Artículo Original

Análise de conteúdos curriculares com BERTopic numa perspectiva interdisciplinar

Analysis of Curriculum Contents with BERTopic from an Interdisciplinary Perspective

Antonio Miguel Faustini Zarth[1], Eliseo Reategui[2]

[1] Instituto Federal de Santa Catarina, Garopaba/SC, Brasil
[2] Universidade Federal do Rio Grande do Sul, Porto Alegre/RS, Brasil

miguel.zarth@ifsc.edu.br, eliseoreategui@gmail.com

Recibido: 10/02/2024 | **Aceptado:** 07/06/2024

Cita sugerida: A.M. Faustini Zarth, E. Reategui, "Análise de conteúdos curriculares com BERTopic numa perspectiva interdisciplinar," *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, no. 39, pp. 148-157, 2024. doi:10.24215/18509959.39.e15.

Resumo

A interdisciplinaridade desafia os educadores a repensar suas abordagens educacionais, visando integrar conhecimentos e métodos de diferentes disciplinas para promover uma compreensão mais holística e significativa dos tópicos de estudo. Este trabalho tem como objetivo utilizar e avaliar a técnica BERTopic na modelagem de tópicos, clusterização e classificação, empregando um projeto de curso integrado ao ensino médio como corpus, visando pontos de convergência entre unidades curriculares. Após análise exploratória dos dados, os resultados mostraram-se promissores na identificação de informações que podem fornecer uma melhor compreensão dos componentes curriculares, destacando as relações entre diferentes disciplinas. Tais resultados também podem auxiliar no desenvolvimento de novas metodologias de ensino que ressaltem o potencial da abordagem interdisciplinar na educação.

Palavras-chave: BERTopic; Modelagem de tópicos; Currículo de curso; Interdisciplinaridade.

Abstract

Interdisciplinarity challenges educators to rethink their educational approaches, aiming to integrate knowledge and methods from different disciplines to promote a more holistic and meaningful understanding of study topics. This work aims to use and evaluate BERTopic in topic modeling, clustering, and classification, employing an integrated high school course project as the corpus, seeking points of convergence between curriculum units. After an exploratory data analysis, the results showed promise in identifying information that can provide a better understanding of curricular components, highlighting the relationships between different disciplines. Such results can also help in the development of new teaching methodologies that highlight the potential of an interdisciplinary approach in education.

Keywords: BERTopic; Topic Modeling; Course Curriculum; Interdisciplinarity.

1. Introdução

Nos últimos anos, a aplicação de técnicas de modelagem de tópicos avançou significativamente no campo da inteligência artificial, particularmente no âmbito do processamento de linguagem natural (NLP). Essas técnicas oferecem uma abordagem sistemática para identificar temas e tópicos em coleções de dados textuais. Tradicionalmente, métodos como Alocação Latente de Dirichlet [1] e Fatorização de Matriz Não-Negativa [2] têm sido prevalentemente usados nesse domínio. Uma abordagem mais recente, o BERTopic [3] trouxe uma nova perspectiva à modelagem de tópicos. Este método facilita o agrupamento de documentos semelhantes, identificando dinamicamente o número ideal de tópicos e eliminando a necessidade de especificação prévia deste valor, como era requerido por métodos anteriores.

Na área da educação, a modelagem de tópicos tem encontrado aplicação em uma variedade de contextos. Estes incluem a organização de textos educacionais [4], a análise de relações entre pesquisa e ensino [5], bem como a complementação de técnicas mais tradicionais de revisão de literatura [6]. Esses estudos ressaltam o potencial da modelagem de tópicos na área educacional. Na pesquisa aqui apresentada, a modelagem de tópicos foi empregada para identificar pontos de convergência entre as unidades curriculares de um mesmo curso. Essa abordagem se baseou na premissa de que a avaliação da interdisciplinaridade e da fragmentação curricular é crucial para verificar se os objetivos delineados nos projetos de curso estão sendo adequadamente alcançados. A relevância da integração entre disciplinas para aprofundar o conhecimento é corroborada em estudos com resultados positivos tanto para docentes [7] quanto para discentes [8].

Além das práticas e atividades metodológicas que permeiam conteúdos interdisciplinares, a organização de um currículo de curso não pode ser feita fundamentalmente como uma grande colcha de retalhos. Neste contexto, o problema que se coloca reside na especificidade das disciplinas e dos saberes dos professores em suas respectivas áreas. Promover atividades interdisciplinares de modo a facilitar a integração curricular, não é uma tarefa simples. Com a predominância de currículos tradicionais, com apenas um professor em sala e com tempo de planejamento escasso, vencer este desafio exigiria uma intersecção de conhecimento muitas vezes oculta entre os pares, na ótica dos formadores. No entanto, Sobrinho [9] corrobora a ideia de que, mesmo em um currículo tradicional por disciplinas, é possível proporcionar a integração, contanto que alunos e professores colaborem para atingir esse propósito. Para isso, é necessário um olhar sobre como podem ser construídas essas relações integradoras. É neste cenário que este estudo se apresenta.

No âmbito da Inteligência Artificial, há uma diversidade de aplicações na Educação, conforme destacado na revisão de literatura conduzida por Chen et al. [10]. Os autores descrevem aplicações como plataformas de educação online, robôs, personalização de currículos e melhorar a qualidade geral das experiências de aprendizagem. No entanto, ainda são escassos os estudos que demonstram como essas ferramentas podem ser úteis para encontrar e explicitar as relações integradoras e interdisciplinares no âmbito da educação, em cursos específicos. Neste contexto, cabe destacar o estudo de Schettini et al. [11]. As pesquisadoras utilizaram mineração de texto através da ferramenta *Voyant* para avaliar o grau de interdisciplinaridade do curso de Relações Internacionais da USP e elaborar mapas de frequências de palavras no currículo. O reduzido grau de interdisciplinaridade apresentado nos resultados das autoras pode estar relacionado justamente às limitações das técnicas de frequência e coocorrência de palavras utilizadas pela ferramenta, pois desconsideram as relações semânticas das palavras em um corpus pequeno. Já Wang et al. [12] apresentaram um framework para encontrar tópicos interdisciplinares em artigos científicos utilizando como ferramenta central o BERTopic. Os autores utilizaram um amplo corpus de 70.384 artigos da base Web of Science (WoS), apresentando resultados empíricos, porém consistentes. Em um projeto de curso, no entanto, o conteúdo programático das disciplinas frequentemente se caracteriza por um vocabulário diversificado, mas ao mesmo tempo extremamente condensado. Muitas vezes, as unidades curriculares são resumidas em palavras-chave que representam o conhecimento a ser adquirido, o que resulta em um corpus escasso, heterogêneo e de utilidade limitada para as técnicas tradicionais de mineração de texto.

Neste contexto, o objetivo deste trabalho é realizar uma análise exploratória da utilização do BERTopic para clusterização e modelagem de tópicos, tendo como base de treinamento os conteúdos dispostos em um único projeto pedagógico de um curso técnico integrado ao ensino médio. Essa análise visa explorar e avaliar *insights* sobre as relações integradoras e interdisciplinares do documento, que podem ser utilizadas para melhorar a qualidade do ensino e da aprendizagem, além da compreensão e aprimoramento do currículo acadêmico por gestores e professores.

2. Ensino Integrado e Interdisciplinaridade

O ensino técnico integrado, cujo projeto de curso é objeto deste estudo, é uma abordagem educacional que busca a indissociabilidade entre educação básica e educação profissional, promovendo um aprendizado mais abrangente e contextualizado. Seu principal objetivo é promover a articulação entre formação geral e conhecimentos técnicos, superando a tradicional dicotomia entre trabalho manual e trabalho intelectual, vislumbrando a formação humana integral [13]. Frequentemente na prática, esse formato de curso se torna uma composição complexa, com lacunas de saberes, “apresentando fronteiras quase intransponíveis entre os seus componentes curriculares” [14]. No entanto, quando componentes curriculares podem cooperar entre si, potencializam seu objetivo comum e contrapõem esta

fragmentação. A interdisciplinaridade surge assim como elemento fundamental.

Conforme indicado por Erez [15], a interdisciplinaridade não possui uma definição única e universalmente aceita. Porém, através de uma investigação sistemática em 90 artigos especializados, identifica dois conceitos mais recorrentes relacionados ao tema, corroborados pelo trabalho de Souza et al. [16]. Estes estudos definem dois pilares conceituais sobre interdisciplinaridade: a primeira abordagem concentra-se principalmente na dimensão epistemológica. Nela, enfatiza-se a precisão conceitual e terminológica, com o objetivo de estabelecer uma hierarquia na integração das disciplinas, que vai desde o nível pluri e multidisciplinar até o inter e transdisciplinar. Esta abordagem tem como preocupação “a geração de conhecimento novo e os obstáculos que a fragmentação disciplinar pode criar nesse sentido” [16]. Na segunda abordagem, a interdisciplinaridade pode ser entendida como uma colaboração ou interação entre profissionais de diversas áreas, onde essas disciplinas estabelecem uma relação de reciprocidade, beneficiando-se mutuamente. Isso cria um senso de copropriedade que facilita o diálogo entre os envolvidos.

Para alcançar uma colaboração intrínseca entre diferentes áreas em um projeto de curso visando à interdisciplinaridade curricular, é essencial que os envolvidos sejam capazes de explorar as fronteiras de cada disciplina, estabelecer conexões entre elas e, por fim, identificar quais conhecimentos podem ser facilitadores desse processo. Nesse sentido, a modelagem de tópicos emerge como uma ferramenta valiosa, pois permite a sistematização, classificação e identificação de padrões e tendências em textos.

3. Modelagem de Tópicos com BERTopic

A partir destes pressupostos teóricos sobre interdisciplinaridade, foi escolhido o BERTopic como ferramenta para se alcançar os objetivos deste estudo. Sua escolha está motivada na capacidade do BERTopic em superar modelos clássicos, como o LDA, ao se trabalhar com textos curtos ou com relações hierárquicas de tópicos [3]. Como o nome sugere, este modelo utiliza o framework SBERT (*Sentence Bidirectional Encoder Representations from Transformers*) [17] para codificar os documentos em vetores de palavras (*word embeddings*) como etapa da extração de tópicos. Ele é *open-source*, modular, e permite múltiplas configurações e combinações de técnicas e algoritmos.

De forma geral, o BERTopic utiliza vetores de palavras pré treinadas por modelos transformadores (como o SBERT) para capturar a relação semântica entre palavras, além de combinar técnicas de redução de dimensionalidade, *clustering* e aplicar uma variação do algoritmo TF-IDF (*Term Frequency – Inverse Document Frequency*) baseadas em classes, o c-TF-IDF, para ponderar a importância das palavras e selecionar os tópicos que representam os documentos [16]. As etapas do algoritmo e o pipeline do BERTopic configurado para este estudo está descrito na figura 1.

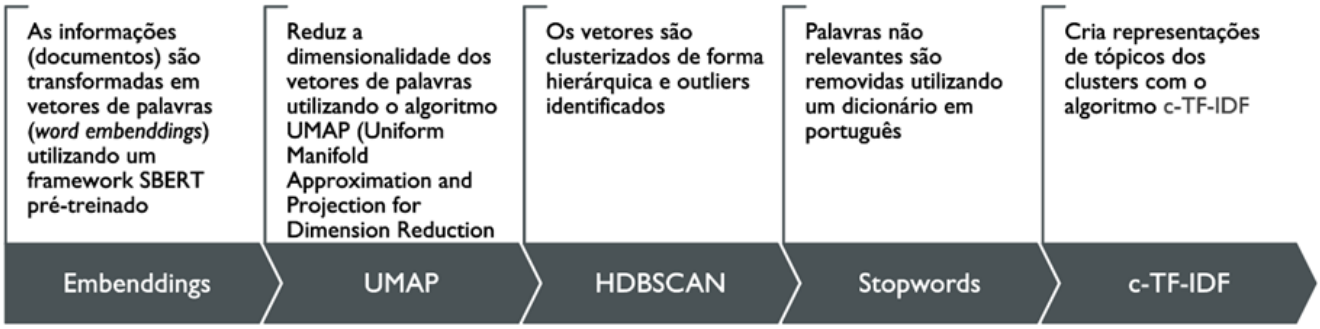

Figura 1. Pipeline BERTopic

O BERTopic tem sido utilizado em diversos contextos, como, por exemplo, para a extração de informações relevantes sobre as tendências de distintos segmentos de uma empresa a partir de dados de notícias [18]. Nestes estudos, o BERTopic foi capaz de extrair tópicos de dados textuais e avaliar sua coerência e diversidade, agrupando documentos e identificando valores atípicos.

A utilização do BERTopic para modelagem de tópicos tem apresentado resultados promissores também em documentos de textos curtos [19], assim como o corpus desta pesquisa, indicando o potencial de ser explorado em campos diversificados de estudo, entre eles a Educação. As próximas seções deste trabalho detalharão sua utilização no contexto da pesquisa.

4. Materiais e Métodos

O documento escolhido para ser avaliado no estudo envolvendo o BERTopic foi um Projeto Pedagógico de Curso (PPC) do Técnico em Informática Integrado ao Ensino Médio do Instituto Federal de Santa Catarina

(IFSC), Brasil. Um dos autores deste estudo participou da equipe de construção do documento e já atuou como coordenador do curso, experiência que contribuiu para uma melhor compreensão e interpretação dos resultados.

O PPC deste curso possui 44 componentes curriculares, sendo a maioria com duração de um ano. As disciplinas pertencem às áreas de conhecimento Linguagens, Códigos e suas Tecnologias, Ciências da Natureza e Matemática e suas Tecnologias, Ciências Humanas e Suas Tecnologias, além da formação específica em Informática. As disciplinas com nomes iguais, mas em anos diferentes (ex: Matemática I e Matemática II) foram agrupadas, consolidando um total de 27 componentes curriculares. Cada uma descreve os conhecimentos (conteúdo a ser abordado), além das habilidades e competências a serem desenvolvidas. Estas três informações (conhecimentos, habilidades e competências) foram extraídas de todo o documento, agrupadas, pré-processadas, e utilizadas como corpus. Os nomes das disciplinas foram descartados evitando assim uma pré-classificação induzida.

Na mineração de texto e modelagem de tópicos, os verbos muitas vezes podem não agregar informação relevante. Embora habilidades e competências remetam a ações, prevalecendo assim o uso de verbos, sua utilização foi considerada pela natureza de sua significação. Como exemplo, os verbos "programar", "diagnosticar", "implementar" e "ler" são próprios de áreas específicas e podem ser úteis na construção das relações de finalidade.

Para geração dos vetores (*embeddings*) do PPC, foi utilizada a base multilíngue pré-treinada "*paraphrase-multilingual-MiniLM-L12-v2*". Somente após a criação das incorporações, foram removidas as *stop words* da língua portuguesa, utilizando o *CountVectorizer* do BERTopic. Essa sequência é recomendada já que os modelos baseados em transformadores precisam de todo o contexto da frase para maior precisão.

Na geração e clusterização recursiva dos vetores de palavras evitando tópicos muito similares, foi utilizada a medida de similaridade dos cossenos, dado pela equação 1:

$$CS(T1,T2) = \frac{\sum_{i=1}^{n}(X_i \vee T_1).(Y_i \vee T_2)}{\sqrt{\sum_{i=1}^{n}(X_i \vee T_1)^2}.\sqrt{\sum_{i=1}^{n}(Y_i \vee T_2)^2}} \quad (1)$$

Equação 1

A equação indica quão similares são dois vetores de tópicos, sendo o valor 1 sendo iguais e 0 totalmente diferentes. X representa a palavra-chave do tópico T_1 enquanto Y a palavra-chave do tópico T_2 a ser comparado.

Para avaliar a qualidade dos tópicos gerados, foi utilizada a medida de coerência de tópicos (*c_v*) descrita em [3]. Essa métrica tem como objetivo avaliar a interpretabilidade e a coerência dos tópicos, quantificando o grau de similaridade semântica entre palavras dentro de cada tópico. Quanto maior a pontuação da coerência de tópicos, mais coeso e interpretável é o tópico. A coerência, cujo intervalo é de -1 a 1, foi avaliada utilizando a informação mútua pontual normalizada (NPMI, do inglês *normalized pointwise mutual information*).

Com o objetivo de avaliar a interdisciplinaridade e como as unidades curriculares compartilham e se apropriam dos tópicos gerados, cada conhecimento, habilidade e competência foi rotulada com o nome da respectiva disciplina e classificada pelo BERTopic frente ao conjunto de tópicos gerados anteriormente utilizando todo o PPC. Para avaliação das características de interdisciplinaridade dos tópicos, foi utilizada a equação proposta por Xu et al. [20], descrita na equação 2.

$$TI = d \times logtf \quad (2)$$

Equação 2

Esta equação quantifica o índice de interdisciplinaridade (*TI*) de cada tópico, onde *d* é a quantidade de disciplinas associadas ao tópico e *tf* representa a frequência de termos nos documentos (conteúdos) associados. Valores maiores de *TI* indicam mais interdisciplinaridade dos termos dos tópicos nas disciplinas

5. Resultados e Discussão

5.1. Identificação e avaliação dos tópicos

A quantidade de tópicos e palavras representativas encontradas pode variar a cada execução, devido à natureza estocástica do algoritmo de redução de dimensionalidade UMAP utilizado no pipeline do BERTopic. Desta forma, foi atribuído ao parâmetro *random_state* o valor 1 para que seja possível a replicação dos resultados. Nesta configuração, foram identificados 19 tópicos a partir dos conteúdos do PPC.

A figura 2 representa os 19 tópicos encontrados, destacando as 6 palavras mais significativas em cada um. A importância de cada palavra para o tópico é dada pelo score atribuído com o algoritmo c-TF-IDF.

Figura 2. Tópicos encontrados

A coerência (c_v), quantificada pelo modelo, foi de 0.76, indicando um valor positivamente relevante de coesão e interpretabilidade dos tópicos. Este resultado reforça a consistência e clareza das temáticas identificadas no conjunto de dados analisado.

A diversidade dos tópicos pode ser avaliada pela matriz de similaridade dos tópicos na figura 3. Essa matriz indica o quanto um tópico é similar a outro através da medida de similaridade de cossenos (equação 1). Quanto menos similar for um tópico com os demais, mais claro é ponto de cruzamento, gerando um mapa de calor. É possível interpretar com este mapa que não há tópicos excessivamente similares ou que poderiam ser considerados redundantes.

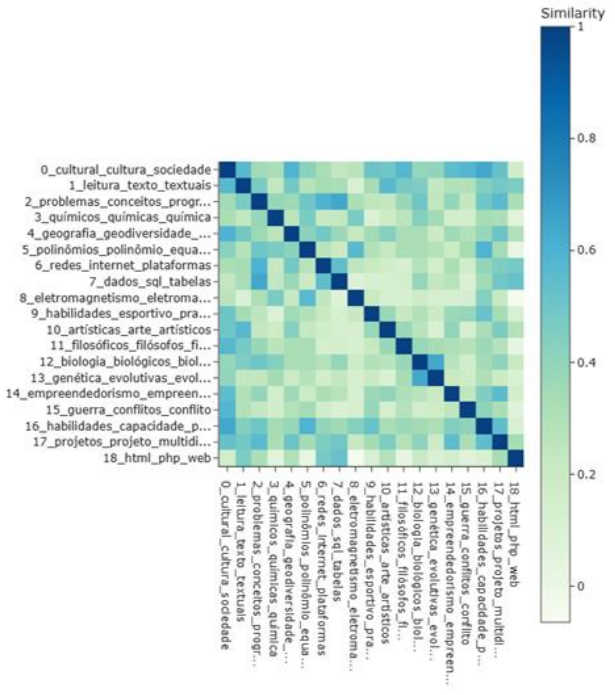

Figura 3. Matriz de Similaridade

5.2. Relações e hierarquia dos tópicos

O mapa de distância inter-tópicos (do inglês *intertopic distance map*) do BERTopic, corresponde a uma visualização que representa a distância entre os tópicos em um espaço bidimensional. Os tópicos mais semelhantes estão mais próximos uns dos outros. No contexto da interdisciplinaridade, esse mapa bidimensional é interessante para a identificação dos grupos de tópicos relacionados. Um grupo de tópicos próximos uns dos outros pode indicar que estejam relacionados ao mesmo assunto. O mapa também pode ser um facilitador para perceber a diversidade de conhecimentos e uma relação de hierarquia entre termos. Se houver um tópico que está no centro de um grupo de tópicos relacionados, isso pode indicar que esse é mais geral do que os outros tópicos no grupo. Por outro lado, o distanciamento dos agrupamentos pode indicar uma falta de coesão entre os assuntos abordados.

Para uma análise mais minuciosa deste mapeamento, foram identificados quatro aglomerados distintos na figura 4, cada um representando conjuntos de tópicos correlatos. Evidencia-se uma clara inclinação à proximidade entre tópicos que abordam áreas afins. Como exemplo, constata-se que tópicos relacionados à tecnologia tendem a se agrupar, assim como em outras áreas, indicando tendências facilitadoras para atividades interdisciplinares. Por se tratar de um curso técnico em informática integrado ao ensino médio, também fica evidente a dificuldade de integração entre a técnica e a propedêutica. O agrupamento na parte superior da figura 4 é essencialmente constituído de tópicos relacionados à formação profissional e mantém-se relativamente distante de tópicos relacionados à formação geral (demais agrupamentos). Essa dicotomia, embora passível de integração, reafirma como é desafiador o objetivo curricular de indissociabilidade entre educação básica e educação profissional.

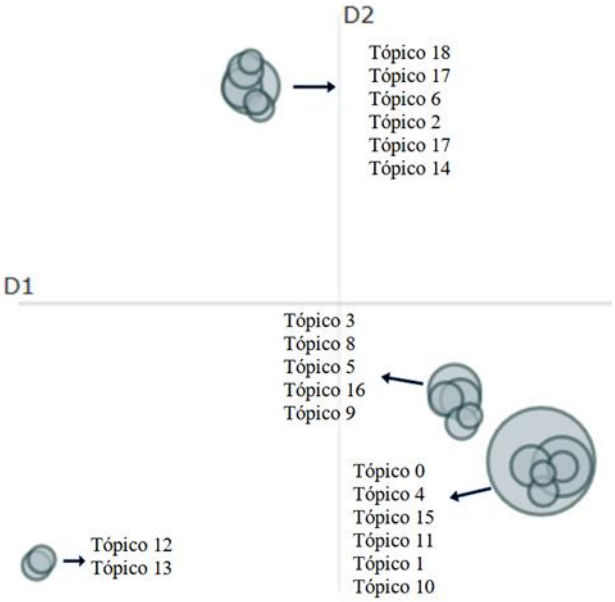

Figura 4: Mapa de distância entre tópicos

Para analisar a integração entre os tópicos e suas relações hierárquicas, também foi aplicado o método *visualize_hierarchy()* do BERTopic para os 19 tópicos obtidos, conforme ilustrado na figura 5. A hierarquia de tópicos é uma representação onde os tópicos mais gerais estão no topo da hierarquia e os mais específicos estão na base. Os tópicos próximos uns dos outros compartilham cores, indicando maior semelhança. Essa representação oferece *insights* sobre como os tópicos compõem o curso como um todo, revelando oportunidades para relações integradoras, interdisciplinares e a construção do conhecimento.

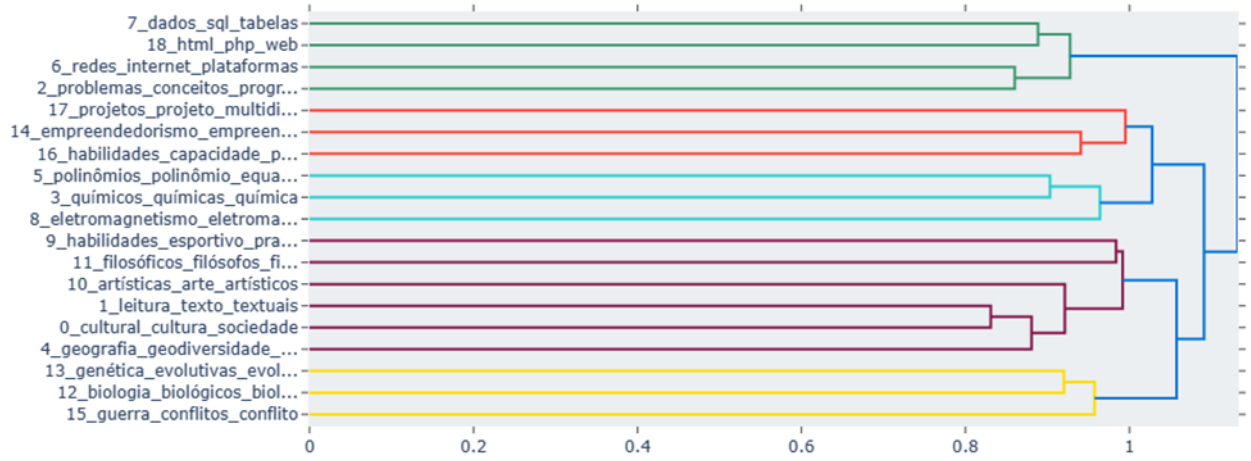

Figura 5. Estrutura hierárquica dos tópicos

Ciente dos tópicos que melhor representam os conteúdos de suas respectivas áreas, professores podem utilizar a estrutura hierárquica para perceber oportunidades na organização e análise de contextos relacionados a outras disciplinas, buscando áreas de convergência ou finalidade. Como exemplo, disciplinas que possuem como interesse conhecimentos relacionados a “dados” e “sql” (tópico 7), podem construir atividades interdisciplinares com conhecimentos de “html” ou “web” (tópico 18), o que, por sua vez, pode se integrar de maneira lógica e hierárquica com termos como “redes” (tópico 6) ou “programação” (tópico 2). Esta estrutura também pode contribuir para a gestão acadêmica, podendo ser uma ferramenta útil para a organização de currículos que fluem de maneira mais natural e interconectada.

5.3. Relações interdisciplinares revisitando o PPC

Com o objetivo de examinar as características interdisciplinares do PPC, foi realizada a classificação e avaliação das disciplinas, bem como de seus documentos correspondentes, levando em conta os tópicos identificados dentro de seu escopo. A tabela 1 exibe o Índice de Interdisciplinaridade (TI) dos 19 tópicos identificados, acompanhado da quantidade de disciplinas e a frequência de ocorrência dos termos no documento. Destaca-se o Tópico 0, caracterizado por termos como “sociedade” e “cultura”, que demonstra o mais elevado índice de interdisciplinaridade e integração, abrangendo 21 disciplinas distintas de um total de 27. A frequência dos termos associados a este tópico é de 268 ocorrências no PPC, um valor substancialmente superior em relação aos demais tópicos. Sob uma perspectiva pedagógica, o referido tópico, seus termos e suas significações, representam uma base interdisciplinar fundamental na elaboração do PPC. Por outro lado, o tópico 18 tendo como principais termos “html”, “web” e “php”, demonstrou ter o menor *TI*, sem relação interdisciplinar com diferentes disciplinas.

Tabela 1. Índice de Interdisciplinaridade

Tópico	Frequência	Disciplinas	TI
0	**268**	**21**	**50,99**
1	83	10	19,19
2	73	9	16,77
3	63	3	5,4
4	38	3	4,74
5	37	4	6,27
6	36	6	9,34
7	28	8	11,58
8	26	4	5,66
9	23	1	1,36
10	21	2	2,64
11	20	4	5,2
12	19	1	1,28
13	17	2	2,46
14	17	5	6,15
15	14	3	3,44
16	14	8	9,17
17	14	3	3,44
18	13	1	1,11

A Figura 6 apresenta a distribuição desses tópicos no conjunto de disciplinas investigadas. Esta visualização permite não apenas compreender a distribuição dos tópicos selecionados, mas também discernir sua relevância em cada unidade curricular, evidenciando aqueles que se destacam por sua frequência, abrangência ou proeminência. Abaixo dos nomes dos tópicos e das disciplinas, o valor da frequência de documentos que os relaciona é indicado. Tópicos que são consistentemente relevantes em várias disciplinas podem ser considerados fundamentais para a

educação interdisciplinar, servindo como pontos de conexão para que educadores desenvolvam suas atividades. A seguir serão discutidos alguns achados interessantes

.

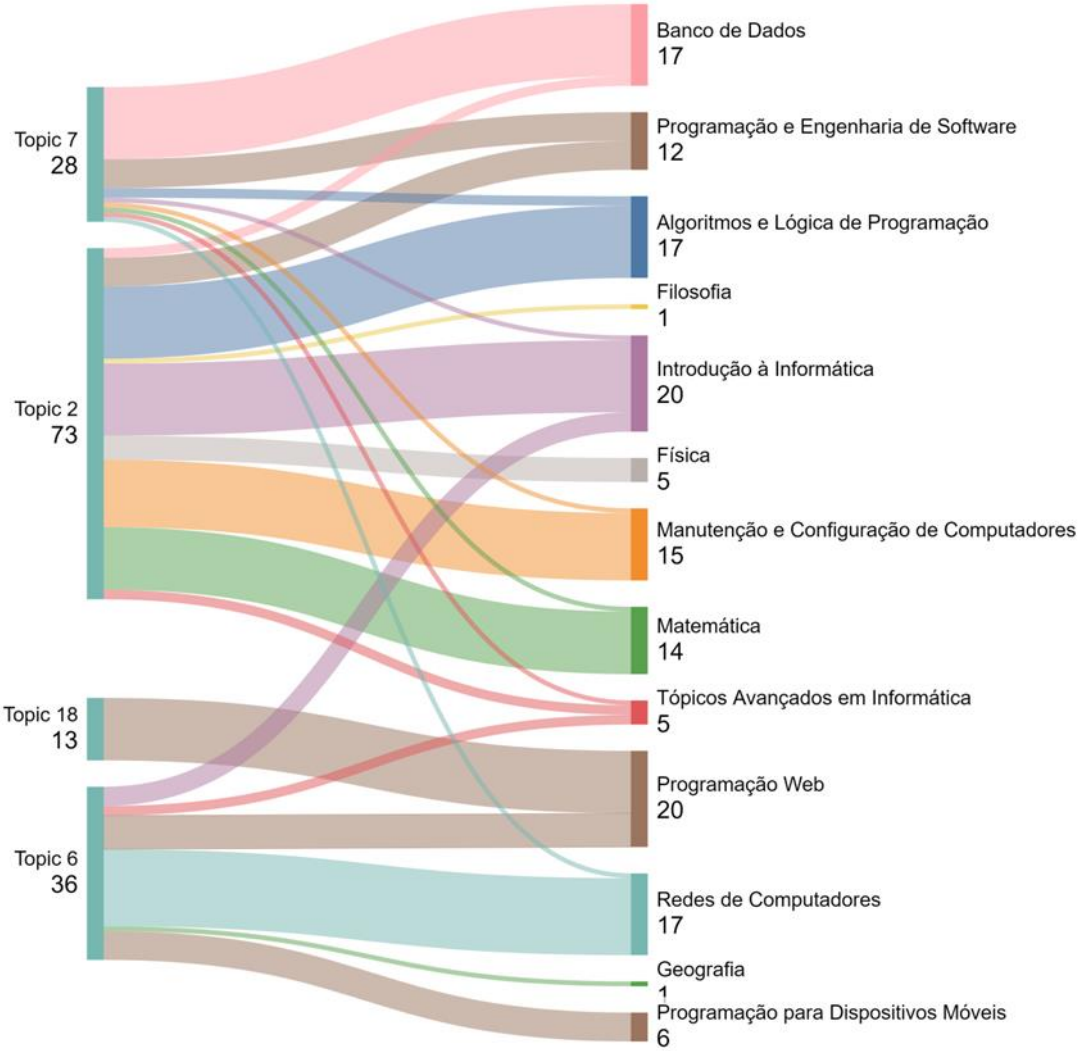

Figura 6. Frequência dos tópicos nas disciplinas

Para uma avaliação mais aprofundada das nuances e interconexões entre diferentes disciplinas e áreas do conhecimento, procedeu-se à seleção de quatro tópicos pertinentes ao âmbito profissionalizante (tópicos 2, 6, 7 e 18). A utilização de vetores de palavras (*word embeddings*) viabilizou a identificação de relações semânticas nos documentos, conferindo-lhes maior poder de generalização, em contraposição aos modelos de coocorrência estatística de palavras, cuja capacidade de abstração é limitada. O conjunto das 4 palavras mais significativas dos tópicos escolhidos pode ser revisitado na tabela 2.

Tabela 2. Representações dos tópicos (parcial)

Tópico	Termos
2	problemas, conceitos, programação e computacionais
6	redes, internet, plataformas e servidores
7	dados, sql, tabelas e software
18	html, php, web e css

O tópico 18 possui integração única com a disciplina “Programação Web”, já apontado como pouco interdisciplinar pelo valor de *TI* na tabela 1. No entanto, esta disciplina apresentou 35% de documentos (frequência 7) relacionadas também ao tópico 6, o que por sua vez, possui alta frequência nas disciplinas de “Redes” e

“Programação para Dispositivos Móveis”, além de uma frequência modesta na disciplina de “Geografia”.

O tópico 2, por sua vez, apresenta relação com 9 disciplinas, com distribuição de frequência significativa e homogênea em outras 4. Idealizando atividades que relacionam teoria, aplicação tecnológica e desenvolvimento, é possível perceber uma linha coerente e viável de integração entre estas disciplinas.

Por fim, o tópico 7 possui frequência predominantemente nas disciplinas de “Banco de Dados” e “Programação e Engenharia de Software”. No entanto, é possível utilizar os termos de sua representação para se encontrar pontos de convergência interdisciplinar em outras 6 disciplinas.

Com base nestes exemplos, os resultados demonstram evidências de que é possível encontrar referências sólidas para se repensar as fronteiras conceituais de cada disciplina. No entanto, para uma avaliação mais aprofundada e uma compreensão mais completa dos resultados em diferentes contextos, é importante incorporar uma metodologia de avaliação qualitativa que envolva professores, gestores e alunos como sujeitos da pesquisa, o que não fez parte do escopo deste trabalho.

Considerações Finais

O BERTopic é um algoritmo recente e a análise de suas potencialidades e limitações ainda é objeto de estudo em diversas áreas, incluindo a Educação. A modelagem de tópicos com o BERTopic, utilizando um projeto de curso como corpus e adotando a perspectiva de relações integradoras e, principalmente, interdisciplinares, demonstrou ser viável, com evidências de resultados sólidos. Essa é a principal contribuição e originalidade desta pesquisa. Estudos que aplicam IA e mineração de texto nesse contexto são escassos, podendo ser este trabalho uma referência inicial para pesquisadores em inteligência artificial e profissionais da educação interessados na curricularização da interdisciplinaridade.

Embora esta pesquisa tenha apresentado resultados promissores, é imperativo realizar estudos adicionais em outros PPCs e utilizando diferentes parâmetros e algoritmos de treinamento do BERTopic. A inexistência de uma base de dados pré-treinada específica para o contexto educacional em português brasileiro também pode ser um limitador para este gênero de estudo.

O modelo também identificou como *outliers* cerca de 25% dos documentos (conhecimentos, habilidades ou competências) do projeto de curso. Estudos mais específicos precisam ser feitos para a compreensão deste resultado. Estes podem estar relacionados à escrita do PPC ou a outro refinamento do modelo não contemplado nesta pesquisa. Pesquisas futuras também podem ser direcionadas na construção de bases de incorporação de palavras treinadas em modelos transformadores mais eficientes, assim como análise da metodologia aplicada em cursos de outros níveis de ensino, como graduação e pós-graduação.

Referências

[1] H. Jelodar, Y. Wang, C. Yuan, et al. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey", in Multimed Tools Appl, vol. 78, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4

[2] M. Luo, F. Nie, Chang, F., Yang, X., Hauptmann, Y. & Q. Zheng, "Probabilistic Non-Negative Matrix Factorization and Its Robust Extensions for Topic Modeling", in Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, 2017, doi: 10.1609/aaai.v31i1.10832

[3] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.

[4] A. Sakhovskiy, E. Tutubalina, V. Solovyev and M. Solnyshkina, "Topic Modeling as a Method of Educational Text Structuring," 2020 13th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, United Kingdom, 2020, pp. 399-405, doi: 10.1109/DeSE51703.2020.9450232.

[5] H. Lee, Kwak, H., Song, J. M. et al. "Coherence analysis of research and education using topic modeling", Scientometrics, vol. 102, pp. 1119–1137, 2015, doi: 10.1007/s11192-014-1453-x

[6] A. M. Grisales A., S. Robledo and M. Zuluaga, "Topic Modeling: Perspectives From a Literature Review," in IEEE Access, vol. 11, pp. 4066-4078, 2023, doi: 10.1109/ACCESS.2022.3232939.

[7] M. A. Salica, “Analítica del aprendizaje interdisciplinar con modalidad d-learning en contexto de COVID-19”, TEyET, n.º 31, p. e1, mar. 2022.

[8] L. de Lima, R. C. Loureiro, y G. Teles, “Interdisciplinaridade e tecnologias digitais na transformacao da compreensao de docencia”, TEyET, n.º 20, pp. p. 16–27, dic. 2017.

[9] S. C. Sobrinho, "Diretrizes institucionais e a perspectiva da integração curricular no IF Farroupilha," in Ensino médio integrado no Brasil: fundamentos, práticas e desafios, Brasília: Ed. IFB, 2017, pp. 106-140.

[10] L. Chen, P. Chen, Z. Lin, "Artificial intelligence in education: A review," IEEE Access, vol. 8, pp. 75264-75278, 2020.

[11] D. C. D. Schettini, M. A. D. T. Lins, M. Nishijima, "Interdisciplinaridade em Bacharelado no Brasil: O Caso de Relações Internacionais da USP," Revista de Graduação USP, vol. 3, no. 1, pp. 3-17, 2018.

[12] Z. Wang et al., "Identifying interdisciplinary topics and their evolution based on BERTopic," Scientometrics, pp. 1-26, 2023.

[13] CONIF, "Diretrizes Indutoras para a Oferta de Cursos Técnicos Integrados ao Ensino Médio na Rede Federal de Educação Profissional, Científica e Tecnológica," Fórum de Dirigentes de Ensino/CONIF, 2018.

[14] L. C. Silva-Pereira, J. R. A. Santos, M. G. Oliveira Neto, "Metodologias integradoras na educação profissional: construindo a ponte entre a Base Comum e as disciplinas técnicas no ensino técnico integrado," in Ensino médio integrado no Brasil: fundamentos, práticas e desafios, Brasília, DF: IFB, 2017.

[15] O. C. Erez, "O que é Interdisciplinaridade? Definições mais comuns em Artigos Científicos Brasileiros," Interseções: Revista de Estudos Interdisciplinares, vol. 20, no. 2, 2019. DOI: 10.12957/irei.2018.39041.

[16] M. A. de Souza et al., "Interdisciplinaridade e práticas pedagógicas: O que dizem os professores," Revista Portuguesa de Educação, vol. 35, no. 1, pp. 4-25, 2022.

[17] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks," in Proc. of the 2021 Conf. of the North American Chapter of the ACL (NAACL-HLT), pp. 2584-2596, 2021.

[18] H. Okazaki, H. Takahashi, "Nowcasting of Corporate Research and Development trends through news article analysis by BERTopic: the case of Japanese electric company," in 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), IEEE, 2022, pp. 1-6.

[19] M. de Groot, M. Aliannejadi, M. R. Haas, "Experiments on Generalizability of BERTopic on Multi-Domain Short Text," arXiv:2212.08459v1, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2212.08459

[20] Xu, H., Guo, T., Yue, Z., Ru, L., & Fang, S. (2016). Interdisciplinary topics of information science: a study based on the terms interdisciplinarity index series. Scientometrics, 106(2), 583-601. https://doi.org/10.1007/s11192-015-1792-2

Informação de Contato dos Autores:

Antonio Miguel Faustini Zarth
Garopaba/SC
Brasil
miguel.zarth@ifsc.edu.br
https://orcid.org/0009-0001-1318-6145

Eliseo Reategui
Porto Alegre/RS
Brasil
eliseoreategui@gmail.com
https://orcid.org/0000-0002-5025-9710

Antonio Miguel Faustini Zarth
Mestre em Ciências da Computação pela Universidade Federal de Pernambuco, é professor do Instituto Federal de Santa Catarina (IFSC), Brasil.

Eliseo Reategui
Doutor em Computação pela Universidade de Londres, Inglaterra (UCL), é professor da Faculdade de Educação da Universidade Federal do Rio Grande do Sul (UFRGS), Brasil.