

Predição da Dificuldade dos Itens do ENEM a partir de Características Textuais

Maria Victória Brandão Barros
USP - ICMC
mariabrandaob@usp.br

Renan Gonzales
USP - ICMC
renan.gonzales@usp.br

Lucas Ivars Cadima Ciziks
USP - ICMC
luciziks@usp.br

Mariana Curi
Departamento de Matemática Aplicada e Estatística
USP - ICMC
mcuri@icmc.usp.br

RESUMO

A TRI é usada no ENEM para calibrar itens, analisando dificuldade e discriminação. Pré-testes garantem qualidade, mas são caros e arriscados. Este estudo propõe métodos para prever a dificuldade dos itens, baseando-se em dados anteriores. O estudo utilizou itens de ciências da natureza do ENEM (2009-2023) e seus parâmetros de dificuldade. Os dados foram pré-processados e, em seguida, vetorizados com o método Word2Vec, criando vetores médios que representam cada enunciado. Dois modelos de regressão (linear e lasso) foram treinados e testados para prever a dificuldade dos itens, usando a raiz do erro quadrático médio (RMSE) para validar a precisão das previsões. Quanto menor o RMSE, melhor o desempenho do modelo. O modelo Lasso obteve melhores resultados entre os testes realizados. Reduzir o tamanho dos vetores dos enunciados teve impacto neutro ou positivo nas previsões, pois evitou redundâncias. Os melhores modelos obtiveram valor de RMSE de 0,5.

Palavras-chave: Regressão; Predição; ENEM; Dificuldade.

*Os três primeiros autores tiveram contribuições iguais para a elaboração deste trabalho

1. INTRODUÇÃO

A Teoria de Resposta ao Item (TRI) é amplamente utilizada em avaliações educacionais de larga escala, como o Exame Nacional do Ensino Médio (ENEM). Este método avalia itens com base em parâmetros como dificuldade, discriminação e acerto casual, permitindo uma análise da proficiência dos respondentes, que representam uma amostra da população de pessoas a realizar o Exame. No contexto do ENEM, a TRI é utilizada em etapas preliminares, os pré-testes, que visam calibrar os itens antes de sua inclusão no banco de itens.

Os pré-testes do ENEM são aplicados para garantir a qualidade e a equidade do exame, uma vez que permitem identificar itens com grau de dificuldade inadequado antes de sua utilização na prova oficial. Nessa etapa, a TRI oferece ferramentas estatísticas para analisar o comportamento dos itens em relação ao desempenho dos candidatos, verificando, por exemplo, se itens muito fáceis ou muito difíceis estão adequadamente discriminando os participantes de diferentes níveis de habilidade.

Contudo, esse processo exige recursos operacionais e financeiros elevados, além do risco causado por uma aplicação prévia do exame, já que a exposição de um item pode comprometer sua validade. Por isso, é coerente a pesquisa de outros métodos que podem ser utilizados em conjunto aos pré-testes, ou que atuem como seu substituto.

Este trabalho desenvolveu métodos estatísticos capazes de prever o parâmetro de dificuldade de um item com base em provas aplicadas anteriormente. O estudo usa como ponto de partida a metodologia proposta por Jaloto et al. (2023), acrescentando análises de outros modelos estatísticos e medidas de avaliação dos resultados.

2. METODOLOGIA

2.1. Dados utilizados

O trabalho aqui apresentado utiliza enunciados dos itens de ciências da natureza das provas do ENEM aplicadas entre 2009 e 2023. Para as provas de 2017 em diante, também utilizamos as provas em formato LEDOR, que possuem metodologia consistente para leitura do conteúdo das imagens dos itens por um profissional, para os participantes com deficiência visual. O motivo desta escolha foi evitar a perda de informações presentes em figuras. Em outras palavras, enunciados a partir de 2017 tendem a ser mais ricos para o modelo de predição, uma vez que serão capazes de capturar informações não apenas do texto do enunciado, mas também de imagens utilizadas nos itens. Além do conteúdo textual, foram utilizadas as estimativas dos

parâmetros de dificuldade de cada item fornecidas nos microdados de cada prova. Os dados utilizados para este trabalho foram coletados do site oficial do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

2.2. Pré-processamento dos dados

Em todos os textos utilizados no estudo foi aplicada uma sequência de transformações dos dados: conversão para letras minúsculas, remoção de pontuação, exclusão de *stopwords* (palavras que pouco acrescentam no tema principal de uma sentença), remoção de números e lematização. Também foram eliminadas palavras duplicadas dentro de cada item, de modo a reduzir o ruído nas representações vetoriais posteriores. O objetivo desta transformação é preparar os dados para a fase de vetorização e garantindo consistência na semântica dos termos representados.

2.3. Vetorização dos enunciados e cálculo de vetor de médias

Utilizando o método de vetorização Word2Vec (Mikolov et al., 2013), cada palavra de cada enunciado foi transformada em um vetor numérico de dimensão pré-determinada que representa aspectos específicos de tal palavra. Os métodos utilizados foram testados com vetores de 50, 100 e 300 dimensões para avaliar a diferença entre os resultados dos modelos estatísticos quando aplicados em representações com diferentes tamanhos.

A partir desta transformação, foi calculado, para cada enunciado, um vetor resultante das médias dos vetores de palavras contidas no texto. Este vetor resultante atua como uma representação geral das informações, temas e ideias presentes em cada item, e foi utilizado como base para a predição do valor do parâmetro de dificuldade.

2.4. Modelos estatísticos de regressão

Para a predição do parâmetro de dificuldade de cada item, foram utilizados dois métodos de regressão, a regressão linear e a regressão lasso (Tibshirani, 1996). Ambos os modelos seguem o pressuposto de que há uma relação linear entre a representação numérica de um enunciado e o valor do parâmetro para aquele item.

Os dados foram separados numa proporção de 80% de treino e 20% de teste. A parcela de treino foi utilizada para modelar os métodos estatísticos, ou seja, encontrar uma forma de calcular o valor do parâmetro de dificuldade de um item a partir do vetor numérico de seu enunciado. Nesta etapa, os modelos têm acesso tanto às variáveis explicativas, que são os

vetores numéricos, como à variável resposta que, neste caso, é o valor do parâmetro, e utilizam ambos para construir uma função linear que melhor prediz a dificuldade do item.

A parcela de teste é usada para validar a função encontrada na etapa anterior. Para isso, o modelo estimado é aplicado às representações numéricas dos enunciados e, dessa forma, calcula-se o valor previsto para o parâmetro de dificuldade destes itens. As previsões são então comparadas com o valor verdadeiro do parâmetro de cada item e, a partir daí, utiliza-se uma medida de erro para avaliar o desempenho do modelo estatístico.

2.5. Validação do modelo

Para validar o modelo, a métrica de erro utilizada foi a raiz do erro quadrático médio (RMSE, do inglês Root Mean Squared Error). Essa métrica é calculada pela raiz quadrada da média dos erros quadráticos, na qual a medida de erro é obtida pela diferença entre o valor fornecido nos microdados e o valor predito do parâmetro de dificuldade. Dessa forma, quanto menor o valor desta métrica, menor é a média dos erros cometidos pelo modelo.

A equação abaixo apresenta a definição matemática da raiz do erro quadrático médio, onde y_i representa cada valor real do parâmetro e \hat{y}_i representa o respectivo valor previsto pela função.

Equação 1: Definição matemática do RMSE

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

3. RESULTADOS E DISCUSSÃO

A tabela a seguir mostra os resultados obtidos para cada modelo testado, utilizando os métodos de regressão linear e regressão lasso, e diferentes tamanhos de vetores. Os resultados também estão divididos entre experimentos realizados apenas com os textos presentes nas provas e experimentos que incluem as provas em formato LEDOR, quando disponíveis.

Tabela 1: Valores de RMSE dos resultados dos experimentos realizados

Dimensões	Sem usar as provas LEDOR		Usando as provas LEDOR	
	Regressão Linear	Regressão Lasso	Regressão Linear	Regressão Lasso

300	1,05	2,26	1,00	0,50
100	0,58	0,51	0,56	0,50
50	0,54	0,50	0,52	0,50

Como é possível observar na Tabela 1, os melhores resultados foram obtidos com os modelos que utilizaram a regressão lasso. Este resultado se dá pelo fato da regressão lasso ser uma forma mais “avançada” do modelo linear clássico.

Também é possível concluir que a diminuição da dimensão dos vetores que representam os enunciados tem impacto neutro ou positivo no desempenho dos modelos. Uma provável explicação para esse fenômeno é que, com muitas dimensões representando cada palavra, vários valores dos vetores se tornam redundantes. Por fim, os melhores resultados obtidos são de modelos que atingiram um valor de RMSE de 0,5.

4. CONCLUSÃO E CONSIDERAÇÕES FINAIS

Este trabalho traz luz à utilização de novas metodologias para o cálculo do parâmetro de dificuldade do TRI, cuja implementação, caso se prove viável, poderia substituir a aplicação do pré-teste do ENEM. Isso resultaria em economia de recursos humanos e financeiros investidos anualmente nesse processo.

Desde a coleta e pré-processamento dos textos até a análise dos resultados, os métodos aplicados mostram-se eficientes no contexto do estudo realizado. Os resultados obtidos podem ser considerados satisfatórios para um trabalho contendo apenas experimentos iniciais, que podem ser aprimorados com a aplicação de métodos mais robustos.

Por fim, a investigação de outras técnicas, como a utilização de outros métodos de vetorização e outros modelos estatísticos, é viável para fornecer uma comparação com os métodos aqui aplicados. É também possível aprimorar os modelos apresentados com a inserção de mais enunciados provenientes de novas provas aplicadas futuramente, aumentando a quantidade de dados a serem utilizados, e também adicionando outras características dos itens, além dos enunciados, melhorando o desempenho dos modelos.

5. REFERÊNCIAS

JALOTO, A. et al. É possível calibrar os itens do Enem sem pré-teste? Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2023.

MIKOLOV, T. et al. Efficient Estimation of Word Representations in Vector Space. Google Inc., 2013.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), v. 58, n. 1, p. 267-288, 1996.