É possível calibrar os itens do Enem sem pré-teste?

Alexandre Jaloto
Inep
alexandre.jaloto@inep.gov.br

Alexandre José de Souza Peres
UFMS
alexandre.peres@ufms.br

Ana Carolina Zuanazzi
IAS
aczuanazzi@ias.org.br

Araê Cainã
USF
araecaina@hotmail.com

Ricardo Primi
USF
rprimi@mac.com

Resumo

O Exame Nacional do Ensino Médio (Enem) utiliza itens pré-testados para garantir a comparabilidade entre as provas de uma edição. No entanto, a pré-testagem de itens envolve alto custo operacional e risco para o processo, pois o exame é de alto impacto. Por isso, este trabalho objetivou verificar a capacidade de predizer o parâmetro de dificuldade de itens do Enem a partir de suas características textuais. Modelamos uma regressão linear com aprendizagem de máquina em que a variável predita foi o parâmetro b (dificuldade) dos itens de Ciências da Natureza e as variáveis preditoras foram as médias do escore do item em cada uma das 300 dimensões de vetores pré-treinado de palavras (*word embedings*). A correlação entre o parâmetro b da base de teste e o valor predito foi 0,50. Discutimos como as palavras podem influenciar na dificuldade do item.
Palavras-chave: processamento da linguagem natural; teoria de resposta ao item; avaliação educacional

Introdução

Os testes educacionais de larga escala que ocorrem em âmbito nacional em geral utilizam a Teoria de Resposta ao Item (TRI) para garantir a comparabilidade entre as provas. Como exemplo, temos os testes do Sistema de Avaliação da Educação Básica (Saeb), o Exame Nacional para Competências de Jovens e Adultos (Encceja) e o Exame Nacional do Ensino Médio (Enem). No caso específico do Enem, garantir a comparabilidade é necessário porque os sujeitos que respondem a provas diferentes concorrem às mesmas vagas nas universidades.

A comparabilidade no Enem é alcançada ao se posicionar todos os itens em uma única métrica, que tem como referência os concluintes regulares de escola pública do Enem 2009. Para posicionar os itens em uma mesma métrica, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) realiza pré-testes com o objetivo de obter os parâmetros dos itens na TRI. O modelo utilizado para o Enem é o logístico de três parâmetros (3PL) (INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, 2021). O pré-teste de um teste de alto impacto como o Enem requer investimento em segurança, pois caso um item seja publicado, a aplicação do exame será comprometida. Isso

porque os sujeitos que possuem acesso prévio a um item terão maior probabilidade de acertá-lo em comparação aos sujeitos de mesmo nível de competência que não tiveram acesso ao item. Adicionalmente, são necessárias pelo menos 750 respostas por item para a calibração no modelo 3PL, a depender do tamanho do teste (ŞAHIN; ANIL, 2017).

Uma alternativa ao pré-teste de itens com centenas ou milhares de sujeitos é a predição da dificuldade do item por meio de ferramentas do processamento da linguagem natural (*Natural Language Processing*, NLP) ao utilizar características do texto como variáveis preditoras. Duas dessas variáveis são o tamanho da palavra e o tamanho da frase, que predizem a dificuldade do item em teste de vocabulário em inglês (SETTLES; LAFLAIR; HAGIWARA, 2020). Outro exemplo desse tipo de variável é a similaridade entre o enunciado e o gabarito, que prediz a dificuldade de um item de múltipla escolha de estudos sociais (HSU et al., 2018, p. 218). Adicionalmente, combinar as características textuais do item com a resposta de poucos alunos aumenta a predição da dificuldade (MCCARTHY et al., 2021). Este trabalho buscou utilizar a potencialidade das ferramentas de NLP para avançar nas possibilidades de calibração de itens do Enem, dada a importância de se conhecer seus parâmetros antes da aplicação e a necessidade de se expor de maneira limitada esses itens em um pré-teste.

Objetivo

Verificar a capacidade de predizer o parâmetro de dificuldade dos itens de Ciências da Natureza do Enem a partir de suas características textuais. Este trabalho avança ao contribuir para uma reflexão sobre a substituição ou complementação de um pré-teste de itens de um exame de alto impacto como o Enem.

Método

Utilizamos 600 itens de Ciências da Natureza de 2009 a 2020 disponibilizados nos microdados do Enem (INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, 2023) que possuíam parâmetros de dificuldade (parâmetro b) entre -3,0 e 3,0. Esse filtro foi utilizado para retirar os itens muito distantes do centro da escala. Construímos um modelo de regressão linear com aprendizagem de máquina em que a variável predita foi o parâmetro b dos itens. As variáveis preditoras foram obtidas a partir de vetores de palavras treinados por Hartmann et al. (2017) e disponíveis no Repositório de Word Embeddings, do Núcleo Interinstitucional de Linguística Computacional (2021). Os valores utilizados como preditores corresponderam à média entre as palavras do item para cada uma das 300 dimensões do vetor pré-treinado. Esses valores foram obtidos com base no protocolo proposto por Primi (2021) e adotou os seguintes passos:

1. Divisão do item em palavras

2. Exclusão de números, palavras de parada (*stopwords*) e palavras duplicadas nos itens, e formatação para todas as palavras ficarem com letras em minúsculo
3. Obtenção dos vetores de palavras com 300 dimensões
4. Cálculo da média do item para cada dimensão, considerando os valores das palavras

Para o modelo de regressão linear, dividimos os itens em duas bases: treino (480 itens) e teste (120 itens) e utilizamos o pacote tidymodels (KUHN; WICKHAM, 2020). Avaliamos o modelo de regressão linear pela correlação entre o valor predito para os itens da base de teste e seu parâmetro de dificuldade. Adicionalmente, verificamos a dimensão com maior contribuição positiva na regressão e a com maior contribuição negativa. Em seguida, das palavras utilizadas nos itens, construímos uma nuvem com as 100 palavras com maior escore de cada dimensão.

As dimensões com maiores contribuições positivas na regressão contribuem mais para o item ser mais difícil. Ou seja, os itens com valores mais altos dessas dimensões tendem a ser mais difíceis. Já as dimensões com contribuições mais negativas contribuem mais para o item ser mais fácil. Ou seja, os itens com maiores valores dessas dimensões tendem a ser mais fáceis.

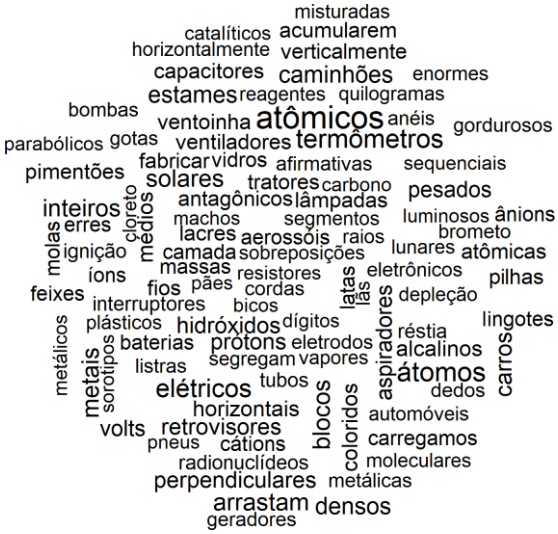
Resultados e discussão

A correlação entre o valor predito para os itens da base de teste e seu parâmetro de dificuldade foi 0,50. Ou seja, nosso modelo explica 25,2% da variância da dificuldade dos itens dessa base. O valor de R^2 superou a maioria dos valores obtidos por McCarthy et al. (2021), que encontraram correlações que variaram de 0,32 a 0,74 entre as dificuldades preditas e as reais de itens de um teste de língua inglesa. Cabe destacar que o trabalho citado utilizou a resposta de alguns sujeitos em algumas modelagens.

A **Figura** 1 mostra a nuvem de palavras da dimensão com a maior contribuição positiva. Essa nuvem contém as 100 palavras com maior escore nessa dimensão. Isso significa que um item que contém essas palavras muito provavelmente terá alto valor nessa dimensão e sua dificuldade deverá ser maior. Nota-se que a nuvem é formada predominantemente por vocabulário especializado da área de Ciências da Natureza (e.g., hidróxidos, prótons, átomos, metais, massas e capacitores) ou por palavras que sugerem um contexto próprio dessa área (e.g., termômetros, retrovisores, feixes, solares, lâmpadas e pimentões). As palavras que configuram vocabulário especializado em geral são mais utilizadas nas disciplinas escolares de Física e Química.

Figura 1

Nuvem das palavras com maior escore na dimensão com maior contribuição positiva na regressão



A Figura 2 mostra o item com a maior dificuldade. Esse item possui dez das 100 palavras mais características da dimensão com maior contribuição positiva. Ou seja, das palavras usadas nesse item, dez estão presentes na nuvem da Figura 1: bateria, baterias, descartadas, elétricos, fabricantes, ingerir metais, pesados, pilhas e sólidos.

Figura 2

Item com maior parâmetro de dificuldade

Cerca de 1% do lixo urbano é constituído por resíduos sólidos contendo elementos tóxicos. Entre esses elementos estão metais pesados como o cádmio, o chumbo e o mercúrio, componentes de pilhas e baterias, que são perigosos à saúde humana e ao meio ambiente. Quando descartadas em lixos comuns, pilhas e baterias vão para aterros sanitários ou lixões a céu aberto, e o vazamento de seus componentes contamina o solo, os rios e o lençol freático, atingindo a flora e a fauna. Por serem bioacumulativos e não biodegradáveis, esses metais chegam de forma acumulada aos seres humanos, por meio da cadeia alimentar. A legislação vigente (Resolução CONAMA n° 257/1999) regulamenta o destino de pilhas e baterias após seu esgotamento energético e determina aos fabricantes e/ou importadores a quantidade máxima permitida desses metais em cada tipo de pilha/bateria, porém o problema ainda persiste.

Disponível em: http://www.mma.gov.br. Acesso em: 11 jul. 2009 (adaptado).

Uma medida que poderia contribuir para acabar definitivamente com o problema da poluição ambiental por metais pesados relatado no texto seria:

A) deixar de consumir aparelhos elétricos que utilizem pilha ou bateria como fonte de energia.
B) usar apenas pilhas ou baterias recarregáveis e de vida útil longa e evitar ingerir alimentos contaminados, especialmente peixes.
C) devolver pilhas e baterias, após o esgotamento da energia armazenada, à rede de assistência técnica especializada para repasse a fabricantes e/ou importadores.
D) criar nas cidades, especialmente naquelas com mais de 100 mil habitantes, pontos estratégicos de coleta de baterias e pilhas, para posterior repasse a fabricantes e/ou importadores.
E) exigir que fabricantes invistam em pesquisa para a substituição desses metais tóxicos por substâncias menos nocivas ao homem e ao ambiente, e que não sejam bioacumulativas.

A Figura 3 mostra a nuvem de palavras da dimensão com a maior contribuição negativa. Essa nuvem contém as 100 palavras com maior escore nessa dimensão. Isso significa que um item que contém essas palavras muito provavelmente terá alto valor nessa dimensão e sua dificuldade deverá ser menor. Nota-se que as palavras se referem a contextos gerais e

configuram vocabulários pouco especializados (e.g., mediados, suínos, ferimentos, motorista e frei). É possível que os itens com altos valores nessa dimensão demandem a aplicação de conceitos científicos em contextos mais comuns e poucos especializados.

Figura 3

Nuvem das palavras com maior escore na dimensão com maior contribuição negativa na regressão

A Figura 4 mostra o item com a menor dificuldade. Esse item possui uma das 100 palavras mais características do tópico com maior contribuição negativa (água).

Figura 4

Item com menor parâmetro de dificuldade

Durante as estações chuvosas, aumentam no Brasil as campanhas de prevenção à dengue, que têm como objetivo a redução da proliferação do mosquito *Aedes aegypti*, transmissor do vírus da dengue.

Que proposta preventiva poderia ser efetivada para diminuir a reprodução desse mosquito?

A) Colocação de telas nas portas e janelas, pois o mosquito necessita de ambientes cobertos e fechados para a sua reprodução.
B) Substituição das casas de barro por casas de alvenaria, haja vista que o mosquito se reproduz na parede das casas de barro.
C) Remoção dos recipientes que possam acumular água, porque as larvas do mosquito se desenvolvem nesse meio.
D) Higienização adequada de alimentos, visto que as larvas do mosquito se desenvolvem nesse tipo de substrato.
E) Colocação de filtros de água nas casas, visto que a reprodução do mosquito acontece em águas contaminadas.

Conclusão

Este trabalho verificou a capacidade de predizer o parâmetro de dificuldade dos itens de Ciências da Natureza do Enem a partir de suas características textuais, sem o uso de respostas de sujeitos. De forma geral, conseguimos explicar 25,2% da dificuldade dos itens por meio de

seus textos. Estudos futuros devem utilizar outros procedimentos de obtenção dos vetores numéricos dos itens, como o BERT. Recomendamos também o uso de outras características dos itens, como a similaridade entre o gabarito e o enunciado, o tamanho médio das frases e o tamanho médio das alternativas do item. Este trabalho tem como limitação o fato de não fazer um tratamento nos itens com imagens, pois utilizamos somente os textos verbais. Esperamos contribuir para a redução dos custos dos testes educacionais e para o aumento de sua segurança.

Referências

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.

HSU, F.-Y. et al. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. **Information Processing & Management**, Automated estimation of item difficulty for multiple-choice tests, v. 54, n. 6, p. 969–984, nov. 2018.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Exame Nacional do Ensino Médio - Enem** procedimentos de análise. Brasília: Inep, 2021. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_procedimentos_de_analise.pdf>. Acesso em: 26 out. 2022.

_____. **Microdados**. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/>. Acesso em: 14 mar. 2023.

KUHN, M.; WICKHAM, H. **Tidymodels** a collection of packages for modeling and machine learning using tidyverse principles. [s.l: s.n.]. Disponível em: <http://www.tidymodels.org>. Acesso em: 12 jun. 2023.

MCCARTHY, A. D. et al. Jump-starting item parameters for adaptive language tests. Em: 2021 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. 2021.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. **Repositório de Word Embeddings do NILC**. Disponível em: <http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc>. Acesso em: 26 jul. 2021.

PRIMI, R. Uso do word-to-vec (word embeddings) para análise de textos. *Em*: FAIAD, C.; BAPTISTA, M. N.; PRIMI, R. (org.). **Tutoriais em análise de dados aplicados a psicometria**. Petrópolis: Vozes, 2021. p. 460–476.

ŞAHIN, A.; ANIL, D. The effects of test length and sample size on item parameters in item response theory. **Educational Sciences: Theory & Practice**, v. 17, n. 1, p. 321–335, 2017.

SETTLES, B.; LAFLAIR, G. T.; HAGIWARA, M. Machine learning–driven language assessment. **Transactions of the Association for Computational Linguistics**, v. 8, p. 247–263, dez. 2020.