Casi 17ª Edição
Congresso de Administração, Sociedade e Inovação

Artigo completo
De 03 a 05 de dezembro de 2025

Mapeando os tópicos das dissertações do PROFIAP utilizando BERTopic

Thiago Duarte de Souza[1]
Ivan Reinaldo Meneghini[2]
Sthéfany Ferreira Alves[3]

Resumo: Este trabalho mapeia o conjunto de 1.258 resumos de dissertações do Mestrado Profissional em Administração Pública (PROFIAP), defendidas entre 2015 e 2023, com o uso do BERTopic, um modelo de tópicos neural que combina embeddings do Sentence-BERT multilíngue, redução de dimensionalidade via UMAP e clusterização pelo HDBSCAN. O corpus foi construído a partir da Plataforma Sucupira (CAPES), seguido de pré-processamento com técnicas de lematização, filtragem de termos e geração de embeddings semânticos. A varredura de hiperparâmetros (min_cluster_size entre 5 e 50) resultou em 28 tópicos estáveis, agrupados em seis macroáreas: organização do trabalho e bem-estar, educação e permanência estudantil, transparência e governança digital, gestão financeira e compras públicas, sustentabilidade ambiental e inovação tecnológica. As análises revelaram tendências temporais relevantes, como o crescimento de estudos sobre teletrabalho durante a pandemia de COVID-19, a intensificação da pesquisa em governança digital após a implementação de políticas de dados abertos e o fortalecimento recente das agendas de sustentabilidade. Os resultados oferecem uma ferramenta de apoio ao redesenho curricular e ao planejamento estratégico do PROFIAP, mas também evidenciam lacunas em áreas como equidade de gênero, raça e políticas urbanas, ainda periféricas na produção do programa. Conclui-se que a abordagem adotada, embora sujeita a limitações inerentes à análise de resumos e ao julgamento humano na rotulação dos tópicos, fornece uma visão inédita e orientada por dados da trajetória temática do PROFIAP, além de apontar caminhos para pesquisas futuras.
Palavras-Chave: profiap; bertopic; pós-graduação; visualização.

1. Introdução

O Mestrado Profissional é uma modalidade de pós-graduação que tem como objetivo integrar o conhecimento acadêmico com a aplicação prática em contextos de trabalho. No Brasil, ao longo das últimas décadas, esses programas experimentaram um rápido crescimento quantitativo, impulsionado por políticas de financiamento alinhadas a agendas neoliberais e pelo aumento da demanda do setor produtivo por pesquisas aplicadas. A literatura documenta não apenas essa expansão numérica, mas também um debate contínuo sobre o rigor científico desses programas e sua capacidade de gerar soluções aplicáveis, destacando a tensão entre a manutenção de uma base teórica sólida e a entrega de respostas práticas e específicas ao contexto (Giacomazzo & Leite, 2014; Camargo, Moraes & Andrade, 2024).

Criado em 2014 sob a coordenação da Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (ANDIFES), o Mestrado Profissional em Administração Pública (PROFIAP) abrangia, até 2023, 29 universidades distribuídas pelas cinco regiões do Brasil. Apesar de sua ampla rede e da importância estratégica de alinhar a formação em administração pública com os desafios do mundo real, ainda há uma ausência notável de análises temáticas abrangentes sobre a produção acadêmica do PROFIAP. Embora alguns estudos tenham examinado regiões ou coortes específicas, como o mapeamento detalhado de dissertações em Minas Gerais realizado por de Deus, Paula e Paiva (2024), até o momento nenhuma pesquisa sintetizou o corpus nacional de resumos de dissertações do PROFIAP para identificar as principais áreas temáticas. Essa lacuna sugere uma oportunidade de aplicar técnicas de modelagem de tópicos ao conjunto completo de resumos até 2023, permitindo a identificação empírica de temas latentes e apoiando, em última instância, o refinamento das linhas do programa e a tomada de decisão institucional baseada em evidências.

A modelagem de tópicos tem sido amplamente aplicada em pesquisas em administração pública para analisar grandes repositórios de textos, seja na mineração de notícias e discursos parlamentares para revelar prioridades latentes (Chen & Wang, 2021; Payson et al., 2022); na implementação e monitoramento, por meio do processamento de registros administrativos e fluxos de redes sociais para detectar gargalos operacionais (Shi, Xu, Ying & Li, 2022; Zha, Ye, Li & Ozbay, 2023); ou ainda na avaliação e no feedback, quantificando o sentimento público e os resultados de políticas (Corrêa, Uriona-Maldonado & Vaz, 2022; Xu et al., 2021). Tradicionalmente, a maioria dos estudos recorre a representações Bag-of-Words combinadas com a Latent Dirichlet Allocation (LDA), apesar de revisões sistemáticas alertarem que essa abordagem frequentemente captura apenas coocorrências superficiais de termos e exige validação manual extensa (Hankar, Kasri & Beni-Hssane, 2025).

Para enfrentar essas limitações, adotamos o BERTopic, uma biblioteca recente de modelagem neural de tópicos que trata a descoberta de tópicos como um problema de agrupamento, aproveitando embeddings de sentenças derivados do BERT, junto a algoritmos de redução de dimensionalidade e clusterização em uma abordagem modular. Ao utilizar embeddings, essa arquitetura pode revelar relações semânticas latentes entre termos que nem sempre são descobertas pelos padrões de coocorrência típicos dos modelos LDA (Grootendorst, 2022).

Evidências empíricas de corpora relacionados a políticas também corroboram a interpretabilidade do BERTopic em cenários reais (Kazemi, Younus, Jeon, Qureshi & Caton, 2023). Essas vantagens motivam a escolha do BERTopic para mapear o panorama temático das dissertações do PROFIAP.

Nossas contribuições são três. Primeiro, apresentamos o primeiro mapeamento temático em escala nacional das dissertações do PROFIAP (2015-2023). Segundo, demonstramos como o BERTopic, ao integrar embeddings BERT, UMAP, K-Means e c-TF-IDF, capta relações semânticas latentes em resumos acadêmicos em português, revelando estruturas de tópicos mais refinadas. Terceiro, rastreamos a evolução temática ao longo da história do programa, fornecendo conhecimento baseado em evidências para subsidiar decisões futuras e pesquisas.

2. Fundamentação teórica

A modelagem de tópicos é uma técnica automática de descoberta de temas latentes em coleções de documentos textuais. Utilizada originalmente em sistemas de recuperação de informação, essa abordagem permite inferir padrões ocultos em um corpus, organizando, compreendendo e resumindo grandes volumes de texto de modo interpretável e não supervisionado (Abdelrazek, 2023). No campo da administração pública, seu uso tem crescido em análises de discursos parlamentares, registros administrativos e produção científica, como forma de identificar prioridades emergentes e monitorar tendências de pesquisa (Chen & Wang, 2021; Payson et al., 2022).

Segundo a revisão de Hankar, Kasri e Beni-Hssane (2025), a trajetória da modelagem de tópicos pode ser dividida em três ciclos evolutivos. O primeiro foi marcado por matrizes algébricas, inauguradas pela Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990), cujo emprego da decomposição de valor singular (SVD) permitiu reduzir dimensionalidade e captar associações semânticas. A segunda fase trouxe os modelos probabilísticos, como o Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) e a amplamente reconhecida Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003), que descrevem cada documento como mistura de distribuições temáticas inferidas por frequências de palavras. Finalmente, a terceira onda incorporou representações vetoriais densas (embeddings), iniciada pelo Word2Vec (Mikolov, Chen, Corrado & Dean, 2013) e aprofundada pelo BERT (Devlin, Chang, Lee & Toutanova, 2019), abrindo espaço para modelos híbridos como o Lda2Vec (Moody, 2016) e o BERTopic, que combina coerência semântica de embeddings com interpretabilidade probabilística (Grootendorst, 2022).

Apesar dos avanços, três desafios permanecem centrais: (i) a definição do número ótimo de tópicos, ainda dependente de ciclos longos de experimentação; (ii) a instabilidade e queda de interpretabilidade em textos curtos, como resumos de dissertações; e (iii) o elevado custo computacional em corpora volumosos. Esses desafios se tornam mais críticos no presente

estudo, que trabalha especificamente com abstracts do PROFIAP, caracterizados por vocabulário disperso e baixo contexto.

Para lidar com essas limitações, o BERTopic integra sentence embeddings, redução de dimensionalidade, clusterização e rotulação. Essa arquitetura elimina a necessidade de fixar *k* a priori, melhora a legibilidade dos tópicos e reduz o impacto da alta dimensionalidade típica dos embeddings. Estudos recentes destacam o bom desempenho do BERTopic em corpora curtos, avaliado por métricas de coerência e diversidade temática (Egger & Yu, 2022), o que justifica sua escolha neste trabalho para explorar os 1.258 resumos de dissertações do PROFIAP (2015-2023). A próxima seção detalha o funcionamento do pipeline adotado.

2.1. Bertopic

A biblioteca Bertopic apresenta um pipeline que possibilita a utilização de diversos modelos para modelagem de tópicos neural. O princípio por trás da biblioteca é a aplicação de uma abordagem de clusterização sobre representações semanticas dos documentos geradas através de modelos de embeddings textuais. O modelo de embedding captura relações semanticas nos documentos e representa em um vetor denso. Em uma segunda etapa é realizada a redução de dimensionalidade desses vetores, facilitando a terceira etapa crucial que é a aplicação de um algoritmo de clusterização. Uma vez gerados estes agrupamentos sob a representação vetorial do texto, é utilizada uma versão adaptada do algoritmo Term Frequency Inverse Document Frequency (TF-IDF), o Class-based Term Frequency Inverse Document Frequency (cTF-IDF) para identificação dos termos mais relevantes para a construção do tópico associado a cada cluster (Borcin, 2024; Grootendorst, 2022).

2.2. UMAP

Na etapa de projeção de embeddings do BERTopic, empregamos o algoritmo Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, & Melville, 2018), padrão da biblioteca. O método constrói um grafo de vizinhança fuzzy no espaço de alta dimensão, em que arestas representam probabilidades de conexão entre pontos, e depois otimiza a posição desses pontos em um espaço de menor dimensão minimizando a divergência entre as distribuições de conectividade nos dois espaços (McInnes, Healy & Melville, 2018; Allaoui, Kherfi & Cheriet, 2020; Wang, 2021). Essa transformação explora a suposição de uniformidade e conectividade local da variedade de dados, garantindo que estruturas de vizinhança próximas no espaço original sejam preservadas na projeção. Ao projetar os embeddings textuais em um espaço de menor dimensão, o UMAP prepara os dados para a clusterização, enfatizando proximidades semânticas locais (McInnes, Healy & Astels, 2017).

2.3. HDBSCAN

Após a projeção dos embeddings com UMAP, utilizamos também o algoritmo padrão de clusterização, o Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). Essa técnica realiza a clusterização com base em estimativas de densidade local, dispensando a definição prévia do número de agrupamentos.

Conforme McInnes, Healy e Astels (2017), o algoritmo emprega a distância de alcançabilidade mútua, definida como o maior valor entre as distâncias ao vizinho k-ésimo mais próximo de

dois pontos e a distância entre eles. Essa métrica é utilizada para construir uma hierarquia de agrupamentos com base em conectividade. Posteriormente, a hierarquia é condensada por meio de critérios de persistência, que selecionam os agrupamentos mais estáveis ao longo das escalas de densidade.

Uma das vantagens do HDBSCAN é a capacidade de rotular pontos com baixa densidade como ruído, evitando sua associação forçada a grupos existentes. No Bertopic estes outliers estão listados por padrão no tópico -1 (Grootendorst, 2022).

3. Método de pesquisa

Todos os procedimentos metodológicos foram implementados em Python 3.10. O processamento foi realizado em um processador i5 de 13ª geração, com 16GB de RAM e placa gráfica NVIDIA RTX3050. O código-fonte e os artefatos de reprodutibilidade estão disponíveis em repositório público: https://github.com/souza-td/bertopic_profiap

3.1. Construção do corpus

Foram recuperadas, a partir da interface de dados abertos da Plataforma Sucupira (CAPES), todas as dissertações desde a criação do PROFIAP em 2014 até 2023, último conjunto de dados disponível (CAPES, 2025). As exportações em formato CSV da plataforma foram baixadas e concatenadas em um único *DataFrame* do Pandas, posteriormente filtrado pelo identificador único do programa PROFIAP. Do *DataFrame* filtrado, foram mantidas apenas as colunas de título, resumo e ano de defesa. Embora a janela de coleta tenha começado em 2014, não houve defesas registradas no ano de criação; assim, os resumos analisados abrangem o período de 2015 a 2023. Foram identificados e removidos dois resumos duplicados, resultando em um conjunto final de 1.258 registros únicos, salvo em um novo arquivo CSV para as análises subsequentes. Como esses registros são de acesso público e destinados a pesquisa, não foi necessária aprovação ética adicional.

3.2. Pré-processamento de Texto e *Sentence Embeddings*

Para capturar nuances sintáticas e lexicais que poderiam não ser detectadas por modelos *bag-of-words*, foram inicialmente gerados *embeddings* semânticos densos com o modelo *paraphrase-multilingual-MiniLM-L12-v2* (Reimers & Gurevych, 2019). Em paralelo, foi construída uma matriz esparsa termo-documento utilizando o *CountVectorizer*, configurado para unigramas e bigramas (*ngram_range = (1,2)*). Tokens que ocorriam em menos de dois resumos foram descartados (*min_df = 2*), enquanto aqueles presentes em mais de 90% do corpus foram removidos (*max_df = 0.9*). Esses limiares ajudaram a reduzir ruído preservando o vocabulário relevante da área.

A tokenização e lematização foram realizadas com o modelo de português do SpaCy, com o *syntactic parsing* e o NER desativados. A lista padrão de *stopwords* foi expandida para incluir termos de alta frequência específicos da área (como *gestão*, *público*, *resultado*, *objetivo*) e também siglas das universidades participantes (UFJF, UFPE etc.).

3.3. Redução de Dimensionalidade

Os *embeddings* de alta dimensão foram projetados em um espaço de menor dimensionalidade com o UMAP (*Uniform Manifold Approximation and Projection*) (Ghojogh, Ghodsi, Karray & Crowley, 2021). Testes preliminares com vizinhanças acima de 5 resultaram em agrupamentos excessivamente amplos, o que levou à configuração final, que privilegia estruturas locais mais refinadas: *n_neighbors = 3*, *n_components = 5* e *min_dist = 0.1*.

3.4. Agrupamento (*Clustering*)

Os tópicos foram induzidos por meio do HDBSCAN. Todos os parâmetros seguiram os padrões do BERTopic, exceto o *min_cluster_size*, que define o tamanho mínimo permitido para um tópico. Foi avaliado um grid de valores para *min_cluster_size* ∈ {5, 10, 15, ..., 50}. Para cada valor, todo o pipeline foi treinado e foram registrados dois critérios: (i) coerência (*C_v*) e (ii) diversidade de palavras do tópico (razão de tokens únicos entre os 10 termos principais). A partir dos resultados, foi selecionada a solução mais interpretável.

3.5. Construção e Rotulação de Tópicos

As representações dos tópicos combinaram o estimador inspirado no KeyBERT e a *Maximal Marginal Relevance*, com coeficiente de diversidade ajustado para 0,3 (Carbonell & Goldstein, 1998). O ranqueamento resultante foi revisado manualmente e apresentou coerência consistente entre os tópicos, abrangendo áreas como serviços públicos digitais, qualidade de vida no trabalho e gestão da saúde pública. De cada lista, foram extraídos os dez termos mais bem ranqueados, analisados em conjunto com os cinco documentos mais próximos ao centroide do cluster. As rotulações descritivas foram atribuídas por consenso entre os autores.

4. Resultados e Discussão

A execução do pipeline do BERTopic sobre os 1.258 resumos de dissertações do PROFIAP produziu 28 tópicos interpretáveis e um grupo residual classificado como “Outliers”, conforme apresentado na Tabela I. A configuração selecionada, *min_cluster_size = 10*, *n_neighbors = 3* e *nr_topics = auto,* alcançou um equilíbrio entre amplitude semântica e interpretabilidade, resultando em valores de coerência de 0,394 e diversidade de 0,729, como mostrado na Tabela II. A clusterização hierárquica dos *embeddings* revelou a formação de seis macroáreas de ordem superior: organização do trabalho e bem-estar, educação e permanência estudantil, transparência e governança digital, gestão financeira e compras públicas, sustentabilidade ambiental e inovação tecnológica. Essas macroáreas são representadas na Figura 5 como seis ramos densos que convergem nos menores níveis de ligação.

Tabela 1. Tópicos identificados.

Tópico	Quantidade	Rótulo	Palavras-chave MMR (top-10)
-1	336	Outliers	aluno, comunicação, risco, custo, participação, planejamento, municipal, brasileiro, evasão,

			competência
0	86	Política Pública Brasileira	brasileiro, governo, legislação, racial, reforma, planejamento, participação, população, climático, comissão
1	78	Sustentabilidade e Agenda Verde	sustentabilidade, sustentável, ambiental, resíduo, água, projeto, consumo, agenda, verde, socioambiental
2	61	Avaliação da Educação Superior	curso, graduação, científico, campus, divulgação, aluno, integração, participante, portal, profiap
3	57	Gestão de Processos Institucionais	contrato, gerenciamento, racionalidade, patrimônio, mapeamento, campus, maturidade, banco, integrar, neoliberalismo
4	53	Estratégias de Governança Digital	governança, digital, brasileiro, militar, governo, legislação, municipal, judiciário, agência
5	51	Expansão da Educação Superior	expansão, estudantil, estudante, universitário, educacional, graduação, beneficiário, socioeconômico, brasileiro, campus
6	47	Inovação em Governança Digital	inovação, tecnológico, digital, governo, cidadão, modernização, transparência, acesso, mídia, cooperação
7	44	Gestão Fiscal Municipal	fiscal, dívida, despesa, brasileiro, endividamento, pagamento, contribuinte, finanças, crédito, econômico
8	42	Violência de Gênero	violência, assédio, feminino, moral, crime, homicídio, vulnerabilidade, vítima, criminalidade, polícia
9	36	Fiscalização de Contratos Públicos	contrato, disciplinar, legislação, normativo, contratação, licitação, legal, sanção, custo, responsabilidade
10	27	Desafios do Trabalho Remoto	teletrabalho, produtividade, laboral, teletrabalhador, desafio, distanciamento, aprendizagem, viabilidade, desvantagem, comunicação
11	27	Desafios em Licitação e Planejamento	compra, licitação, indicador, fornecedor, planejamento, desperdício, custo, eficiência, passivo, comprador
12	27	Transparência na Governança	transparência, corrupção, acesso, portal, brasileiro, eleitoral, democracia, accountability, cidadão, governamental
13	27	Avaliação da Qualidade no Trabalho	saúde, satisfação, estresse, profissional, percepção, emoção, mental, cultura, enfermagem, liderança
14	26	Iniciativas de Energia Sustentável	sustentável, sustentabilidade, fotovoltaico, solar, brasileiro, elétrico, socioambiental, florestal, renovável, amazônia
15	26	Estudos de Satisfação Organizacional	restaurante, brasileira, competência, independente, brasileiro, sociodemográfico, financeiro, estudante, influência, insatisfação
16	26	Desafios do Orçamento Público	orçamento, evasão, participativo, orçamentário, despesa, contrato, regulação, pagar, aluno, financeiro
17	22	Competências do Setor Público	competência, mapeamento, profissional, administrador, inovação, scorecard, capacitação, decreto, qualificação, docente
18	18	Políticas de Educação	estudante, agência, comissão, evasão, escola,

		Superior	democrático, fundação, reserva, produtivismo, rendimento
19	18	Governança em Saúde Pública	saúde, hospital, paciente, agricultor, profissional, agricultura, atendimento, maternidade, escolar, associação
20	18	Gestão de Resíduos Urbanos	resíduo, urbano, reciclável, mobilidade, transporte, sustentável, carro, comissão, ambientalmente, sanitário
21	18	Análise da Governança Pública	governança, ouvidoria, cultural, participação, gerenciamento, accountability, governança, agência, regulatório, municipal
22	18	Estratégias de Resposta à Pandemia	pandemia, teletrabalho, coronavírus, escolar, emergencial, desafio, mundo, accountability, pandemia, doença
23	16	Avaliações de Clima Organizacional	competência, percepção, satisfação, profissional, atributo, aluno, opinião, geografia, relacionamento, equipe
24	15	Gestão no Setor Público	projeto, maturidade, gerenciamento, hospitalar, transparência, privado, eficiência, institucionalização, isomorfismo, sedimentação
25	13	Políticas de Assistência Estudantil	estudantil, cultural, monitoramento, cultura, restaurante, socioeconômico, pnaes, ifrs, decreto, rendimento
26	13	Iniciativas de Pesquisa no Brasil	patente, inovação, tecnológico, depósito, profissional, pesquisador, fiocruz, importação, logístico, sustentável
27	12	Accountability e Transparência Institucional	transparência, accountability, integridade, compliance, portal, ifsp, blockchain, fundação, acesso

Fonte: Elaborado pelo autor.

4.1. Palavras-chave relevantes

A Figura 1 apresenta os cinco termos mais representativos dos tópicos 0 a 3. A lista completa de tópicos, acompanhada dos dez termos mais relevantes identificados pelo algoritmo de *Maximum Marginal Relevance* (MMR) com diversidade ajustada para 0,3, encontra-se na Tabela 1. Palavras recorrentes de alto peso, como *teletrabalho* e *produtividade* (organização do trabalho), *portal* e *transparência* (governança digital), além de *sustentabilidade* e *resíduo* (estudos ambientais), evidenciam uma separação temática clara entre as macroáreas.

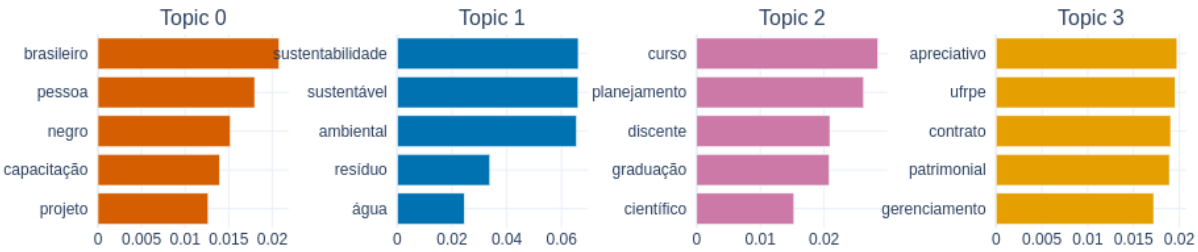

Figura 1. Termos mais representativos dos tópicos 0 a 3.
Fonte: Elaborado pelo autor.

4.2. Dinâmica temporal

A variação temporal da frequência dos tópicos apresentados na Tabela I revelou três pontos de inflexão importantes, ilustrado nas Figuras 2 e 3. Entre 2018 e 2019 houve um crescimento acentuado das dissertações sobre governança digital, alinhado temporalmente à promulgação das políticas de dados abertos no Brasil. Já no período de 2020 a 2021 observou-se um aumento de 2,5 vezes nos estudos sobre trabalho remoto, impulsionado pela pandemia de COVID-19, chegando a superar temporariamente a produção relacionada às finanças públicas. Finalmente, entre 2022 e 2023 verificou-se um crescimento contínuo nas pesquisas sobre sustentabilidade, que atingiram um patamar semelhante ao da governança digital.

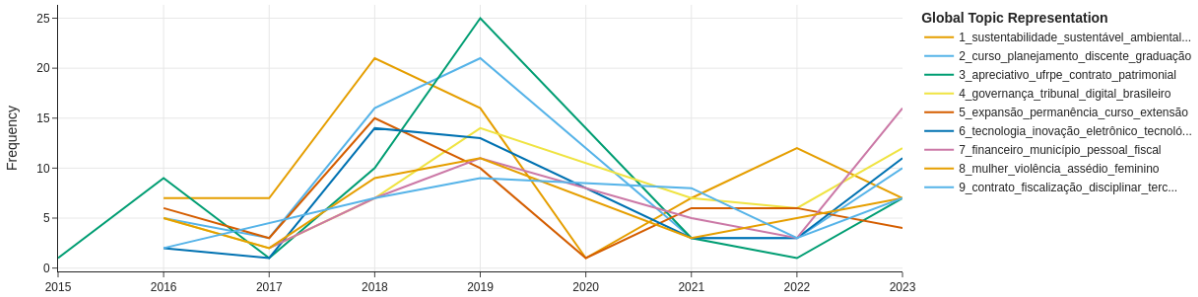

Figura 2. Distribuição temporal dos tópicos 1 a 9.
Fonte: Elaborado pelo autor.

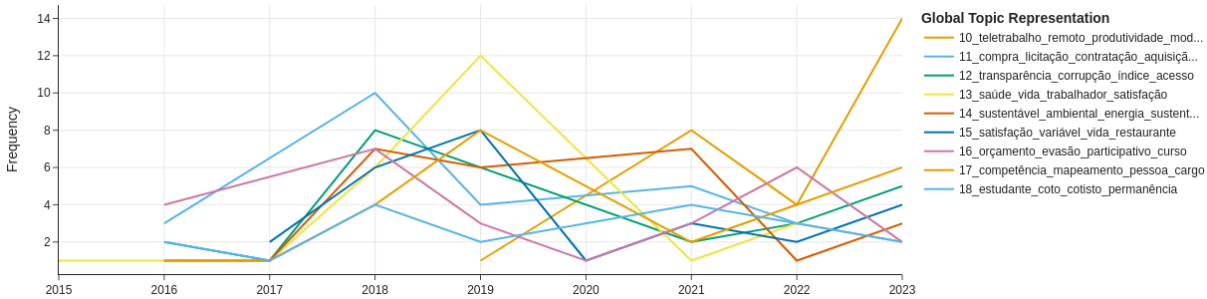

Figura 3. Distribuição temporal dos tópicos 10 a 18.
Fonte: Elaborado pelo autor.

4.3. Temas sub-representados

Os tópicos centrados em violência de gênero e inclusão racial reuniram menos de 50 documentos cada e surgiram em ramos periféricos do dendrograma (Figura 4). Já temas clássicos de políticas sociais, como habitação, mobilidade urbana e segurança alimentar, não

formaram clusters distintos; ao contrário, seus termos-chave ficaram dispersos em diferentes tópicos.

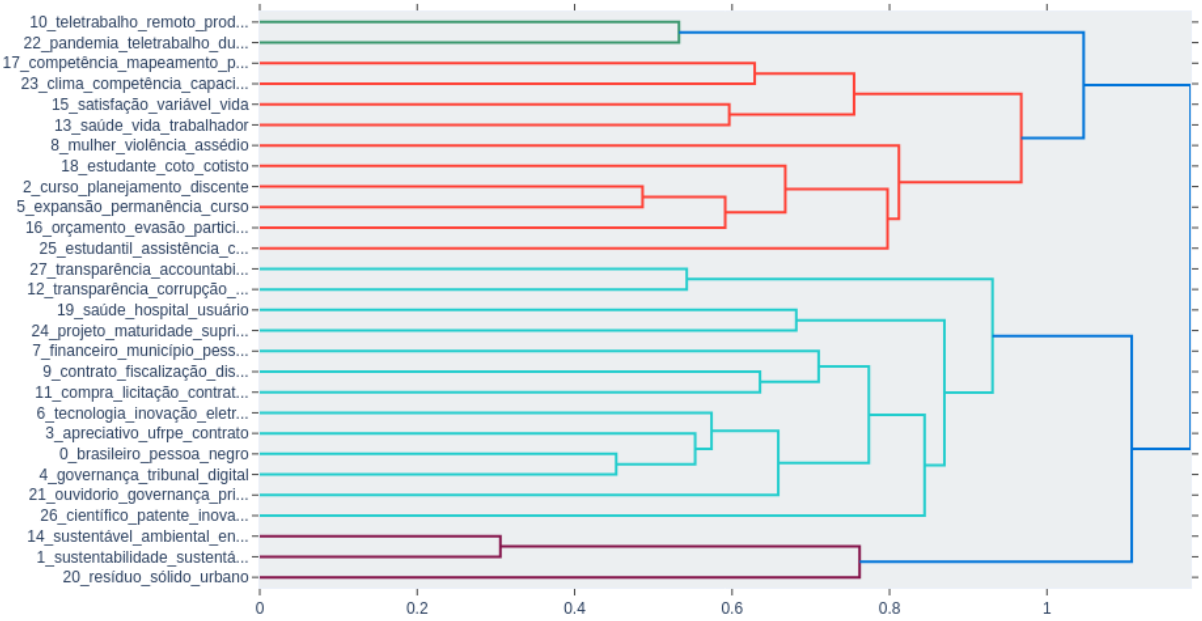

Figura 4. Dendrograma da representação hierárquica dos tópicos.
Fonte: Elaborado pelo autor.

4.1. Diagnóstico do modelo

A Tabela 2 apresenta uma comparação entre as execuções do modelo variando os valores de *min_cluster_size* entre 5 e 50. Observa-se que a coerência atinge seu pico quando o tamanho mínimo de cluster é 40 (Cv = 0,444), enquanto a diversidade cresce de forma monotônica à medida que o tamanho mínimo diminui. A configuração adotada, com *min_cluster_size = 10*, maximizou o número de tópicos sem provocar uma queda excessiva na coerência, preservando, assim, a interpretabilidade do modelo.

Tabela 2. Métricas por min_cluster_size.

min_cluster_size	n_topics	cv	diversidade
5	70	0,367103	0,692857
10	28	0,393727	0,728571
15	6	0,359499	0,833333
20	4	0,326891	0,925000
25	3	0,336211	0,933333
30	2	0,390035	0,950000
35	2	0,390035	0,950000
40	2	0,443669	0,950000
45	2	0,344276	1,000000
50	2	0,355507	1,000000

Fonte: Elaborada pelo autor.

5. Conclusões

A aplicação do BERTopic aos 1.258 resumos de dissertações do PROFIAP permitiu traçar um panorama inicial da produção acadêmica do programa. O modelo identificou vinte e oito tópicos estáveis, posteriormente agrupados em seis macroáreas: organização do trabalho, educação, governança digital, finanças públicas, sustentabilidade ambiental e inovação tecnológica. Esse mapeamento revela que a pesquisa desenvolvida no âmbito do PROFIAP cobre um leque amplo, mas ao mesmo tempo interconectado, de temáticas relevantes para a administração pública.

As análises temporais mostraram que fatores externos e mudanças de políticas públicas influenciam a distribuição dos temas ao longo do tempo. Houve intensificação de estudos sobre governança digital após a implementação de iniciativas de dados abertos no Brasil, crescimento expressivo de pesquisas sobre teletrabalho durante a pandemia de COVID-19 e fortalecimento das discussões sobre sustentabilidade nos anos mais recentes. Em contrapartida, temas relacionados à equidade, como gênero e raça, além de áreas clássicas da política urbana, como habitação, mobilidade e segurança alimentar, permaneceram periféricos, sugerindo campos ainda pouco explorados e que poderiam ser estimulados.

Essas observações, embora não prescrevam ações específicas, oferecem subsídios importantes para discussões internas do programa. Coordenadores podem, por exemplo, considerar a formalização de linhas de pesquisa mais consolidadas em torno da integridade digital e da gestão do trabalho, ao mesmo tempo em que exploram formas de incentivar investigações em domínios sub-representados. De modo mais amplo, a abordagem analítica empregada, combinando modelagem de tópicos não supervisionada, análise hierárquica e dinâmica temporal, demonstra potencial como ferramenta diagnóstica flexível para outros programas de pós-graduação interessados em refletir sobre seus portfólios de pesquisa com base em evidências.

Apesar dessas contribuições, o estudo apresenta limitações que precisam ser consideradas. A análise foi baseada exclusivamente em resumos, o que permitiu o processamento em larga escala, mas inevitavelmente deixou de fora detalhes metodológicos e resultados específicos contidos nos textos completos das dissertações. Além disso, a atribuição de nomes aos tópicos exigiu julgamento humano, prática comum na área, mas que introduz certo grau de subjetividade na interpretação. Outro ponto a destacar é que a agregação de quase uma década de produção em uma única base amplia o poder estatístico, mas pode mascarar mudanças mais sutis e trajetórias emergentes. A configuração escolhida do BERTopic, bem como o conjunto limitado de métricas adotado, coerência C_v e diversidade c-TF–IDF, também restringem o alcance dos resultados, uma vez que diferentes embeddings, algoritmos de clusterização ou critérios de validação poderiam gerar estruturas alternativas de tópicos. Por fim, as escolhas

específicas de pré-processamento para o português acadêmico, como listas de *stopwords* e esquemas de ponderação, podem ter influenciado a forma final do mapa temático.

Essas limitações abrem espaço para trabalhos futuros. Uma possibilidade é estender a análise para o texto completo das dissertações, o que enriqueceria a representação semântica e permitiria identificar também tendências metodológicas. A incorporação de técnicas semi-supervisionadas ou de rotulação com apoio de especialistas poderia reduzir a subjetividade na nomeação dos tópicos. Outra perspectiva é adotar um enfoque longitudinal, modelando separadamente as diferentes coortes ou aplicando modelos de tópicos dinâmicos, o que permitiria acompanhar em tempo real o surgimento, declínio ou reconfiguração de agendas de pesquisa, favorecendo ajustes curriculares mais ágeis. Aplicar o mesmo pipeline a dissertações de outros programas profissionais também possibilitaria comparações e *benchmarking* entre diferentes iniciativas de pós-graduação. Por fim, estudos de robustez que variem embeddings, algoritmos de clusterização e métricas de avaliação, aliados a validações qualitativas com docentes e egressos, poderiam fornecer um retrato ainda mais confiável e acionável da produção acadêmica do PROFIAP.

Referências

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020, June). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study. In *International conference on image and signal processing* (pp. 317-325). Cham: Springer International Publishing.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Borčin, M., & Jose, J. M. (2024, March). Optimizing BERTopic: Analysis and reproducibility study of parameter influences on topic modeling. In *European Conference on Information Retrieval* (pp. 147-160). Cham: Springer Nature Switzerland.

Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity- based reranking for reordering documents and pro- ducing summaries. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 335–336.

Chen, C. C., & Wang, H. -C. (2021). Adapting the influences of publishers to perform news event detection. Journal of Information Science, 49(5), 1277–1292.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Diretoria de Avaliação. (2025). Dados Abertos CAPES: Avaliação da Pós-Graduação Stricto Sensu [Acesso em: 31 mar. 2025]. https://dadosabertos. capes.gov.br/organization/diretoria-de-avaliacao

Corrêa, K. C., Uriona-Maldonado, M., & Vaz, C. R. (2022). The evolution, consolidation and future challenges of wind energy in uruguay. Energy Policy, 161, 112758.

de Camargo, A. M. M., Moraes, M. E. C., & Andrade, A. C. (2024). A expansão da pós-graduação e a modali- dade de mestrados profissionais: Como transcorreu a evolução da oferta?: The expansion of postgraduate studies and professional master's degrees: How did the offer evolved? Revista Cocar, (29).

de Deus, L. A., Paula, C. E. A., & de Paiva, A. L. (2024). Análise das dissertações do mestrado profissional em administração pública em rede nacional (profiap) do estado de minas gerais. Perspectivas em Políticas Públicas, 17(34), 8–34.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, *7*, 886498.

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). Uni- form manifold approximation and projection (umap) and its variants: Tutorial and survey. arXiv preprint arXiv:2109.02508.

Giacomazzo, G., & Leite, D. (2014). O mestrado profissional no âmbito das políticas públicas na educação: Reflexões Fig. 5. Hierarchical representation of the topics. sobre a produção do conhecimento científico. ETD Educação Temática Digital, 16(3), 475–493.

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

Hankar, M., Kasri, M., & Beni-Hssane, A. (2025). A com- prehensive overview of topic modeling: Techniques, applications and challenges. Neurocomputing, 129638.

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57).

Kazemi, A., Younus, A., Jeon, M., Qureshi, M. A., & Caton, S. (2023). Inéire: An interpretable nlp pipeline summa- rizing inclusive policy making concerning migrants in ireland. IEEE Access, 11, 88807–88823.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, *2*(11), 205.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Payson, J., Casas, A., Nagler, J., Bonneau, R., & Tucker, J. A. (2022). Using social media data to reveal patterns of policy engagement in state legislatures. State Politics & Policy Quarterly, 22(4), 371–395.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

Shi, C., Xu, T., Ying, Z., & Li, H. (2022). How policy mix choices affect the covid-19 pandemic response outcomes in chinese cities: An empirical analysis. International Journal of Environmental Research and Public Health, 19(13), 8094.

Xu, S., Sun, K., Yang, B., Zhao, L., Wang, B., Zhao, W., Wang, Z., & Su, M. (2021). Can public participation in haze governance be guided by government? – evidence from large-scale social media content data mining. Journal of Cleaner Production, 318, 128401.

Zha, W., Ye, Q., Li, J., & Ozbay, K. (2023). A social media data-driven analysis for transport policy response to the covid-19 pandemic outbreak in wuhan, china. Transportation Research Part A: Policy and Practice, 172, 103669.

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems, 112,* 102131.

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In *International Conference on Image and Signal Processing* (pp. 317–325). Springer.

Assembly, U. N. General, & others. (2015). *Transforming our world: The 2030 agenda for sustainable development.*

BleI, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

Borcin, M., & Jose, J. M. (2024). Optimizing BERTopic: Analysis and reproducibility study of parameter influences on topic modeling. In *ECIR (4)* (pp. 147–160).

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL, 30,* 31–40.

Camargo, A. M. M. de, Moraes, M. E. C., & Andrade, A. C. (2024). A expansão da pós-graduação e a modalidade de mestrados profissionais: Como transcorreu a evolução da oferta? *Revista Cocar, 29,* 1–25.

Campagnolo, J. M., Duarte, D., & Dal Bianco, G. (2022). Topic coherence metrics: How sensitive are they? *Journal of Information and Data Management, 13*(4).

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 335–336).

CAPES – Diretoria de Avaliação. (2025). *Dados abertos CAPES: Avaliação da pós-graduação stricto sensu*. https://dadosabertos.capes.gov.br/organization/diretoria-de-avaliacao