26º ENANCIB
Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
A Amazônia na sociedade global em rede: assimetrias e potencialidades informacionais, tecnológicas e ambientais
Belém, Pará, 09 a 13 de novembro de 2026
Universidade Federal do Pará

**ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO**
**ISSN 2177-3688**

**GT 8 – Dados, Informação e Tecnologia**

# MINERAÇÃO DA PRODUÇÃO CIENTÍFICA DA UNIVERSIDADE FEDERAL DE MINAS GERAIS

## *MINING THE SCIENTIFIC OUTPUT OF THE FEDERAL UNIVERSITY OF MINAS GERAIS*

**Mardochée Ogécime** – Universidade Federal de Minas Gerais (UFMG)
**Maria Aparecida Moura** – Universidade Federal de Minas Gerais (UFMG)
**Eduardo Valadares da Silva** – Universidade Federal de Minas Gerais (UFMG)
**Hugo Avelar Cardoso Pires** – Universidade Federal de Minas Gerais (UFMG)
**Leonardo Borges Rodrigues Chagas** – Universidade Federal de Minas Gerais (UFMG)

**Modalidade: Trabalho Completo**

**Resumo:** Este estudo analisa a produção científica depositada no Repositório Institucional da Universidade Federal de Minas Gerais entre os anos de 2021 e 2025, utilizando mineração semântica e modelagem de tópicos baseada em representações vetoriais contextualizadas. O objetivo consiste em identificar estruturas temáticas e padrões tipológicos da produção científica institucional por meio de arquiteturas de Processamento de Linguagem Natural. A coleta, realizada por meio do protocolo Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) do repositório institucional, resultou em 18.743 registros correspondentes ao período analisado. Destes, 12.037 documentos receberam atribuição temática válida na modelagem. Metodologicamente, o estudo combinou representações vetoriais semânticas, técnicas de redução de dimensionalidade, agrupamento por densidade e modelagem de tópicos. Os resultados permitiram identificar 48 agrupamentos temáticos substantivos, organizados principalmente em macrocampos relacionados à educação, saúde, ciências naturais e humanidades. A modelagem também evidenciou diferenciações entre regimes documentais da produção científica institucional, particularmente na distribuição de artigos, dissertações e teses entre os agrupamentos identificados. As métricas de estabilidade estrutural e compactação semântica indicaram consistência interna da solução analítica obtida. Conclui-se que abordagens baseadas em mineração semântica oferecem contribuições metodológicas relevantes para a análise informacional de grandes repositórios científicos, especialmente em aplicações relacionadas à organização da informação, ao monitoramento temático e à inteligência científica.

**Palavras-chave:** mineração semântica; modelagem neural; repositórios institucionais; produção científica; Ciência da Informação.

**Abstract:** This study analyzes the scientific output deposited in the Institutional Repository of the Federal University of Minas Gerais between 2021 and 2025 using semantic mining and topic modeling based on contextualized vector representations. The objective is to identify thematic structures and typological patterns in institutional scientific production through Natural Language Processing architectures. Data harvesting, performed through the repository's Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), yielded 18,743 records corresponding to the period analyzed. Of these, 12,037 documents received valid thematic assignments in the topic modeling process. Methodologically, the study combined semantic vector representations, dimensionality reduction techniques, density-based clustering, and topic modeling. The results enabled the identification of 48 substantive thematic clusters, mainly organized into broad fields related to education, health, natural sciences, and the humanities. The modeling also revealed differences among documentary regimes of institutional scientific production, particularly in the distribution of journal articles, master's dissertations, and doctoral theses across the identified clusters. Structural stability and semantic compactness metrics indicated the internal consistency of the analytical solution obtained. It is concluded that semantic mining approaches provide relevant methodological contributions to the informational analysis of large scientific repositories, especially for applications related to information organization, thematic monitoring, and scientific intelligence.
**Keywords:** semantic mining; neural modeling; institutional repositories; scientific production; Information Science.

## 1 INTRODUÇÃO

Os Repositórios Institucionais (RI) consolidaram-se como infraestruturas sociotécnicas estratégicas para preservação, organização e circulação da produção científica universitária, articulando diferentes regimes documentais, sistemas classificatórios e formas de representação do conhecimento (Bowker; Star, 1999; Leite, 2009). No entanto, a crescente volumetria e heterogeneidade dessas coleções impõem desafios analíticos à Ciência da Informação, especialmente diante das limitações de abordagens bibliométricas tradicionais centradas em produtividade e citação, frequentemente insuficientes para capturar regularidades semânticas latentes em grandes corpora multidisciplinares (Devlin; Chang; Lee; Toutanova, 2019; Wu; Chin, 2014).

Avanços recentes em Processamento de Linguagem Natural (PLN) e aprendizado profundo, particularmente a partir das arquiteturas *Transformer* (Vaswani *et al*., 2017), ampliaram as possibilidades de exploração semântica de coleções textuais. Modelos como BERT (*Bidirectional Encoder Representations from Transformers*) e Sentence-BERT permitem gerar representações vetoriais contextualizadas de documentos científicos, possibilitando

tarefas de agrupamento semântico e modelagem neural de tópicos (Devlin; Chang; Lee; Toutanova, 2019; Grootendorst, 2022; Reimers; Gurevych, 2019).

No contexto brasileiro, embora os repositórios institucionais estejam amplamente disseminados, ainda são relativamente escassos estudos que utilizem arquiteturas de aprendizado profundo para exploração semântica sistemática da produção científica institucional (Café *et al*., 2003; Moutinho; Martins; Alencar; Coneglian, 2024; Silva; Café; Catapan, 2010; Shintaku; Suaiden, 2015; Shintaku; Vidotti, 2016). Diante dessa lacuna, questiona-se: em que medida arquiteturas baseadas em *embeddings* semânticos e modelagem neural de tópicos permitem identificar estruturas temáticas da produção científica institucional?

O objetivo deste estudo consiste em aplicar uma arquitetura de mineração semântica à produção científica depositada no Repositório Institucional da Universidade Federal de Minas Gerais (UFMG), no período de 2021 a 2025, combinando representações semânticas baseadas em Sentence-BERT, redução de dimensionalidade com UMAP (*Uniform Manifold Approximation and Projection*), agrupamento por densidade com HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) e modelagem de tópicos com BERTopic, para identificação de agrupamentos temáticos e regularidades semânticas da coleção analisada. Parte-se da hipótese de que arquiteturas contemporâneas de mineração semântica permitem produzir representações semanticamente consistentes da produção científica institucional e identificar estruturas temáticas que se mantêm estáveis frente a variações controladas de parametrização, oferecendo contribuições metodológicas relevantes para a análise informacional de grandes repositórios científicos.

## 2 REPOSITÓRIOS INSTITUCIONAIS, ORGANIZAÇÃO E ANÁLISE INFORMACIONAL

Como objetos de estudo e intervenção na Ciência da Informação, os Repositórios Institucionais (RI) caracterizam-se por sua natureza multidimensional, articulando funções de preservação, organização, representação, acesso e comunicação da produção científica. Esses ambientes constituem infraestruturas sociotécnicas nas quais documentos, metadados, sistemas classificatórios e políticas institucionais participam da construção de condições de visibilidade e circulação do conhecimento acadêmico (Bowker; Star, 1999; Buckland, 1991; Leite, 2009).

Essa perspectiva aproxima a análise de repositórios dos problemas clássicos da

organização e da representação da informação e do conhecimento. Bräscher e Café (2008) distinguem a organização da informação, relacionada à descrição e ao arranjo dos objetos informacionais, da organização do conhecimento, orientada à estruturação conceitual dos domínios. De modo convergente, Alvarenga (2003) situa a representação como processo de mediação presente na produção dos registros de conhecimento, na organização dos sistemas documentais e nas condições de acesso à informação. Assim, a estrutura observável de um repositório resulta tanto das características dos documentos armazenados quanto dos procedimentos institucionais e conceituais empregados para descrevê-los e organizá-los.

A análise de domínio amplia esse enquadramento ao considerar que a organização do conhecimento se constitui em contextos disciplinares, discursivos e institucionais específicos. Para Hjørland (2002), os domínios podem ser investigados por diferentes abordagens, entre elas estudos bibliométricos, epistemológicos, documentais, terminológicos e das estruturas de comunicação científica. No contexto brasileiro da organização da informação, Guimarães (2014) destaca a análise de domínio como perspectiva capaz de identificar estruturas conceituais e relações que caracterizam campos específicos de conhecimento. Desse modo, regularidades observadas em coleções científicas devem ser interpretadas em relação aos contextos epistemológicos e documentais nos quais a produção do conhecimento ocorre.

Nesse sentido, métodos computacionais de exploração semântica podem acrescentar uma camada analítica à organização previamente expressa pelos metadados e sistemas documentais. A identificação automática de proximidades entre documentos possibilita examinar concentrações temáticas, diferenciações entre tipos de produção e outras regularidades que não são imediatamente observáveis pela inspeção dos registros individuais. Tais representações, contudo, possuem caráter analítico e exploratório: os agrupamentos inferidos computacionalmente não equivalem a sistemas formais de organização do conhecimento, mas constituem modelos empíricos derivados das propriedades textuais da coleção. Essa compreensão é compatível com a proposta de Börner e Scharnhorst (2009) de articular modelos conceituais da ciência, mensuração e evidências empíricas para representar estruturas da atividade científica.

A análise informacional contemporânea de grandes repositórios requer, portanto, métodos capazes de explorar estruturas semânticas subjacentes à produção científica e complementar formas consolidadas de descrição e mensuração. Embora abordagens bibliométricas e cientométricas sejam fundamentais para o estudo da produção e da

comunicação científica, medidas centradas em produtividade, citação ou frequência lexical podem ser complementadas por técnicas capazes de representar relações semânticas contextualizadas em *corpora* multidisciplinares e documentalmente heterogêneos (Devlin; Chang; Lee; Toutanova, 2019; Fortunato *et al*., 2018).

## 2.1 MODELAGEM TEMÁTICA NEURAL E MINERAÇÃO SEMÂNTICA

Avanços recentes em Processamento de Linguagem Natural (PLN), particularmente aqueles associados às arquiteturas *Transformer* (Vaswani *et al*., 2017), ampliaram as possibilidades de representação computacional do conteúdo semântico de grandes coleções textuais. Modelos como BERT (Devlin; Chang; Lee; Toutanova, 2019) e Sentence-BERT (Reimers; Gurevych, 2019) permitem representar textos por meio de *embeddings* densos e contextualizados, nos quais documentos semanticamente próximos tendem a ocupar regiões próximas de um espaço vetorial multidimensional.

Essa forma de representação oferece possibilidades distintas das abordagens baseadas predominantemente em frequência lexical, uma vez que incorpora relações contextuais entre os termos utilizados nos documentos. Em coleções científicas, *embeddings* podem ser empregados em tarefas de recuperação de informação, mensuração de similaridade e agrupamento documental, possibilitando investigar estruturas temáticas latentes em conjuntos textuais de grande escala (Reimers; Gurevych, 2019).

Nesse contexto, a modelagem de tópicos baseada em *embeddings* combina representações vetoriais contextualizadas, redução de dimensionalidade, técnicas de agrupamento e procedimentos de representação lexical dos grupos identificados. A arquitetura BERTopic (Grootendorst, 2022), utilizada neste estudo, articula essas etapas por meio da integração entre UMAP, HDBSCAN e representação lexical baseada em c-TF-IDF[1]. O UMAP preserva relações de vizinhança durante a redução do espaço vetorial, enquanto o HDBSCAN identifica agrupamentos por densidade sem exigir a definição prévia do número de tópicos. Posteriormente, o c-TF-IDF permite caracterizar lexicalmente os agrupamentos

[1] O *class-based Term Frequency–Inverse Document Frequency* (c-TF-IDF) consiste em uma adaptação do algoritmo TF-IDF para contextos de agrupamento documental. Nesse procedimento, os documentos pertencentes a um mesmo ***cluster*** são tratados conjuntamente como um "documento de classe", possibilitando identificar os termos que melhor distinguem cada agrupamento dos demais. Enquanto o TF-IDF convencional estima a relevância lexical em documentos individuais, o c-TF-IDF produz uma representação lexical dos tópicos ou classes resultantes da modelagem, favorecendo sua interpretação (Grootendorst, 2022).

resultantes mediante a identificação dos termos mais distintivos de cada tópico.

Sob a perspectiva da Ciência da Informação, o interesse dessa arquitetura reside na possibilidade de transformar proximidades semânticas calculadas entre registros documentais em estruturas passíveis de interpretação informacional. Os tópicos resultantes podem revelar concentrações temáticas, especializações e relações entre formas documentais da comunicação científica. Sua interpretação, entretanto, depende do *corpus*, das representações vetoriais utilizadas e das decisões de parametrização da arquitetura. Por essa razão, a consistência dos agrupamentos deve ser examinada por procedimentos específicos de validação e interpretada em articulação com as características documentais e institucionais da coleção analisada.

## 3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa caracteriza-se como um estudo exploratório-computacional de mineração semântica. O corpus compreende a produção científica depositada no Repositório Institucional da UFMG com recorte temporal de 2021 a 2025.

A coleta foi operacionalizada em 29 de novembro de 2025 por meio do protocolo *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) do Repositório Institucional da UFMG, utilizando o formato de metadados QDC (*metadataPrefix="qdc"*), garantindo a extração padronizada de metadados (título, autoria, tipo documental, descritores, palavras-chave e resumo). O procedimento recuperou inicialmente 74.367 registros. Como o recorte temporal empregado pelo OAI-PMH incide sobre o *datestamp* dos registros do repositório, realizou-se posteriormente a filtragem pelo campo correspondente ao ano de publicação, mantendo-se 18.743 registros publicados entre 2021 e 2025. A modelagem de tópicos foi aplicada a esse conjunto. Desses documentos, 12.037 (64,2%) receberam atribuição temática válida pelo BERTopic, enquanto 6.706 (35,8%) foram classificados como observações periféricas ou ruído pelo HDBSCAN (tópico -1). As análises temáticas subsequentes consideraram os 12.037 documentos com atribuição a tópicos substantivos.

A representação textual foi construída mediante a concatenação dos campos palavras-chave, assunto, título e resumo/descrição. Na composição utilizada para a modelagem, o campo de palavras-chave recebeu ponderação adicional mediante sua repetição três vezes na representação textual, com o objetivo de ampliar sua contribuição para a caracterização temática dos documentos. O pré-processamento envolveu o preenchimento de campos

textuais ausentes com valores vazios e a normalização de espaços em branco.

Os procedimentos computacionais foram implementados em Python 3 no ambiente Google Colab, com aceleração por GPU. A execução registrada utilizou uma GPU NVIDIA A100-SXM4-40GB. Entre as principais bibliotecas empregadas estão *pandas*, *NumPy*, *PyTorch*, *sentence-transformers*, *BERTopic*, *umap-learn*, *hdbscan* e *scikit-learn*.

A vetorização dos documentos utilizou o modelo Sentence-BERT (*all-MiniLM-L6-v2*)[2], arquitetura baseada em *Transformers* que gera *embeddings* semânticos contextualizados (Reimers; Gurevych, 2019). Os documentos foram projetados em um espaço vetorial de 384 dimensões. A escolha dessa arquitetura considerou sua representação vetorial compacta, com 384 dimensões, e seu menor custo computacional em tarefas de similaridade e agrupamento. Entretanto, o *all-MiniLM-L6-v2* foi desenvolvido predominantemente para língua inglesa, enquanto a coleção analisada apresenta forte presença de documentos em português (Sentence Tranformers, 2024). Essa diferença linguística é assumida como limitação metodológica, razão pela qual as métricas empregadas neste estudo são interpretadas como avaliações da consistência interna da solução obtida, e não como validação da adequação linguística do modelo ao português.

A arquitetura de mineração semântica foi estruturada em quatro etapas integradas:

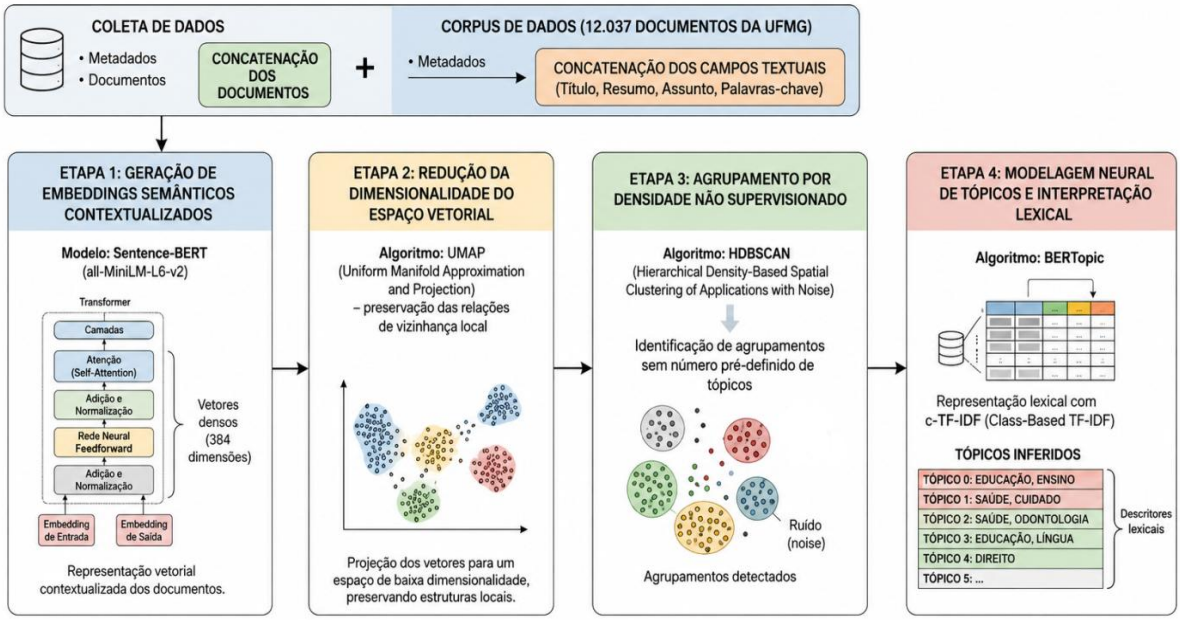
a) geração de *embeddings* semânticos com *Sentence-BERT*;
b) redução de dimensionalidade com *UMAP (Uniform Manifold Approximation and Projection)*;
c) agrupamento por densidade com *HDBSCAN*; e
d) modelagem de tópicos com BERTopic (Grootendorst, 2022).

O *UMAP* foi utilizado para preservação das relações de vizinhança no espaço vetorial reduzido (McInnes; Healy; Melville, 2018), enquanto o *HDBSCAN* permitiu a identificação não supervisionada de agrupamentos sem necessidade de definição prévia do número de tópicos, com maior tolerância a ruídos e documentos semanticamente periféricos (Campello; Moulavi;

[2] O modelo *all-MiniLM-L6-v2* corresponde a uma variante compacta do Sentence-BERT baseada na arquitetura MiniLM, otimizada para tarefas de similaridade semântica e recuperação de informação. O modelo produz *embeddings* densos de 384 dimensões, com menor custo computacional em comparação com arquiteturas *Transformer* de maior porte. Embora seja amplamente utilizado em tarefas de *clustering* e busca semântica, sua otimização predominante para textos em língua inglesa deve ser considerada na interpretação dos resultados obtidos em um corpus majoritariamente constituído por documentos em português.

Sander, 2020). A interpretação lexical dos agrupamentos foi realizada pelo *BERTopic*, utilizando ponderação *c-TF-IDF* para extração dos descritores mais representativos de cada tópico. Os documentos classificados pelo HDBSCAN como ruído foram identificados pelo rótulo -1 e não foram considerados tópicos substantivos nas análises temáticas. A solução final compreendeu, assim, 48 tópicos temáticos válidos. A Figura 1 sintetiza o *pipeline* metodológico empregado na mineração semântica da coleção analisada.

**Figura 1** - Arquitetura metodológica da mineração semântica




**Fonte**: elaboração própria (2026).

A robustez analítica da arquitetura foi examinada por três estratégias complementares de validação:

a) A primeira consistiu na avaliação da compactação semântica intra-tópico, estimada pela similaridade cosseno média entre os *embeddings* documentais e os centróides vetoriais dos respectivos agrupamentos temáticos. Para tópicos com mais de 300 documentos, foi utilizada uma amostra aleatória máxima de 300 registros, com semente aleatória fixada em 42, enquanto tópicos menores foram avaliados integralmente;

b) A segunda estratégia envolveu análise da estabilidade estrutural da modelagem por meio do *Adjusted Rand Index (ARI)* e do *Normalized Mutual Information (NMI)*, calculados a partir de diferentes parametrizações da arquitetura BERTopic. Foram comparadas duas configurações, com min_topic_size igual a 30 e 50, respectivamente,

sendo as métricas calculadas sobre os documentos que receberam atribuições temáticas válidas em ambas as soluções;

c) Por fim, realizou-se avaliação supervisionada auxiliar da representação vetorial produzida, utilizando classificação do tipo documental como tarefa *downstream*[3]. Foram comparados modelos baseados exclusivamente nos *embeddings* semânticos e modelos combinando *embeddings*, tópicos inferidos e variáveis temporais. A tarefa foi implementada por meio de Regressão Logística (*max_iter=1000*), comparando uma matriz composta exclusivamente pelos 384 atributos dos *embeddings* com uma matriz expandida pela representação *one-hot* dos tópicos e pelo ano de publicação normalizado. Os dados foram divididos de forma estratificada em conjuntos de treinamento (70%), validação (15%) e teste (15%), utilizando *random_state=42*.

## **4 RESULTADOS E DISCUSSÃO:** PANORAMA DO CORPUS

Do universo inicial de 18.743 registros, 12.037 (64,2%) receberam atribuição temática válida. Os 6.706 registros restantes (35,8%) foram classificados pelo HDBSCAN como observações periféricas ou ruído (tópico -1) e, por isso, não integraram as análises temáticas subsequentes. Os 12.037 documentos com atribuição válida constituíram, assim, o corpus analítico utilizado para caracterizar as estruturas temática e tipológica da produção científica da UFMG.

### 4.1 ESTRUTURA TEMÁTICA GLOBAL DA PRODUÇÃO CIENTÍFICA

A modelagem revelou uma organização composta por 48 tópicos substantivos**,** além da categoria de ruído representada pelo tópico -1. Para fins de síntese e interpretação, a Tabela 1 apresenta os dez agrupamentos de maior cardinalidade. Esses tópicos reúnem 6.906 documentos, correspondendo a 57,4% dos 12.037 registros com atribuição temática válida. Os 38 tópicos restantes, individualmente menos numerosos, compõem uma "longa cauda" temática equivalente a 42,6% do corpus analisado.

[3] Em Processamento de Linguagem Natural, *downstream tasks* correspondem a tarefas aplicadas de avaliação realizadas a partir de representações previamente aprendidas por modelos linguísticos, como classificação, agrupamento, recuperação de informação ou análise semântica (Devlin et al., 2019; Reimers; Gurevych, 2019). No contexto deste estudo, a classificação do tipo documental foi utilizada como tarefa *downstream* para avaliar a capacidade dos *embeddings* semânticos em preservar informações relevantes sobre a estrutura tipológica da produção científica analisada.

**Tabela 1** - Estrutura temática dominante (Top 10 tópicos, 2021-2025)

| Tópico | n_docs | Tipo predominante | Área temática inferida |
|---|---|---|---|
| 0 | 1.586 | Artigo | Educação / Ensino |
| 1 | 1.009 | Artigo | Saúde / Cuidado |
| 2 | 1.008 | Monografia | Língua Portuguesa / Ensino |
| 3 | 591 | Artigo | COVID-19 |
| 4 | 568 | Dissertação | Recursos Hídricos / Mineração |
| 5 | 518 | Dissertação | Química / Espectrometria |
| 6 | 513 | Artigo | Ciências Biológicas |
| 7 | 399 | Artigo | Saúde Bucal |
| 8 | 362 | Dissertação | Direito |
| 9 | 352 | Dissertação | Crítica Literária |

**Fonte**: elaboração própria (2026).

Os dados permitem reconhecer três conjuntos temáticos predominantes entre os dez maiores agrupamentos:

a) Educação e Linguagem (Tópicos 0 e 2): reúnem produções associadas ao ensino, à formação e aos estudos da língua e da literatura;
b) Saúde e Ciências Biomédicas (Tópicos 1, 3 e 7): concentram produções relacionadas ao cuidado, à COVID-19 e à saúde bucal;
c) Ciências Ambientais e Exatas (Tópicos 4, 5 e 6): agregam produções relacionadas a recursos hídricos e mineração, química e espectrometria e ciências biológicas.

## 4.2 REGIMES DE PRODUÇÃO DE CONHECIMENTO

A análise tipológica por tópico permite examinar como diferentes tipos documentais se distribuem entre os agrupamentos temáticos. O Tópico 0 (Educação/Ensino), maior núcleo do corpus, apresenta composição documental diversificada (Tabela 2).

**Tabela 2** - Composição tipológica do Tópico 0 (Educação/Ensino)

| Tipo documental | N | % |
|---|---|---|
| ARTIGO DE PERIÓDICO | 487 | 30,7 |
| DISSERTAÇÃO DE MESTRADO | 418 | 26,4 |
| TESE DE DOUTORADO | 221 | 13,9 |
| CAPÍTULO DE LIVRO | 179 | 11,3 |
| MONOGRAFIA DE ESPECIALIZAÇÃO | 118 | 7,4 |
| TRABALHO EM EVENTO | 104 | 6,6 |
| LIVRO | 59 | 3,7 |
| **TOTAL** | **1.586** | **100** |

**Fonte**: elaboração própria (2026).

Este perfil evidencia que a Educação na UFMG atua como um núcleo epistêmico-pedagógico, integrando a formação de alto nível (com produções no mestrado e doutorado, representando 40,3%), a disseminação científica regular (30,7%) e uma produção editorial consolidada (15%). Tal configuração tipológica sugere relativa maturidade institucional na articulação entre processos formativos, produção científica e regimes de circulação do conhecimento.

Este perfil evidencia que o agrupamento Educação/Ensino combina diferentes modalidades de produção acadêmica: produções de mestrado e doutorado representam 40,3%, artigos de periódico correspondem a 30,7%, enquanto capítulos de livro e livros somam 15,0%. A distribuição sugere que esse agrupamento articula produtos associados à formação pós-graduada, à comunicação periódica e à circulação editorial do conhecimento.

## 4.3 VALIDAÇÃO E ESTABILIDADE DO MODELO: COERÊNCIA SEMÂNTICA INTRA-TÓPICO

A compactação semântica global, estimada pela similaridade cosseno média entre os documentos e os centróides de seus respectivos tópicos, atingiu o índice de 0,671. Esse resultado indica proximidade semântica moderada entre os documentos e os centros vetoriais de seus agrupamentos no conjunto analisado. Os tópicos com maior similaridade ao centróide (Tabela 3) apresentam maior homogeneidade interna no espaço vetorial.

**Tabela 3** - Tópicos com maior similaridade média ao centróide

| Tópico | n_docs | Similaridade média |
|---|---|---|
| 35 | 81 | 0,798 |
| 29 | 103 | 0,775 |
| 26 | 109 | 0,775 |
| 44 | 64 | 0,748 |
| 36 | 75 | 0,740 |

**Fonte**: elaboração própria (2026).

Em sentido inverso, os tópicos apresentados na Tabela 4 registraram os menores valores de similaridade média ao centróide, variando entre 0,584 e 0,607.

**Tabela 4** - Tópicos com menor similaridade média ao centróide

| Tópico | Nº docs. | Similaridade média |
|---|---|---|
| 25 | 115 | 0,584 |
| 10 | 336 | 0,587 |
| 24 | 119 | 0,604 |

| | | |
|---|---|---|
| 6 | 513 | 0,605 |
| 32 | 94 | 0,607 |

**Fonte**: elaboração própria (2026).

Esses resultados indicam maior dispersão semântica interna nesses agrupamentos em comparação com os tópicos apresentados na Tabela 3, sem que isso implique, por si só, inadequação da solução temática.

### 4.3.1 Estabilidade estrutural (ARI e NMI)

A robustez da arquitetura foi testada via comparação de partições sob diferentes parametrizações de granularidade (*min_topic_size* = 30 e = 50). Os resultados de *Adjusted Rand Index (ARI)* e *Normalized Mutual Information (NMI)* são apresentados na Tabela 5.

**Tabela 5** - Indicadores de estabilidade estrutural

| Métrica | Valor |
|---|---|
| ARI | 0,9325 |
| NMI | 0,9539 |

**Fonte**: elaboração própria (2026).

O índice ARI ≈ 0,93 indica concordância significativa entre as atribuições documentais produzidas pelas duas parametrizações, enquanto o NMI ≈ 0,95 evidencia quantidade consistente de informação compartilhada entre as respectivas estruturas de agrupamento. Em conjunto, os resultados indicam estabilidade representativa da solução temática frente às variações de parametrização examinadas, reforçando sua consistência interna.

### 4.3.2 Validação supervisionada e poder preditivo

Para avaliar se os tópicos identificados capturam dimensões associadas à estrutura documental do corpus, realizou-se uma tarefa supervisionada de classificação tipológica utilizando Regressão Logística. Foram comparados dois modelos: um baseline composto exclusivamente pelos *embeddings* semânticos dos documentos e um modelo expandido, incorporando adicionalmente as variáveis de tópico dominante e ano de publicação. Os resultados de acurácia são apresentados na Tabela 6.

**Tabela 6** - Desempenho na classificação tipológica

| Modelo | Acurácia |
|---|---|
| EMBEDDINGS *(BASELINE)* | 0,5083 |
| EMBEDDINGS + TÓPICO + ANO | 0,5133 |

**Fonte**: elaboração própria (2026).

A inclusão das variáveis de tópico dominante e ano de publicação elevou a acurácia de 0,5083 para 0,5133, correspondendo a um incremento absoluto de 0,005, ou 0,5 ponto

percentual. Dada sua pequena magnitude, esse resultado não sustenta, isoladamente, a existência de ganho preditivo substantivo decorrente das variáveis adicionais. A análise deve ser interpretada como uma avaliação exploratória complementar, indicando contribuição incremental limitada do tópico e da temporalidade para a discriminação do tipo documental. O desempenho também sugere que parcela relevante da informação utilizada pelo classificador já se encontra incorporada às representações semânticas dos documentos.

### 4.4 IMPLICAÇÕES INFORMACIONAIS DA MODELAGEM SEMÂNTICA

Os resultados indicam que arquiteturas baseadas em *embeddings* semânticos e modelagem de tópicos podem contribuir para análise temática de grandes repositórios científicos institucionais. A diferenciação observada entre agrupamentos associados predominantemente a artigos, dissertações e teses sugere que a modelagem capturou regularidades semânticas que se articulam à composição documental da produção científica universitária.

Do ponto de vista da Ciência da Informação, os achados reforçam o potencial da mineração semântica para aplicações em organização da informação, monitoramento temático e inteligência científica. A identificação de agrupamentos associados à COVID-19 também evidencia a capacidade da abordagem para tornar observáveis concentrações temáticas relacionadas a eventos de crucial repercussão científica e social no período analisado. Em consonância com a fundamentação adotada, esses agrupamentos devem ser compreendidos como representações analíticas derivadas das propriedades textuais e documentais da coleção, e não como substitutos das estruturas formais de organização do conhecimento empregadas pelo repositório.

## 5 CONSIDERAÇÕES FINAIS

Este estudo aplicou uma arquitetura de mineração semântica à produção científica depositada no Repositório Institucional da Universidade Federal de Minas Gerais entre os anos de 2021 e 2025, buscando examinar sua organização temática, distribuição tipológica e estrutura semântica. A integração entre *embeddings* semânticos baseados em *Sentence-BERT*, redução de dimensionalidade com UMAP, agrupamento por densidade com *HDBSCAN* e modelagem de tópicos com *BERTopic* permitiu estruturar um *pipeline* analítico voltado à exploração de grandes coleções documentais institucionais.

Os resultados indicaram que a arquitetura foi capaz de identificar 48 agrupamentos temáticos substantivos, com consistência interna significativa e estabilidade frente às variações de parametrização examinadas, organizados principalmente em macrocampos relacionados à educação, saúde, ciências naturais e humanidades. A modelagem também possibilitou identificar diferenciações entre regimes documentais da produção científica institucional, evidenciando variações na distribuição de artigos, dissertações, teses e demais formatos acadêmicos entre os agrupamentos temáticos.

Do ponto de vista metodológico, o estudo evidencia a viabilidade da combinação entre representações semânticas baseadas em *Transformers*, redução de dimensionalidade, agrupamento por densidade e modelagem de tópicos para análise informacional de repositórios científicos institucionais. A abordagem permitiu tornar observáveis regularidades temáticas e documentais presentes na coleção, oferecendo uma camada analítica complementar às formas institucionalizadas de organização e representação da informação. Nesse sentido, os resultados apontam possibilidades de aplicação relacionadas à inteligência científica, monitoramento temático e gestão da informação científica.

Como limitações, destacam-se a dependência da qualidade dos metadados depositados no repositório e a utilização exclusiva da dimensão textual dos documentos analisados. Acrescenta-se a utilização do modelo *all-MiniLM-L6-v2*, desenvolvido predominantemente para língua inglesa, em uma coleção com forte presença de documentos em português, aspecto que pode afetar a fidelidade de determinadas relações semânticas produzidas no espaço vetorial. As métricas de compactação e estabilidade empregadas avaliam a consistência interna dos agrupamentos obtidos, não constituindo validação específica da adequação linguística do modelo ao corpus. Investigações futuras podem comparar os resultados com modelos multilíngues ou especificamente ajustados ao português, além de integrar análise semântica, redes de colaboração científica e modelagem temporal, permitindo examinar simultaneamente dimensões temáticas, relacionais e evolutivas da produção científica institucional.

## REFERÊNCIAS

ALVARENGA, Lídia. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 8, n. 15, p. 18-40, 2003. Disponível

em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2003v8n15p18?utm_source=chatgpt.com. Acesso em: 5 nov. 2026.

BÖRNER, Katy; SCHARNHORST, Andrea. Visual conceptualizations and models of science. **Journal of Informetrics**, Amsterdam, v. 3, n. 3, p. 161-172, 2009. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1751157709000315. Acesso em: 12 maio 2026.

BOWKER, Geoffrey C.; STAR, Susan Leigh. **Sorting things out**: classification and its consequences. Cambridge: MIT Press, 1999.

BRÄSCHER, Marisa; CAFÉ, Lígia. Organização da informação ou organização do conhecimento? *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo. **Anais** [...]. São Paulo: ANCIB, 2008. p. 1-14. Disponível em: https://cmapspublic.ihmc.us/rid=1KR7TM7S9-S3HDKP-5STP/BRASCHER%20CAF%C3%89(2008)-1835.pdf. Acesso em: 5 nov. 2026.

BUCKLAND, Michael K. Information as thing. **Journal of the American Society for Information Science**, Hoboken, v. 42, n. 5, p. 351-360, 1991. Disponível em: https://ppggoc.eci.ufmg.br/downloads/bibliografia/Buckland1991.pdf. Acesso em: 10 maio 2026.

CAFÉ, Lígia; MELO, Bianca Amaro de; BARBOZA, Elza Maria Ferraz; NUNES, Eny Marcelino de Almeida *et al*. Repositórios institucionais: nova estratégia para publicação científica na Rede. *In*: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 26., 2003, Belo Horizonte. **Anais**: Congresso Brasileiro De Ciências Da Comunicação. Belo Horizonte: INTERCOM, 2003. p. 1-12. Disponível em: http://www.intercom.org.br/papers/nacionais/2003/www/pdf/2003_ENDOCOM_TRABALHO_cafe.pdf. Acesso em: 12 maio 2026.

CAMPELLO, Ricardo José Gabrielli Barreto; MOULAVI, Davoud; SANDER, Jörg. Density-based clustering. **Wiley Interdisciplinary Reviews**: Data Mining and Knowledge Discovery, Hoboken, v. 10, n. 2, e1343, 2020. Disponível em: https://findresearcher.sdu.dk/ws/files/171664269/widm.1343.pdf. Acesso em: 12 maio 2026.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: pre-training of deep bidirectional transformers for language understanding. *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2019, Minneapolis. **Anais** [...]. Stroudsburg: Association for Computational Linguistics, 2019. p. 4171-4186. Disponível em: https://aclanthology.org/N19-1423/. Acesso em: 12 maio 2026.

FORTUNATO, Santo; BERGSTROM, Carl T.; BÖRNER, Katy; EVANS, James A. *et al*. Science of science. **Science**, Washington, DC, v. 359, n. 6379, art. eaao0185, 2018. Disponível em: https://pubmed.ncbi.nlm.nih.gov/29496846/. Acesso em: 12 maio 2026.

GROOTENDORST, Maarten. **BERTopic**: neural topic modeling with a class-based TF-IDF procedure. Ithaca: arXiv, 2022. Disponível em:

https://www.semanticscholar.org/reader/5493aefc07e331045be76003ea516aef57f76cb9. Acesso em: 8 abr. 2026.

GUIMARÃES, José Augusto Chaves. Análise de domínio como perspectiva metodológica em organização da informação. **Ciência da Informação**, Brasília, v. 43, n. 1, p. 13-21, 2014. Disponível em: https://doi.org/10.18225/ci.inf.v43i1.1415?utm_source=chatgpt.com. Acesso em: 5 de nov. 2026.

HJØRLAND, Birger. Domain analysis in information science: eleven approaches: traditional as well as innovative. **Journal of Documentation**, Bingley, v. 58, n. 4, p. 422-462, 2002. Disponível em: https://doi.org/10.1108/00220410210431136?utm_source=chatgpt.com. Acesso em: 5 nov. 2026.

LEITE, Fernando César Lima. **Como gerenciar e ampliar a visibilidade da informação científica brasileira**: repositórios institucionais de acesso aberto. Brasília: IBICT, 2009. 124 p. Disponível em: http://www.ibict.br/anexos_noticias/repositorios.institucionais.F.Leite_atualizado.pdf. Acesso em: 12 maio 2026.

MCINNES, Leland; HEALY, John; MELVILLE, James. **UMAP**: uniform manifold approximation and projection for dimension reduction. Ithaca: arXiv, 2018. Disponível em: https://arxiv.org/abs/1802.03426. Acesso em: 12 maio 2026.

MOUTINHO, Sônia Oliveira Matos; MARTINS, Paulo George Miranda; ALENCAR, Danila Fernandes; CONEGLIAN, Caio Saraiva. Ciência da Informação e ciência de dados: convergências interdisciplinares. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 29, p. 1-26, 2024. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/99127. Acesso em: 12 maio 2026.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: sentence embeddings using Siamese BERT-networks. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 9., 2019, Hong Kong. **Anais** [...]. Stroudsburg: Association for Computational Linguistics, 2019. p. 3982-3992. Disponível em: https://aclanthology.org/D19-1410/. Acesso em: 12 maio 2026.

SENTENCE TRANSFORMERS. **all-MiniLM-L6-v2**. [*S. l.*]: Hugging Face, 2024. Disponível em: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Acesso em: 9 set. 2026.

SHINTAKU, Milton; SUAIDEN, Emir José. Repositório institucional como componente de sistemas de informação gerencial para universidades. **BIBLOS**: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande, v. 29, n. 1, 2015. Disponível em: https://periodicos.furg.br/biblos/article/view/4905. Acesso em: 11 maio 2026.

SHINTAKU, Milton; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Bibliotecas e repositórios no processo de publicação digital. **BIBLOS**: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande, v. 30, n. 1, 2016. Disponível em: https://periodicos.furg.br/biblos/article/view/5762. Acesso em: 12 maio 2026.

SILVA, Edna Lúcia da; CAFÉ, Lígia; CATAPAN, Araci Hack. Os objetos educacionais, os metadados e os repositórios na sociedade da informação. **Ciência da Informação**, Brasília, v. 39, n. 3, p. 94-104, 2010. Disponível em: http://revista.ibict.br/ciinf/article/view/1269/1447. Acesso em: 10 mar. 2026.

VASWANI, Ashish; HAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob *et al*. Attention is all you need. *In*: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2017, Long Beach. **Proceedings Advances In Neural Information Processing Systems**. Red Hook: Curran Associates, 2017. p. 5998-6008. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Acesso em: 12 maio 2026.

WU, Zhaohui; CHIN, Ooi Beng. From big data to data science: a multi-disciplinary perspective. **Big Data Research**, Amsterdam, v. 1, p. 1, ago. 2014. Disponível em: https://www.sciencedirect.com/science/article/pii/S2214579614000082. Acesso em: 12 maio 2026.