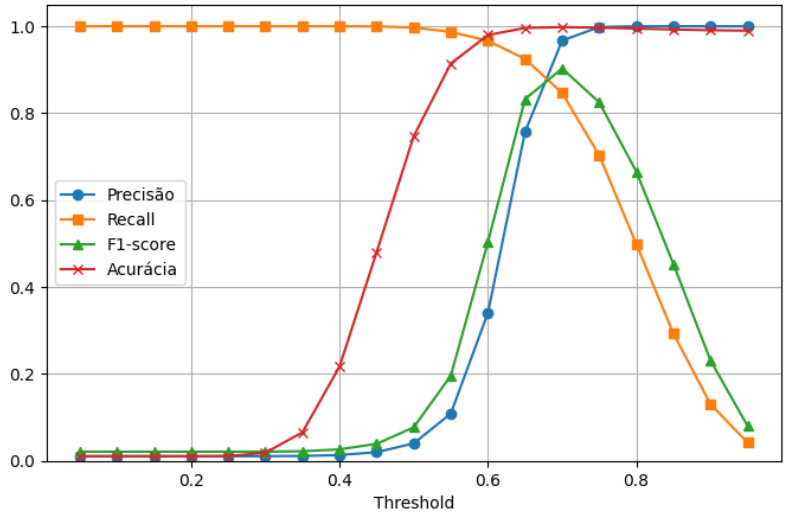

Figura 2. Variação das métricas de precisão, *recall*, F_1-score e acurácia em função do limiar de decisão (*threshold*).

Valores baixos de limiar maximizam o *recall* e reduzem a precisão, indicando excesso de correspondências sugeridas. À medida que o limiar aumenta, a precisão cresce e o *recall* diminui, caracterizando o *trade-off* esperado. O valor $\tau = 0{,}7$ foi selecionado por maximizar o F_1-score e manter acurácia elevada, configurando ponto de equilíbrio

operacional adequado ao objetivo do artigo: identificar correspondências confiáveis sem perder oportunidades relevantes de compras compartilhadas.

Para $\tau = 0{,}7$ (Equação 2), o modelo alcançou precisão de 96,71%, *recall* de 84,65%, F_1-score de 90,28% e acurácia de 99,81%. Esses resultados indicam alta confiabilidade nas correspondências sugeridas e capacidade consistente de recuperar pares semanticamente equivalentes, alinhando desempenho quantitativo e aplicabilidade institucional.

5.3. Comparação com Modelos Baseados em LLMs

Além da abordagem baseada em *embeddings* proposta neste trabalho, foi conduzido um experimento exploratório com modelos de linguagem de grande porte (LLMs) para a tarefa de verificação de similaridade semântica entre itens de licitação. Nessa configuração, a decisão é realizada diretamente a partir do texto, sem etapa explícita de vetorização, caracterizando um regime de classificação semântica por *prompting*.

Tabela 3. Comparação de desempenho entre o ComprasMatch e modelos baseados em LLMs.

Modelo	Precisão	*Recall*	Acurácia	F_1-score
ComprasMatch	96,71%	84,65%	99,81%	90,28%
GPT-5 mini	55,19%	15,00%	98,98%	23,59%
DeepSeek-chat	100,00%	9,12%	99,05%	16,72%
Llama 3	0,81%	0,55%	98,25%	0,65%

Foram utilizadas as APIs da OpenAI (GPT-5 mini), DeepSeek (*deepseek-chat*) e Meta (Llama 3). Os modelos receberam um *prompt* padronizado contendo um item de referência e um conjunto de itens candidatos, devendo indicar aqueles semanticamente similares ao item analisado. A avaliação foi conduzida com uma única configuração de *prompting*, sem ajuste fino ou otimização específica para cada modelo, com o objetivo de fornecer uma comparação inicial entre abordagens.

O experimento utilizou uma amostra aleatória de 335 itens extraída de uma população de aproximadamente 2.700 descrições da *baseline* sintética. Esse tamanho amostral garante estimativas de proporções com intervalo de confiança de 95% e margem de erro máxima de 5%, assumindo máxima variância ($p = 0{,}5$) [Cochran 1977]. As respostas foram comparadas à matriz de similaridade da *baseline*, permitindo o cálculo das métricas descritas na Seção 5.1. Esse protocolo permitiu a obtenção dos resultados apresentados na Figura 2.

Os LLMs apresentaram comportamentos distintos, porém convergentes quanto à limitação em recuperar pares semanticamente equivalentes. Em geral, os modelos obtiveram alta acurácia, mas baixos valores de *recall* e F_1-score, indicando dificuldade em identificar equivalências reais em um cenário desbalanceado.

Os LLMs apresentaram comportamentos distintos, porém convergentes quanto à limitação em recuperar pares semanticamente equivalentes. Em geral, os modelos obtiveram alta acurácia, mas baixos valores de *recall* e F_1-score, indicando dificuldade em identificar equivalências reais em um cenário desbalanceado. Conforme discutido na subseção 5.1, a acurácia isolada não é adequada para avaliar essa tarefa, pois pode permanecer elevada mesmo com baixa recuperação de equivalências.

Sob esse critério, a abordagem baseada em *embeddings* apresentou desempenho superior em termos de F_1-score quando comparada às soluções baseadas exclusivamente em *prompting*. Entretanto, os resultados devem ser interpretados sob a perspectiva de uma investigação exploratória, uma vez que os experimentos com GPT-5 mini, DeepSeek-chat e Llama 3 foram conduzidos com uma única configuração de *prompting*, sem estratégias adicionais de otimização, como engenharia de *prompt*, *few-shot prompting* ou ajuste fino dos modelos.

Assim, a comparação realizada não busca estabelecer um limite definitivo de desempenho para modelos baseados em LLMs, mas fornecer uma análise inicial entre diferentes abordagens para o problema de similaridade semântica em compras públicas. Trabalhos futuros podem investigar estratégias mais avançadas de adaptação e otimização desses modelos no domínio de licitações públicas.

5.4. Limitação do Modelo

A identificação de sinergias entre itens via similaridade de cosseno, embora matematicamente direta, impõe um gargalo crítico de escalabilidade devido à natureza de sua arquitetura computacional. Enquanto a geração de *embeddings* apresenta um custo linear $O(n)$, a etapa de comparação transversal exige o confronto de cada item contra todos os demais da base, o que resulta em uma complexidade quadrática $O(n^2)$. Esse fenômeno de "explosão combinatória" implica que o volume de operações cresce exponencialmente em relação ao número de itens, elevando drasticamente o tempo de execução e o consumo de memória RAM, o que torna a análise exaustiva de grandes volumes de dados proibitiva em hardware convencional sem o emprego de técnicas de redução de dimensionalidade ou estruturas de busca aproximada.

Entretanto, os testes de performance demonstraram que o modelo é aplicável a ambientes restritos com um número limitado de itens analisados, coadunando com o propósito de uso do modelo em situações reais. Esse cenário é comprovado na formatação da Baseline (seção 4.5) com 2.700 itens, em hardware pequeno (CPU AMD Ryzen 7 5700, GPU NVIDEA GeForce RTX 4060 TI, RAM 16 GB), com tempos de execução inferiores a 30 minutos, o que demonstra sua aplicabilidade.

6. Conclusão

O estudo demonstrou que o ComprasMatch é capaz de identificar similaridades semanticamente relevantes entre descrições textuais de itens no planejamento pré-contratual, apoiando a identificação de oportunidades de compras compartilhadas. Sob o limiar de similaridade $\tau = 0{,}7$, o modelo apresentou desempenho quantitativo elevado, com bom equilíbrio entre precisão e recall em um cenário caracterizado por forte desbalanceamento entre pares equivalentes e não equivalentes.

Na análise comparativa com três LLMs de referência, o ComprasMatch apresentou desempenho competitivo, especialmente em termos de F_1-score, refletindo um equilíbrio consistente entre confiabilidade das correspondências e capacidade de recuperação da classe positiva. Esse resultado é relevante no contexto institucional, no qual falsos positivos podem acarretar riscos administrativos, enquanto falsos negativos limitam o potencial de agregação de demandas e ganhos de escala.

Como limitação, destaca-se o custo computacional associado à comparação par a par, cuja complexidade quadrática restringe a escalabilidade do modelo em bases muito extensas. Ainda assim, os resultados indicam viabilidade técnica e aplicabilidade prática em cenários com volume moderado de itens, nos quais a qualidade semântica das correspondências é prioritária.

Trabalhos futuros podem explorar técnicas de *approximate nearest neighbors* (ANN), indexação vetorial eficiente e redução de dimensionalidade para mitigar o custo computacional. Também se aponta como extensão natural a validação em bases reais anotadas manualmente, bem como novas comparações com arquiteturas emergentes e experimentos de ajuste fino supervisionado, visando aprofundar a análise de robustez semântica em diferentes domínios.

Adicionalmente, recomenda-se que um órgão pública conduza um estudo de caso aplicando o ComprasMatch nas fases de planejamento da contratação e pesquisa de preços, avaliando a identificação de demandas convergentes com outros órgãos no PNCP, o potencial de economia de escala por meio de adesão a atas ou contratações conjuntas, bem como as resistências institucionais à adoção de compras compartilhadas, tais como o receio de perda de autonomia decisória sobre o processo licitatório e a relutância em compartilhar informações estratégicas de demanda com órgãos externos.

Todo o material utilizado neste estudo — incluindo código-fonte, scripts de processamento, dados sintéticos e documentação experimental — encontra-se disponível publicamente no repositório oficial do projeto: `https://github.com/KalistaS2/compraMatch/tree/main/.github`

Referências

Abdullahi, B., Ibrahim, Y. M., Ibrahim, A. D., Bala, K., Ibrahim, Y., and Yamusa, M. A. (2024). Development of machine learning models for categorisation of Nigerian government's procurement spending to UNSPSC procurement taxonomy. *International Journal of Procurement Management*, 19(1):106–121.

Appel, A. P., Silva, A. L. d. P., Silva, A. R., Santos, C. D., Silva, T. L. d., Araujo, R. P. d., and Aquino, L. C. F. d. (2022). Item matching using text description and similarity search. *arXiv preprint arXiv:2206.14097*.

Banco Interamericano de Desenvolvimento (2018). *Melhores gastos para melhores vidas: Como a América Latina e o Caribe podem fazer mais com menos*. BID, Washington, D.C.

Bednar, P., Vanko, J. I., and Žiak, E. (2025). Deep learning methods for multilingual classification of business documents. In *International Symposium on Applied Machine Intelligence and Informatics (SAMI)*.

Brum, P. P. V., Silva, M. O., Oliveira, G. P., Costa, L. G. L., Lacerda, A., and Pappa, G. (2024). Unsupervised grouping of public procurement similar items: Which text representation should I use? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17176–17185, Torino, Italy. ELRA and ICCL.

Cai, C. and Wang, Y. (2018). A simple yet effective baseline for non-attributed graph classification. *arXiv preprint arXiv:1811.03508*.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York, 3 edition.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Google DeepMind (2025). Gemma 3: Open models for advanced language understanding. Technical Report.

Guida, M., Caniato, F., and Moretto, A. (2025). Ai meets spend classification: A new frontier in information processing. *Journal of Purchasing and Supply Management*, 31(3):100993.

Góis, L. M. M. d. and Mendonça, C. M. C. d. (2023). Compras públicas centralizadas: vantagens e fatores críticos para o sucesso. *Fórum de Contratação e Gestão Pública*, 22(261):103–131.

Li, Y., Li, J., Suhara, Y., Doan, A., and Tan, W.-C. (2021). Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment*, 14(1):50–60.

Lima, W. E. M., da Silva, V. R., da Silva, J. C., Rabelo, R. d. A. L., and de Paiva, A. C. (2025). A deep learning approach for identifying similarities in contracted object description in public procurement. *Procedia Computer Science*, 264:125–136.

Mendes, C. M. T. (2023). A economia de escala alcançada nas compras compartilhadas: um panorama das aquisições governamentais no âmbito do consórcio público conisul. Observatório da Nova Lei de Licitações.

Ministério da Gestão e da Inovação em Serviços Públicos (2025). Estratégia nacional de contratações públicas para o desenvolvimento sustentável. Texto-base para consulta pública.

Moiraghi, F., Palmonari, M., Allavena, D., and Morando, F. (2024). Zero-shot hierarchical classification on the common procurement vocabulary taxonomy. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 273–278.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., and Raghavendra, V. (2018). Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)*.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Sokolova, M. and Lapalme, G. (2012). A comparative study of several classification metrics and their performances on data. *International Journal of Computer Applications*.

Vale, A. H., Santos, P., Soares, H., and Moura, R. S. (2023). Automatic classification of public expenses in the fight against COVID-19: A case study of TCE/PI. In *Proceedings of the XIX Brazilian Symposium on Information Systems (SBSI '23)*.