Semantic Clustering of Civic Proposals: A Case Study on Brazil's National Participation Platform

Ronivaldo Ferreira[1], **Guilherme da Silva**[2], **Carla Rocha**[2], **Gustavo Pinto**[1]

Faculdade da Computação
Universidade Federal do Pará (UFPA)
Belém – PA – Brazil

Faculdade do Gama
Universidade de Brasília (UnB)
Brasília – DF – Brazil

ronivaldo.junior@icen.ufpa.br

Abstract. *Promoting participation on digital platforms such as Brasil Participativo has emerged as a top priority for governments worldwide. However, due to the sheer volume of contributions, much of this engagement goes underutilized, as organizing it presents significant challenges: (1) manual classification is unfeasible at scale; (2) expert involvement is required; and (3) alignment with official taxonomies is necessary. In this paper, we introduce an approach that combines BERTopic with seed words and automatic validation by large language models. Initial results indicate that the generated topics are coherent and institutionally aligned, with minimal human effort. This methodology enables governments to transform large volumes of citizen input into actionable data for public policy.*

Resumo. *Promover a participação em plataformas digitais, como o Brasil Participativo, tornou-se uma prioridade para governos em todo o mundo. No entanto, devido ao enorme volume de contribuições, grande parte desse engajamento fica subutilizada, pois sua organização apresenta desafios significativos: (1) a classificação manual é inviável em escala; (2) é necessária a participação de especialistas; e (3) é preciso o alinhamento com taxonomias oficiais. Neste artigo, apresentamos uma abordagem que combina BERTopic com palavras-semente e validação automática realizada por grandes modelos de linguagem. Resultados iniciais indicam que os tópicos gerados são coerentes e alinhados institucionalmente, com esforço humano mínimo. Essa metodologia permite aos governos transformar grandes volumes de contribuições dos cidadãos em dados acionáveis para a formulação de políticas públicas.*

1. Introdução

A promoção da participação digital emergiu como uma agenda prioritária para governos em escala global, refletindo a crescente necessidade de modernização dos processos democráticos. Nesse contexto, surgem iniciativas voltadas à criação de plataformas tecnológicas que viabilizam a coleta, organização e análise de propostas da sociedade civil para a formulação de políticas públicas. Essas plataformas buscam não apenas ampliar o

acesso dos cidadãos aos processos decisórios, mas também fortalecer a transparência, a legitimidade e a responsividade das ações governamentais.

Nesse contexto, surgiu em 2023 a plataforma Brasil Participativo (BP) [Aguiar et al. 2024], com foco na ampliação do debate público mediante participação digital. A adoção dessa plataforma visou operacionalizar a iniciativa Plano Plurianual Participativo (PPA), resultando em uma participação massiva. Posteriormente, foram executados processos participativos de grande relevância política e social. O conjunto desses processos gerou mais de 1,4 milhão de cidadãos cadastrados e mais de 8 mil propostas elaboradas pelos cidadãos participantes (mais na Seção 2).

No processo de formulação de políticas públicas, a classificação das propostas apresentadas pela sociedade civil é crucial para definir tanto as ações a serem implementadas quanto os responsáveis por sua execução [Clemente 2018, Saravia and Ferrarezi 2007]. Todavia, esse procedimento manual em cenários de participação digital de larga escala enfrenta duas limitações que comprometem sua eficácia. Primeiramente, o volume expressivo de contribuições, que pode atingir dezenas ou centenas de milhares de propostas. Além disso, a classificação depende de conhecimentos especializados e multidisciplinares, abrangendo desde as especificidades temáticas de cada área de governo até os marcos regulatórios e as estruturas organizacionais, o que eleva custos temporais e financeiros.

Nesse contexto, técnicas de processamento de linguagem natural podem viabilizar o processamento em escala do acervo do BP, garantindo agilidade, consistência e rastreabilidade na categorização. Essas soluções apoiam o mapeamento semântico em taxonomias institucionais e favorecem a incorporação estruturada dos resultados nos ciclos de elaboração de políticas públicas.

Desenvolvemos um pipeline de extração de tópicos baseado em BERTopic [Grootendorst 2022] para organizar e interpretar o corpus. Além da abordagem não supervisionada, investigamos duas estratégias semi-supervisionadas: (i) *palavras-semente*, que incorporam termos extraídos do corpus para orientar a formação de clusters segundo categorias institucionais; e (ii) tópicos guiados, que impõem rótulos predefinidos para construir uma hierarquia semântica. Um Modelo de Linguagem de Grande Escala (LLM) foi empregado para validar automaticamente os tópicos e gerar interpretações, reduzindo a intervenção manual. A qualidade dos modelos foi medida por métricas de coerência, diversidade e alinhamento com a taxonomia.

Este estudo é guiado pelas seguintes questões de pesquisa:

1. **RQ1.** Quais ajustes nos parâmetros do BERTopic maximizam a coerência semântica e a diversidade temática dos tópicos extraídos?
2. **RQ2.** Em que grau a incorporação de *palavras-semente* do VCGE fortalece o alinhamento semântico com as categorias oficiais?

Este artigo apresenta a plataforma Brasil Participativo e seu vocabulário (Seção 2), a metodologia de modelagem de tópicos e validação (Seção 3), os resultados e respostas às questões de pesquisa (Seção 4), discussão sobre impactos e evolução (Seção 5), trabalhos relacionados (Seção 6) e conclusão com direções futuras (Seção 7). Artefatos disponíveis em: `https://github.com/BERTopic/bertopic_bp`.

2. A Plataforma Brasil Participativo

A plataforma Brasil Participativo (BP) foi instituída em 2023 pela Secretaria-Geral da Presidência da República como ambiente digital integrado para coleta, organização e priorização de contribuições cidadãs ao Plano Plurianual (PPA) 2024–2027, instrumento previsto na Constituição Federal[1]. Durante a execução do processo participativo, a plataforma registrou cerca de 1,4 milhão de acessos e mais de 8.000 propostas submetidas pelos cidadãos[2].

A plataforma apoia-se nas experiências federais de participação social, organizando-se em três frentes institucionais (planos, conferências e consultas) e oferecendo, para cada uma, ferramentas de propostas, enquetes e eventos. Dentre elas, as propostas despontaram como o principal canal de engajamento. O Brasil Participativo unifica esses instrumentos num portal colaborativo e interinstitucional, reunindo órgãos governamentais, Dataprev, UnB e a comunidade Decidim-Brasil.

2.1. Vocabulário Controlado de Governo Eletrônico (VCGE)

O VCGE é a taxonomia oficial mantida pela Secretaria de Logística e Tecnologia da Informação do Ministério do Planejamento, Orçamento e Gestão, criada para uniformizar a indexação de conteúdos informacionais no âmbito do Governo Federal[3]. Está organizado em um nível superior (N1) com 26 domínios temáticos, cada um subdividido em termos de segundo nível (N2); a Tabela 1 apresenta dez desses domínios como exemplo, e a lista completa está disponível no repositório do projeto. Cada conceito recebe um identificador numérico único e uma URI estável, assegurando interoperabilidade semântica e facilitando a integração de bases de dados, portais de consulta pública e APIs governamentais. Em plataformas de participação digital como a BP, o VCGE mapeia as propostas dos cidadãos em linguagem institucional consensual, reduzindo ambiguidades e inconsistências, permitindo referências precisas em relatórios, portais de transparência e sistemas de gestão de políticas, e apoiando a agregação e análise de dados.

Tabela 1. Mapeamento resumido de 10 categorias do VCGE (versão 2.1.0)

Nível 1 (N1)	Nível 2 (N2)
Agropecuária e Pesca	Defesa e vigilância sanitária, Produção agropecuária, Outros em Agropecuária
Comércio e Serviços	Comércio externo, Defesa do Consumidor, Turismo, Outros em Comércio e Serviços
Comunicações	Comunicações Postais, Telecomunicações, Outros em Comunicações
Cultura	Difusão Cultural, Patrimônio Cultural, Outros em Cultura
Defesa Nacional	Defesa Civil, Defesa Militar, Outros em Defesa Nacional
Energia	Combustíveis, Energia elétrica, Outros em Energia
Esporte e Lazer	Esporte comunitário, Esporte profissional, Lazer, Outros em Esporte e Lazer
Habitação	Habitação Rural, Habitação Urbana, Outros em Habitação
Indústria	Mineração, Produção Industrial, Propriedade Industrial, Outros em Indústria
Saneamento	Saneamento Básico Rural, Saneamento Básico Urbano, Outros em Saneamento

[1] https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/L14802.htm

[2] https://www.gov.br/planejamento/documentos-hospedados-para-gerar-qrcodes/relatorio-ppaparticipativo

[3] https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/registros-de-referencia/vocabulario-controlado-do-governo-eletronico

3. Metodologia

O desenho metodológico (Fig. 1) engloba quatro etapas: (i) extração e pré-processamento de dados; (ii) validação interna e ajuste de hiperparâmetros; (iii) modelagem de tópicos, mesclando descoberta não supervisionada e refinamento semi-supervisionado por tópicos-semente; e (iv) validação externa, comparando clusters ao VCGE via métricas de concordância e análise qualitativa do LLM.

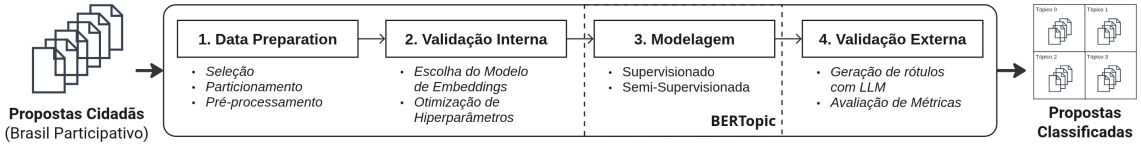

Figura 1. Pipeline de categorização temática com BERTopic

3.1. Seleção, Particionamento e Pré-processamento de Dados

Foram extraídas 10.186 propostas dos processos "Plano Clima", "Plano Plurianual Participativo" e "Congresso da Juventude". Após remoção de 164 duplicatas e registros vazios, restaram 10.022 amostras, divididas em 80% para treino (8.014) e 20% para teste (2.008). Para preservar a representação temática entre os conjuntos, o primeiro autor realizou uma análise manual das categorias selecionadas pelos proponentes na plataforma.

No conjunto de treino, o LLM Gemma 3 (12 bilhões de parâmetros)[4] gerou rótulos automáticos nos níveis N1 (categorias gerais) e N2 (subcategorias), adotados como *gold standard* na etapa de avaliação. O pré-processamento, aplicado a todo o corpus, incluiu normalização para minúsculas, remoção de pontuação e espaços duplicados, eliminação de *stopwords* em português, tokenização e lematização (por exemplo, `cidadãos` reduzido a `cidadão`) para evitar que variações morfológicas fossem tratadas como termos distintos.

3.2. Validação Interna

Nesta etapa, buscamos garantir que tanto a representação semântica quanto a configuração de tópicos estejam ajustadas ao nosso corpus. Seguimos o baseline de validação conforme [Hott et al. 2023], dividindo o trabalho em duas fases: (i) escolha do modelo de embeddings mais adequado e (ii) otimização dos hiperparâmetros do BERTopic.

Escolha do Modelo de Embeddings: Inicialmente, avaliamos quatro modelos de embeddings (ver Tabela 2). O conjunto de treino foi submetido a cada um desses modelos, gerando vetores de alta dimensão que capturam a semântica de cada proposta. Em seguida, para cada conjunto de embeddings, instanciamos e executamos o BERTopic variando o número de tópicos-alvo (`nr_topics`) nos valores 10, 30, 50, 70, 90, 110 e 130, além do modo `auto`, que determina automaticamente o melhor número de temas. Cada combinação foi repetida dez vezes para avaliar a consistência dos resultados. Em cada execução, calculamos as métricas Coerência Normalizada (NC) e Diversidade Normalizada (ND). Com esses dois indicadores, definimos a Pontuação Ponderada (WS), como:

$$WS = 0{,}8 \times NC + 0{,}2 \times ND.$$

[4] `https://ai.google.dev/gemma/docs/core`

Tabela 2. Modelos de Embeddings Avaliados

Modelo	Descrição
LaBSE [Feng et al. 2020]	Modelo multilíngue BERT (109 idiomas), com treino agnóstico ao idioma para representar semântica com alta qualidade.
Sentence-BERT [Reimers and Gurevych 2019]	Duas variantes do Sentence-BERT para múltiplos idiomas (MiniLM-L12-v2 e mpnet-base-v2), ambas otimizadas para similaridade semântica.
BERTimbau [Souza et al. 2020]	Modelos PT-BR base e large, treinados em corpus nativo; large foca em capacidade, base em eficiência.
LegalBERTPT-br [Silva et al. 2021]	Modelo jurídico em PT-BR, com SimCSE sobre BERTimbau, focado em nuances semânticas legais.

Otimização de Hiperparâmetros. Após definir o modelo de embeddings, otimizamos os principais hiperparâmetros do BERTopic por meio de busca em grade. Testamos dois intervalos de n-gramas (unigramas e bigramas) além de variações em `nr_topics` (entre 10 e 130, incluindo o modo `auto`). Paralelamente, investigamos diferentes limites para o parâmetro `min_topic_size`, estabelecendo os valores 3, 5, 10, 15, 20, 25 documentos por tópico, a fim de evitar tópicos com poucas amostras ou temas genéricos demais.

Personalização de Parâmetros ao Contexto do Corpus. Em [Hott et al. 2023], os intervalos de `nr_topics` foram (10, 13, 14, 15, 16, 17, 19, 20, auto) e os de `min_topic_size`, de 10 a 100 em passos de 10. Optamos por faixas distintas por duas razões: (1) o baseline usou documentos longos (licitações em PDF), enquanto trabalhamos com textos curtos, muitas vezes parágrafos; e (2) nossa plataforma abrange diversos processos, como PPA, G20 e Plano Clima, além de 55 categorias selecionáveis. Essa combinação de concisão e diversidade temática tende a reduzir significativamente o número de tópicos extraídos quando se utilizam valores elevados de `min_topic_size` ($>$30), comprometendo a granularidade e a representatividade dos temas.

3.3. Modelagem de Tópicos

Utilizamos uma abordagem não supervisionada e uma estratégia semi-supervisionada, que incorpora conhecimento institucional extraído do VCGE.

3.3.1. Abordagem Não Supervisionada

No modelo não supervisionado, iniciamos com o conjunto de documentos pré-processados e os embeddings BERTimbau-large correspondentes. Ao instanciar o BERTopic, definimos explicitamente `min_topic_size` $=$ 10, `nr_topics` $=$ 70 e `n_gram_range` $= (1, 1)$. O método `fit_transform` foi então aplicado, resultando em dois vetores principais: `topics_train`, que contêm o rótulo de tópico atribuído a cada documento, e `probs_train`, que indica a confiança do modelo em cada atribuição. Após a inferência, construímos uma tabela detalhada que relaciona cada tópico ao número de documentos atribuídos, ao nome (gerado automaticamente a partir das palavras mais representativas) e à lista de palavras-chave que caracterizam aquele tema. Essa tabela

serve de base para análises posteriores, permitindo identificar quais assuntos emergiram sem qualquer orientação prévia.

3.3.2. Abordagem Semi-Supervisionada

Este processo foi conduzido em três etapas. Na primeira etapa, construímos um dicionário `VCGE_TAXONOMY`, em que cada chave corresponde a uma categoria N1 e cada valor é a lista de suas subcategorias N2 (ver Tabela 1). Na segunda etapa, para cada categoria, geramos uma lista reduzida de `seed_words`, composta pelo próprio nome da categoria e por até cinco de suas subcategorias. Também removemos sistematicamente os termos `Outros`, pois não trazem contribuição semântica relevante para o alinhamento.

Os termos N2 foram escolhidos para representar de forma concisa cada domínio de N1, garantindo que as palavras mais representativas recebam peso diferenciado no cálculo de relevância do nosso *c-TF–IDF* personalizado. Por exemplo, a categoria `administração` foi resumida em `compras governamentais`, `orçamento`, `patrimônio`, `serviços públicos` e `recursos humanos`, enquanto `economia e finanças` ficou exemplificada por `defesa da concorrência`, `política econômica` e `sistema financeiro`, e assim sucessivamente para as demais categorias.

Por fim, na terceira etapa, instanciamos um objeto `ClassTfidfTransformer` personalizado, definindo `seed_multiplier` $= 2$ para reforçar o peso das `seed_words` no cálculo de relevância. Por fim, configuramos o BERTopic com os parâmetros `min_topic_size` $= 10$, `nr_topics` $= 70$ e `n_gram_range` $= (1, 1)$, além de incluir as variáveis `ctfidf_model` e `seed_topic_list`, que contêm, respectivamente, o transformer personalizado e o conjunto completo de tópicos-semente.

3.4. Validação Externa

A validação externa do pipeline de análise textual concentrou-se em duas avaliações complementares. A primeira consistiu na rotulação automática das propostas pelo LLM Gemma 3:12B, enquanto a segunda mediu a aderência dos tópicos inferidos pelo BERTopic em relação aos rótulos oficiais do VCGE.

3.4.1. Rotulação Automática com LLM

Para automatizar a atribuição de rótulos, carregamos duas listas com os termos válidos do VCGE (`vcge_n1_option` e `vcge_n2_option`). Em seguida, construímos um prompt com a sequência das opções de nível 1, de nível 2 e até 1.500 caracteres do documento, truncados para preservar a coerência. O prompt instruía o modelo Gemma 3:12B a responder com dois termos exatos, um de cada nível, no formato `<nível 1>, <nível 2>`, sem informações adicionais.

```
CLASSIFIQUE este texto usando APENAS UM destes
termos oficiais do VCGE: $termos_VCGE.
TEXTO: $text_limited_in_1500_chars
RESPONDA APENAS com o termo exato.  Nada além.
```

Utilizamos `temperature=0.2` e `num_ctx=2048`. Quando o modelo sugeria termos fora das listas, o rótulo `no_match` era atribuído. As respostas foram adotadas como *gold standard* para comparação com métodos não supervisionados e semi-supervisionados. Para facilitar a interpretação dos tópicos, empregamos o GPT o3-mini em prompt zero-shot para nomear os clusters extraídos pelo BERTopic. O modelo foi escolhido por sua capacidade de reasoning. O tópico -1"era rotulado como "Outliers"; os demais, com rótulos claros de até três palavras. A saída esperada era um dicionário com os IDs dos tópicos como chaves e seus respectivos rótulos como valores.

3.4.2. Avaliação de Tópicos

Para avaliar a qualidade dos tópicos gerados pelo BERTopic, utilizamos o conjunto de teste (20% dos documentos) com o mesmo pré-processamento descrito na Seção **??**. Em seguida, aplicamos os modelos não supervisionado e semi-supervisionado, obtendo `topics_test` e `probs_test`. Outliers (`topic` = –1) foram filtrados, e para cada proposta registramos: texto limpo, rótulos do LLM (`VCGE_N1` e `VCGE_N2`), tópicos inferidos e suas probabilidades.

Para quantificar a concordância com os rótulos oficiais, calculamos as métricas Adjusted Rand Index (ARI) e Normalized Mutual Information (NMI), separadamente para os níveis N1 e N2. Adicionalmente, construímos matrizes de contingência entre `topic` e rótulo de referência, complementadas por heatmaps com contagens absolutas e proporções normalizadas por categoria. Essas visualizações indicaram o grau de alinhamento entre os tópicos e as categorias do VCGE, além de possíveis lacunas ou sobreposições temáticas, oferecendo subsídios para aprimoramentos no pipeline de classificação.

4. Resultados

4.1. RQ1 – Quais ajustes nos parâmetros do BERTopic maximizam a coerência semântica e a diversidade temática dos tópicos extraídos?

Para esta questão, efetuou-se a escolha do modelo de embeddings mais adequado por meio de validação interna. A Figura 2 compara seis modelos, em que o eixo horizontal indica o número de tópicos-alvo (`nr_topics`) e o eixo vertical apresenta (WS). Observou-se que o BERTimbau-large supera consistentemente as demais alternativas em todas as faixas, refletindo sua capacidade de gerar vetores semânticos que equilibram coerência interna e diversidade temática. Em contraste, o Legal-BERT registrou desempenho inferior, sugerindo que seu treinamento no domínio jurídico não se generaliza à variedade temática das propostas da plataforma.

Com o BERTimbau-large selecionado como o melhor modelo de embeddings para representar nosso corpus, realizou-se a próxima etapa que buscou a melhor configuração de hiperparâmetros do BERTopic. A Tabela 3 resume as dez melhores configurações obtidas por meio da busca em grade. Cada configuração apresenta os valores dos seguintes parâmetros: intervalo de n-gramas utilizado no CountVectorizer (`n_gram_range`), número-alvo de tópicos (`nr_topics`), tamanho mínimo de tópico (`min_topic_size`), número efetivo de tópicos gerados (`topics`) e a Pontuação Ponderada Final (WS).

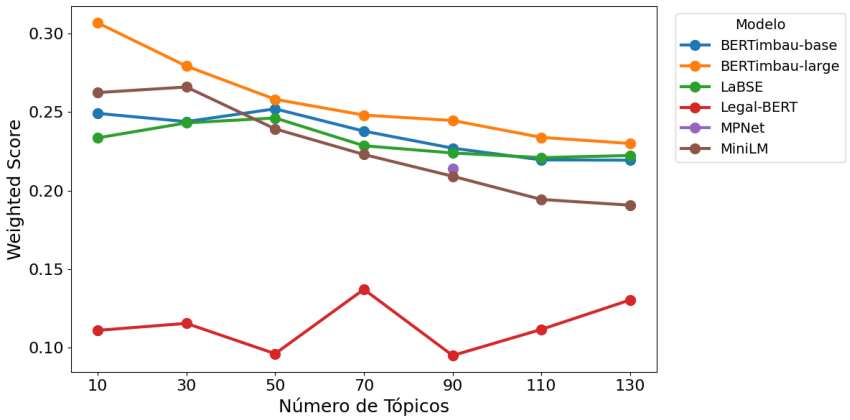

Figura 2. Weighted Score por Modelo e Número de Tópicos

Tabela 3. Top-10 Resultados da Validação Externa

n_gram_range	nr_topics	min_topic_size	topics	NC	ND	WS
(1,1)	70	10	56.1	0,11711	0,86234	0,26615
(1,2)	70	3	69.0	0,08720	0,89522	0,24880
(1,1)	110	10	72.0	0,09156	0,85527	0,24429
(1,2)	90	3	89.0	0,08028	0,88826	0,24188
(1,2)	70	5	69.0	0,08194	0,84971	0,23549
(1,2)	70	10	69.0	0,07758	0,84638	0,23133
(1,2)	110	3	109.0	0,06627	0,87550	0,22811
(1,1)	90	10	85.4	0,07566	0,83252	0,22702
(1,1)	70	5	69.0	0,07143	0,83580	0,22430
(1,1)	130	10	88.8	0,07273	0,82973	0,22413

Pode-se observar, entre todas as configurações testadas, o melhor resultado foi alcançado quando definimos `n_gram_range = (1,1)`, `nr_topics = 70` e `min_topic_size = 10`, produzindo 56 tópicos finais com $WS = 0{,}2661$. Essa configuração equilibra a necessidade de granularidade ao extrair um número de temas compatível com a diversidade do corpus e a robustez dos tópicos, evitando clusters muito pequenos ou excessivamente dispersos.

4.2. RQ2 – Em que grau a incorporação de tópicos-semente do VCGE fortalece o alinhamento semântico com as categorias oficiais?

Para avaliar o quanto a incorporação de tópicos-semente, extraídos do VCGE, reforça o alinhamento semântico na geração de tópicos com o BERTopic, foram comparados dois cenários: um modelo não supervisionado, sem qualquer reforço de termos, e outra configuração semi-supervisionada, que incorpora listas de termos-semente alinhadas às categorias oficiais do VCGE.

A Tabela 4 consolida os resultados obtidos em cada cenário para as métricas internas de qualidade de tópicos (NC, ND e WS) e para as métricas externas de alinhamento (ARI e NMI nos níveis hierárquicos N1 e N2), indicando também as diferenças absolutas e percentuais decorrentes da aplicação da semi-supervisão.

De forma geral, observa-se que a semi-supervisão eleva a coerência (NC), enquanto mantém a diversidade (ND) em patamar próximo ao não supervisionado, culminando em um ganho consolidado no WS. No aspecto de alinhamento externo, todos os

Tabela 4. Valores de Métricas Internas e Externas para os cenários não supervisionado e semi-supervisionada. Δ (%) refere-se à variação percentual de semi-supervisionada em relação a não supervisionado.

Métrica	Unsup	Semi-sup	Dif.	Δ (%)
NC	0,0953	0,1166	+0,0213	+22,4%
ND	0,8522	0,8420	−0,0101	−1,2%
WS	0,2467	0,2617	+0,0150	+6,1%
ARI (N1)	0,2095	0,3089	+0,0994	+47,5%
NMI (N1)	0,5366	0,5495	+0,0129	+2,4%
ARI (N2)	0,2105	0,2992	+0,0887	+42,1%
NMI (N2)	0,6088	0,6220	+0,0132	+2,2%

indicadores (ARI e NMI) apresentam melhorias, especialmente no ARI, que reflete maior aderência às categorias oficiais.

A Figura 3 nos ajuda a comparar melhor esses resultados obtidos. No lado esquerdo, as barras ilustram a diferença entre NC e ND ao adotar tópicos-semente. Já o lado direito detalha a composição de WS em cada cenário, considerando a ponderação entre as métricas internas também para cada cenário executado.

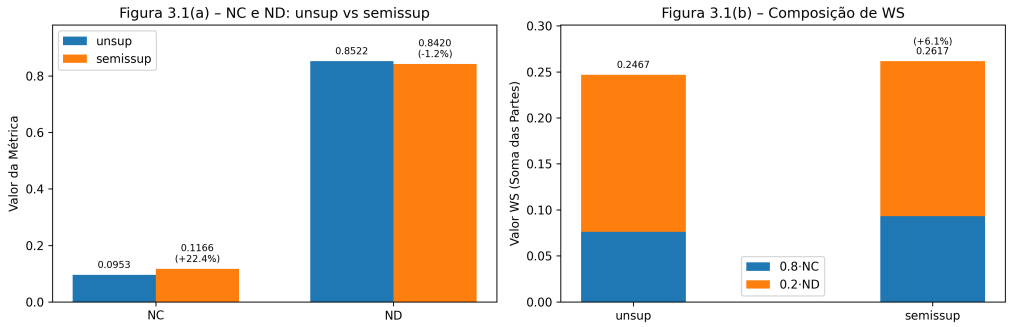

Figura 3. Comparação entre não-supervisionado e semi-supervisionado.

Pode-se observar que a medida de NC passa de 0.0953 no cenário não supervisionado para 0.1166 quando aplicamos a semi-supervisão com tópicos-semente, o que equivale a um aumento de 22.4%. Já a métrica ND, por sua vez, sofre uma leve queda de 0.8522 para 0.8420 (-1.2%). Quando combinamos essas duas medidas através do WS, observamos um crescimento de 0.2467 para 0.2617, ou seja, +6,1%, indicando que o ganho em coerência supera a pequena perda em diversidade.

A Figura 4 ilustra o efeito da inclusão de tópicos-semente nas métricas de alinhamento dos clusters gerados. À esquerda, são comparadas as correspondências entre tópicos e categorias de nível N1 do VCGE; à direita, o mesmo exercício para as subcategorias de nível N2.

Quanto às métricas externas, que quantificam o alinhamento entre tópicos gerados e categorias oficiais do VCGE, o ARI no nível N1 salta de 0.2095 para 0.3089 (+47.5%), enquanto o ARI no nível N2 cresce de 0.2105 para 0.2992 (+42.1%). Esses aumentos demonstram que a semi-supervisão aproxima substancialmente os clusters das categorias

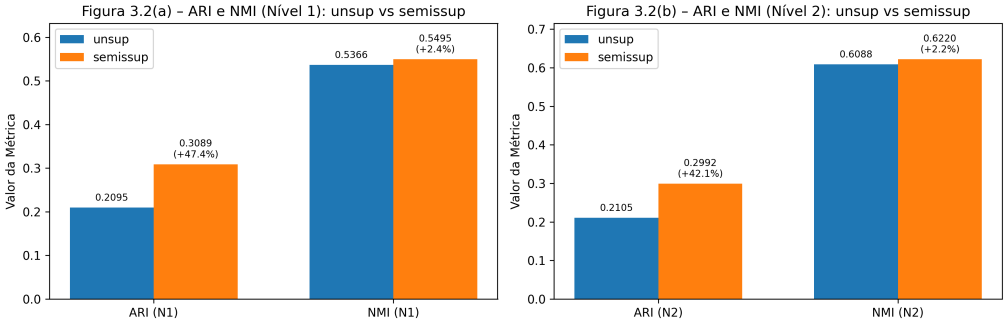

Figura 4. Métricas externas de alinhamento (ARI e NMI) nos níveis N1 e N2.

gerais (N1) e das subcategorias (N2) definidas pelo VCGE. A NMI no nível N1 passa de 0,5366 para 0,5495 (+2,4%) e a NMI no nível N2 de 0,6088 para 0,6220 (+2,2%), refletindo um alinhamento mais consistente em termos de informação mútua normalizada.

Portanto, a semi-supervisão por *tópicos-semente* fortalece de modo mensurável o mapeamento semântico entre os clusters gerados e as categorias oficiais do VCGE, equilibrando coerência interna e diversidade temática sem comprometer a cobertura da taxonomia institucional.

5. Discussões

5.1. Estratégia e Aplicabilidade

Os resultados indicam que a combinação de embeddings especializados em português brasileiro com semi-supervisão baseada em vocabulários oficiais é a estratégia mais eficaz para transformar grandes volumes de propostas cidadãs em insumos acionáveis: a adoção do BERTimbau-large para gerar representações semânticas, aliada ao uso de *palavras-semente* do VCGE, produziu tópicos que conciliam riqueza linguística e alinhamento institucional. Em termos operacionais, esse pipeline pode ser integrado imediatamente a plataformas de engajamento digital. Ao automatizar a categorização, ele reduz substancialmente o trabalho manual, libera especialistas para análises aprofundadas e permite a detecção em tempo real de tendências emergentes (por exemplo, picos de interesse em saúde ou meio ambiente).

5.2. Impactos e Beneficiários

Este método oferece dois benefícios principais: acelera a classificação de propostas, otimizando tempo e recursos humanos; e eleva a qualidade das informações, pois as categorias geradas são consistentes com taxonomias oficiais, garantindo rastreabilidade. Os principais beneficiários são equipes de formulação de políticas públicas, que passam a dispor de relatórios temáticos confiáveis para embasar decisões estratégicas, e desenvolvedores de plataformas cívicas, que podem integrar um módulo plugável de navegação temática para aprimorar interfaces de consulta e engajamento. Além disso, o alinhamento com vocabulários institucionais reforça a legitimidade do processo participativo: quando cidadãos percebem que suas contribuições são corretamente reconhecidas e agrupadas em categorias oficiais, fortalece-se a confiança e o engajamento democrático.

5.3. Replicabilidade e Evolução Contínua

Este estudo demonstra a viabilidade de IA orientada por vocabulários governamentais: mais do que extrair padrões estatísticos, o pipeline respeita e reforça estruturas institucionais. A metodologia proposta é replicável em diferentes contextos estaduais, municipais ou setoriais sem grandes ajustes, graças ao uso de embeddings em português e vocabulários oficiais amplamente disponíveis. Por fim, introduz-se a perspectiva de evolução contínua por meio de ciclos de feedback humano-modelo. Avaliações periódicas de especialistas podem refinar automaticamente as seed words e ajustar parâmetros, mantendo o sistema alinhado às mudanças nas demandas cidadãs. Essa abordagem garante adaptabilidade e relevância em um ambiente de participação digital dinâmico.

6. Trabalhos Relacionados

A aplicação de técnicas de Processamento de Linguagem Natural em documentos administrativos tem ganhado destaque, com evidências de que modelos especializados aumentam a performance em diversas tarefas. [Silveira et al. 2021] demonstraram que o uso do LEGAL-BERT em decisões judiciais melhora a coerência dos tópicos extraídos. [Silva et al. 2022] desenvolveram o LiPSet, corpus anotado de licitações, enquanto [Constantino et al. 2022] aplicaram aprendizado ativo na segmentação de diários oficiais, alcançando 85% de acurácia com menos dados rotulados. No pré-treinamento adaptativo de domínio (DAPT), [Silva et al. 2024a] mostraram que corpora alinhados ao domínio governamental elevam a precisão de modelos baseados em BERT, reforçando a importância da escolha do conjunto de dados. Complementarmente, [Hott et al. 2023] compararam embeddings como BERTimbau, LaBSE e LiBERT-SE em tópicos de compras públicas, evidenciando a vantagem de modelos em português. Seguindo essa linha, [Silva et al. 2024b] apresentaram o GovBERT-BR, treinado com textos oficiais de órgãos públicos brasileiros, que superou modelos generalistas em tarefas de classificação e segmentação. Apesar desses avanços, ainda há pouca investigação sobre modelagem de tópicos em plataformas participativas de larga escala e sua relação com taxonomias institucionais; lacuna que este trabalho busca preencher por meio de um pipeline com validação cruzada e uso de vocabulários oficiais.

7. Conclusão

Este trabalho propôs e avaliou um pipeline de modelagem de tópicos para a classificação de propostas públicas submetidas à plataforma Brasil Participativo. Ao combinar BERTopic com quatro diferentes embeddings, ajustes de hiperparâmetros e conhecimento institucional incorporado por meio de *palavras-semente* extraídos do VCGE, foi possível elevar a coerência e o alinhamento semântico dos tópicos gerados sem comprometer sua diversidade. A utilização de LLMs para rotulação automática mostrou-se eficaz para reduzir o esforço manual de validação, favorecendo a escalabilidade e a eficiência do processo. Ainda assim, mantém-se a necessidade de validação humana para casos de baixa confiança e a adaptação do vocabulário-semente e dos parâmetros quando o método for transferido para outros contextos. Em suma, o pipeline oferece uma solução operacional para transformar contribuições cidadãs em insumos acionáveis para a formulação de políticas públicas.

Referências

Aguiar, C. S. R., Alves, I., Gomes, L., Pinos, B., Bellix, L., and Parra, H. (2024). Colaboração multissetorial para desenvolvimento e manutenção de soluções tecnológicas de participação: o caso do brasil participativo.

Clemente, A. J. (2018). Leonardo secchi. análise de políticas públicas: Diagnóstico de problemas, recomendação de soluções. são paulo: Cengage learning, 2016.

Constantino, K., Cruz, V. A. L., Zucheratto, O. M., França, C., Carvalho, M., Silva, T. H., Laender, A. H., and Gonçalves, M. A. (2022). Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 304–316. SBC.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hott, H. R., Silva, M. O., Oliveira, G. P., Brandão, M. A., Lacerda, A., and Pappa, G. (2023). Evaluating contextualized embeddings for topic modeling in public bidding domain. In *Brazilian Conference on Intelligent Systems*, pages 410–426. Springer.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Saravia, E. and Ferrarezi, E. (2007). Políticas públicas. *Coletâneas. Volumes*, 1.

Silva, M. O., Oliveira, G. P., Costa, L. G., and Pappa, G. L. (2024a). Evaluating domain-adapted language models for governmental text classification tasks in portuguese. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 247–259. SBC.

Silva, M. O., Oliveira, G. P., Costa, L. G., and Pappa, G. L. (2024b). Govbert-br: A bert-based language model for brazilian portuguese governmental data. In *Brazilian Conference on Intelligent Systems*, pages 19–32. Springer.

Silva, M. O., Paula, A. F., Oliveira, G. P., Vaz, I. A., Hott, H., Gomide, L. D., Reis, A. P., Mendes, B. M., Bacha, C. A., Costa, L. L., et al. (2022). Lipset: Um conjunto de dados com documentos rotulados de licitações públicas. In *Dataset Showcase Workshop (DSW)*, pages 13–24. SBC.

Silva, N. F. d., Silva, M. C. R., Pereira, F. S., Tarrega, J. P. M., Beinotti, J. V. P., Fonseca, M., Andrade, F. E. d., and de Carvalho, A. C. d. L. (2021). Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 104–120. Springer.

Silveira, R., Fernandes, C. G., Araujo Monteiro Neto, J., Furtado, V., and Pimentel Filho, J. E. (2021). Topic modelling of legal documents via legal-bert. *Topic Modelling of Legal Documents via LEGAL-BERT*.

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.