Interdisciplinaridade no Enem: um estudo de caso das Ciências Humanas a partir do Processamento de Linguagem Natural

Ester Pereira Neves de Macedo

Inep
Brasília, DF, Brasil
ester.macedo@inep.gov.br

Flávia Ghignone Braga Ribeiro

Inep
Brasília, DF, Brasil
flavia.ribeiro@inep.gov.br

Alexandre Jaloto

Inep
Brasília, DF, Brasil
alexandre.jaloto@inep.gov.br

Danielle de Oliveira Costa

Inep
Brasília, DF, Brasil
danielle.costa@inep.gov.br

Resumo

Um dos pilares do Exame Nacional do Ensino Médio (Enem) é a interdisciplinaridade. Uma das possibilidades de se verificar a similaridade dos textos é a utilização de vetores de palavras, uma das técnicas do Processamento da Linguagem Natural (PLN). Neste trabalho objetivamos caracterizar a interdisciplinaridade na prova do Enem a partir do PLN. Para isso, analisamos a similaridade dos textos dos itens por meio de uma análise exploratória gráfica dos seus vetores numéricos, usando os itens da prova principal do Enem 2016. Utilizamos vetores de palavras treinados com o algoritmo Glove com 300 dimensões. Calculamos a média dos valores das palavras dos itens para cada dimensão e posicionamos cada item em um gráfico de duas dimensões. Por último, classificamos cada item de Ciências Humanas e suas Tecnologias (CHT) quanto ao componente curricular predominante. Os achados apontam para uma identidade tanto de cada área do Enem, quanto dos componentes curriculares.

1 Introdução

O Exame Nacional do Ensino Médio (Enem) foi criado em 1998 e, desde então, é aplicado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), com o objetivo de "aferir o desenvolvimento de competências fundamentais ao exercício pleno da cidadania" (BRASIL. INEP, 2002, p. 5). Observa-se que a avaliação de competências pressupõe uma abordagem contextualizada, que articule o conhecimento escolar e a realidade social, assim, em sua primeira versão, a elaboração do exame era orientada por uma matriz de referência única fundamentada na "colaboração, complementaridade e integração entre os conteúdos das diversas áreas do conhecimento" (BRASIL. INEP, 2002, p. 13).

Conforme Machado (2005), a ênfase na proposta interdisciplinar, apoiada na busca de superar a fragmentação do saber, alcançou certo consenso, mas diversos obstáculos relacionados à organização escolar tradicional, que é mais multidisciplinar, ainda se impunham sobre o seu pleno entendimento: "como mero incremento das relações entre as disciplinas, mantidos seus respectivos objetivos/objetos, e mantidas as relações determinadas pelo sistema que constituem, as ações interdisciplinares têm produzido efeitos apenas paliativos" (MACHADO, 2005, p. 52).

Em 2009, o Enem se consolidou como um dos principais meios de acesso ao ensino superior no país e passou por uma reformulação significativa, em que se destaca a elaboração de novas matrizes de referência, agora por área do conhecimento, que tinham o "objetivo de ampliar e evidenciar os objetos de conhecimento avaliados nas provas" (BRASIL. INEP, 2013, p. 13). Ressalta-se que não houve uma alteração na concepção pedagógica do exame, mas em um estudo posterior, que se detinha sobre as provas de Ciências Humanas (CH) do Enem até 2015, Lima (2017) observou que, a partir de 2009, houve "o avanço do processo de isolamento e aprofundamento das disciplinas dos conteúdos, teorias e recursos solicitados, cada vez mais voltados para áreas internas das disciplinas" (LIMA, 2017, p. 138), reconhecendo-se mais uma vez os desafios à abordagem interdisciplinar efetiva.

Diante desse cenário, se propõe a análise da interdisciplinaridade na prova do Enem de 2016, com o objetivo de registrar estratégias e tendências na construção dessa abordagem. Essa análise por sua vez será também interdisciplinar, aliando uma análise pedagógica baseada nas Ciências Humanas e suas Tecnologias (CHT) com uma análise computacional, a partir do processamento da linguagem natural (PLN).

O PLN é uma área da computação que integra um conjunto de técnicas de manipulação computacional da linguagem natural. A linguagem natural é aquela utilizada pelos humanos para se comunicarem no cotidiano. Em oposição a linguagens artificiais, como linguagens de programação, as linguagens naturais evoluem ao serem transmitidas pelas gerações e são difíceis de terem uma definição com regras explícitas (BIRD; KLEIN; LOPER, 2009).

As técnicas do PLN têm sido utilizadas para explicar a dificuldade dos itens segundo suas características linguísticas (BENEDETTO et al., 2020) e pontuar redações de alunos (PASSERO; FERREIRA; DAZZI, 2019). Para executar os métodos estatísticos e compreender os textos, os programas designam valores

numéricos para as palavras, que é a forma como elas são representadas (PRIMI, no prelo). Uma das formas de representar as palavras em números é na forma de vetores (MIKOLOV et al., 2013). Os vetores de palavras codificam informações de similaridades entre palavras a partir de sua coocorrência no cotidiano.

2 Objetivo

Caracterizar a interdisciplinaridade na prova do Enem por meio do PLN. Para isso, analisamos a similaridade dos textos dos itens do Enem 2016 a partir de uma análise exploratória gráfica dos seus vetores numéricos das palavras que compõem seus textos, acompanhado de uma análise pedagógica dos resultados obtidos, tendo como foco a área de CHT.

3 Metodologia

Para a realização da análise de similaridade, utilizamos vetores de palavras treinados para posicionar os itens em um gráfico de duas dimensões. As análises, baseadas na proposta de Primi (no prelo), foram realizadas a partir da preparação do texto e da obtenção dos vetores de palavras, seguidas da visualização gráfica dos itens. Utilizamos o ambiente R (R CORE TEAM, 2019) para este fim.

Na etapa de preparação do texto, foram utilizados os 180 itens da aplicação principal do Enem 2016. Primeiramente, dividimos os itens em suas palavras constituintes, ou seja, transformamos o texto de um item em uma tabela com uma coluna e “n” linhas, em que “n” corresponde à quantidade de palavras. Em seguida, retiramos as palavras duplicadas de cada item e mantivemos somente as palavras lexicais (substantivos, verbos, adjetivos e advérbios). A lista de palavras não lexicais foi obtida do pacote stopwords (BENOIT; MUHR; WATANABE, 2020).

Para a obtenção dos vetores de palavras treinados por Hartmann et al. (2017), recorremos ao Repositório de Word Embeddings disponibilizado pelo Núcleo Interinstitucional de Linguística Computacional (2017). O algoritmo utilizado para o treinamento foi o Glove, com 300 dimensões. Uma vez que cada item é formado por várias palavras, e cada palavra é formada por um vetor, calculamos a média das palavras para cada dimensão. Com isso, cada item foi representado por um único vetor de 300 dimensões. Por último, utilizamos o pacote Rtsne (VAN DER MAATEN; HINTON, 2008) para posicionar os itens em um gráfico de duas dimensões.

Em uma segunda etapa do estudo, selecionamos a área de CHT para uma análise pedagógica mais aprofundada da disposição dos itens no gráfico. Classificamos os quarenta e cinco itens de CHT da prova de 2016 considerando o(s) componente(s) curricular(es) predominantemente mobilizado(s) em cada um deles (Geografia, História, Sociologia e Filosofia). A partir da localização desses itens no gráfico, descrevemos como eles se articulam em termos de disciplinaridade e interdisciplinaridade nesta edição do exame.

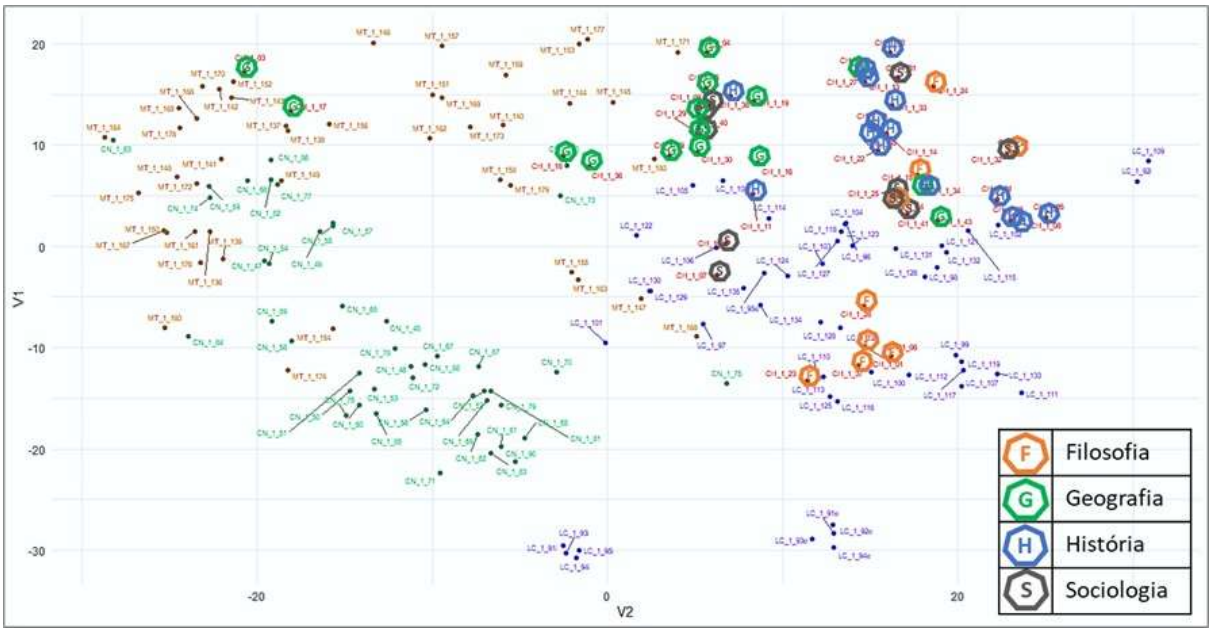
4 Resultados e discussão

A Figura 1 apresenta o gráfico com o posicionamento dos 180 itens da aplicação principal do Enem em 2016 em termos da similaridade de seus textos. Cada

cor dos pontos representa uma área do conhecimento, com destaque para os 45 itens de CHT, categorizados pelo componente curricular predominante: Geografia, História, Sociologia e Filosofia.

Embora haja pontos de intersecção entre as áreas, é possível reconhecer a predominância de uma das quatro áreas do conhecimento avaliadas no Enem em determinados setores do gráfico. Isso aponta para uma identidade nos textos dos itens de uma determinada área. De maneira semelhante, no que diz respeito à área de CHT, é possível também verificar a consistência da identidade dos componentes curriculares que integram esta área do conhecimento – Geografia, História, Filosofia e Sociologia – a partir da análise das localizações agregadas de seus itens e levando-se em consideração a categorização da cobertura predominante do componente curricular em cada um dos itens. É possível notar blocos de itens característicos de cada disciplina de maneira isolada, bem como um bloco com itens dos quatro componentes, num bloco mais interdisciplinar dentro da área de CHT. Constata-se assim que, embora a interdisciplinaridade encontre espaço no Enem, há blocos disciplinares ainda bastante consolidados.

Figura 1: **Dispersão dos itens do Enem 2016 segundo a similaridade de seus textos**



Nota: Os pontos menores denotam as quatro grandes áreas do conhecimento: a cor marrom corresponde a Matemática e suas Tecnologias (MT); a verde, a Ciências da Natureza e suas Tecnologias (CN); a cinza, a Linguagens e Códigos e suas Tecnologias (LC); e a vermelha, a Ciências Humanas e suas Tecnologias (CH). A etiqueta do ponto indica a área, a aplicação principal (1) e a posição do item no caderno azul.

Considerando de maneira mais aprofundada cada um dos componentes curriculares que compõem as Ciências Humanas e suas Tecnologias, embora os itens com temas próprios da Geografia estejam posicionados próximos aos demais itens de CHT, eles apresentam maior similaridade com os itens de Matemática e suas Tecnologias (MTT) do que com os de outras áreas. Destacam-se o uso de técnicas de medida e da geometria mobilizado para a localização e a compreensão de fenômenos físicos e naturais, a quantificação e proporcionalidade utilizadas no estudo

das escalas, representações, projeções e coordenadas cartográficas, bem como a mobilização de dados estatísticos para o estudo de fenômenos demográficos e sociais. No que diz respeito ao componente curricular de História, nota-se dois blocos de itens caracteristicamente associados a este componente, além de dois itens com características mais interdisciplinares. Há um grupo maior com itens relacionados a política e poder que se conectam com os outros três componentes curriculares de CHT. Há um segundo grupo de itens com predominância do componente História que abordam questões relacionadas ao imperialismo, com aproximação temática de um único item de LCT Entre os itens associados à história dispostos mais próximos aos itens associados à Geografia, temos itens relacionados à mundialização e à produção.

O componente curricular de Filosofia aparece de forma concentrada em um bloco de cinco itens mais próximos da área de LCT, e mais afastados dos outros componentes curriculares das CHT, com itens com maior ênfase na leitura de textos, Por outro lado, entre um bloco mais interdisciplinar de itens de CHT, encontramos mais alguns itens com temáticas associadas à filosofia próximos dos outros componentes de Ciências Humanas, em temas como a modernização da cultura e implicações da tecnologia na organização da sociedade e questões sobre filosofia política contemporânea. De forma similar, os itens associados ao componente curricular de Sociologia e, situados em geral internamente ao agrupamento de CHT, aproximam-se das outras três subáreas que compõem essa área de conhecimento, Para além da área de CHT, a aproximação mais frequente apresentada pela Sociologia é com a área de Linguagens, o que pode demonstrar uma ênfase em abordagens ancoradas na interpretação, seja do pensamento de autores reconhecidos na área, seja da realidade social, foco do exame, por meio de fontes atuais e diversas.

5 Conclusões e Considerações Finais

Na prova principal do Enem de 2016, observa-se uma identidade dos textos dos itens de determinada área, dada a sua similaridade. Dentro da área de CHT, os itens também apresentam uma identidade no que diz respeito aos componentes curriculares. Para essa área, as análises apontam para a predominância de blocos mais específicos para cada um dos quatro componentes, bem como para um conjunto de itens que abordam esses quatro componentes de maneira mais interdisciplinar.

Os itens de CHT que apresentaram maior similaridade com os itens de MTT trazem conteúdos comuns às disciplinas escolares de Geografia e MTT. Já os itens de Sociologia e Filosofia apresentaram maior similaridade com os itens de LCT e demandam habilidades de interpretação de textos. Os itens relacionados ao componente curricular História se apresentaram mais distanciados das outras áreas do conhecimento, interagindo somente com os outros componentes curriculares de CHT nesta edição do exame.

A análise deste trabalho revela o potencial das técnicas de PLN como ferramentas de suporte à elaboração de itens e montagem de testes não segregados por área de conhecimento, com potencial de concretização da interdisciplinaridade valorizada nos documentos curriculares e diminuição do tamanho dos testes ao identificar temáticas de convergência textual entre as áreas do conhecimento. Entre

as limitações deste estudo, está a de não incluir a análise de imagens (como gráficos e fotografias). Outra limitação se refere ao componente aleatório na geração das posições dos itens no gráfico, ao transformar as 300 dimensões dos textos em duas. Ao se gerar novamente o gráfico, é possível que as distâncias entre os itens sejam alteradas. Recomendamos estudos que avancem nesses pontos, bem como os que incluam a análise de outras edições do Enem com o objetivo de identificar padrões de similaridade entre os textos das quatro provas do Enem ao longo dos anos, e também análises direcionadas para cada uma das áreas do conhecimento e componentes curriculares que compõem o exame.

6 Referências

BENEDETTO, L. et al. Introducing a Framework to Assess Newly Created Questions with Natural Language Processing. *In*: BITTENCOURT, I. I. et al. (org.). *Artificial Intelligence in Education*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020. v. 12163p. 43–54.BENOIT, K.; MUHR, D.; WATANABE, K. *stopwords:* Multilingual Stopword Lists. Disponível em: <https://CRAN.R-project.org/package=stopwords>. Acesso em: 24 jul. 2021.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python.* Sebastopol: O'Reilly Media, 2009.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais (Inep). *Exame Nacional do Ensino Médio (Enem)*: documento básico. Brasília: Inep, 2002. Disponível em: <http://www.dominiopublico.gov.br/download/texto/me000115.pdf>.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais (Inep). *Exame Nacional do Ensino Médio (Enem)*: relatório pedagógico 2009-2010. Brasília: Inep, 2013. Disponível em: <http://portal.inep.gov.br/documents/186968/484421/Relat%C3%B3rio+Pedag%C3%B3gico+ENEM+2009-2010/70890e24-a78a-44f8-a909-b235f02948f2?version=1.1>.

HARTMANN, N. et al. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. *arXiv:1708.06025 [cs]*, Portuguese Word Embeddings, 20 ago. 2017. Disponível em: <http://arxiv.org/abs/1708.06025>. Acesso em: 26 jul. 2021.

LIMA, A.J.C. A Sociologia nas Matrizes Curriculares do Ensino Médio e no Enem: temas, teorias e conceitos. In: SILVA, Ileizi Fiorelli; GONÇALVES, Danyelle Nilin (Orgs.). *A Sociologia na Educação Básica*. São Paulo: Annablume, 2017.

MACHADO, N.J. Interdisciplinaridade e contextuação. In: BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). *Exame Nacional do Ensino Médio (Enem)*: fundamentação teórico-metodológica. Brasília: Inep, 2005.

MIKOLOV, T. et al. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*, 16 out. 2013. Disponível em: <http://arxiv.org/abs/1310.4546>. Acesso em: 12 jul. 2021.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL - NILC. Repositório de Word Embeddings do NILC. NILC, USP, 2017. Disponível em:

https://www.nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc. Acesso em: 26 jul.2021.

PASSERO, G.; FERREIRA, R.; DAZZI, R. L. S. Off-Topic Essay Detection: A comparative study on the Portuguese language. *Revista Brasileira de Informática na Educação,* Off-Topic Essay Detection, v. 27, n. 03, p. 177–190, 31 dez. 2019.

PRIMI, R. Uso do word-to-vec (word embeddings) para análise de textos. *In*: CRISTIANE FAIAD; MAKILIM NUNES BAPTISTA; RICARDO PRIMI (org.). *Tutoriais em análise de dados aplicados a psicometria.* Petrópolis: Vozes, no prelo.

R CORE TEAM. *R:* a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2019. Disponível em: <https://www.R-project.org/>. Acesso em: 25 out. 2019.VALE, J.M.F. do; MAGNONI JÚNIOR, L. Geografia e Matemática: possíveis aproximações. Ciência Geográfica - Bauru - Ano XXIII - Vol. XXIII - (2): Janeiro/Dezembro – 2019.

VAN DER MAATEN, L.; HINTON, G. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, v. 9, n. 86, p. 2579–2605, 2008.