Geração automática de questões de múltipla escolha com níveis de dificuldade

Palavras-Chave: aprendizagem ativa, questões e respostas, inteligência artificial generativa

Autores:
Lucca de Souza Amodio, IC – Unicamp
Prof. Dr. Julio Cesar dos Reis (orientador), IC - Unicamp

INTRODUÇÃO

A condução de avaliações estruturadas é fundamental para o acompanhamento do processo de ensino-aprendizagem. Muito mais que apenas uma forma de avaliar o conhecimento em determinado assunto, esse processo auxilia na memorização de informação. As avaliações permitem diagnosticar o nível de conhecimento dos alunos, identificar lacunas específicas de aprendizagem e adaptar intervenções pedagógicas de forma mais precisa e eficaz.

A realização de avaliações é parte fundamental do processo de aprendizagem, pois, muito mais do que apenas uma forma de mensurar o conhecimento em determinado assunto, esse processo também auxilia na memorização e retenção de informações por meio do chamado “efeito de ser testado”, levando o aluno, ao ser submetido a avaliações contínuas sobre um mesmo tema, a aumentar o tempo de estudo e aprimorar suas estratégias com vistas a obter melhores resultados. Para a neurobiologia, inclusive, o estresse e a ansiedade decorrentes da realização de testes podem agir como moduladores positivos do aprendizado. Em uma pesquisa conduzida por Joëls e Wiegert [1], concluiu-se que hormônios e neurotransmissores liberados em resposta ao estresse podem ativar circuitos neurais relacionados ao processamento de informações, facilitando mecanismos de memória e aprendizado. Uma das formas mais utilizadas para essa testagem dos estudantes é por meio dos questionários gerados manualmente pelos professores, uma tarefa que demanda muita criatividade e tempo.

É evidente que uma forma automatizada de geração de questionários *online* a partir de conteúdos de diferentes mídias é de extrema utilidade para o aprendizado de estudantes de diversos níveis escolares, assim como para que docentes possam fazer o acompanhamento dos alunos, podendo inclusive utilizar os questionários como parte da avaliação de suas disciplinas. Ao utilizar as perguntas geradas automaticamente em exames avaliativos, os docentes podem aplicar mais tempo para melhorias no processo de aprendizagem dos alunos ou para pesquisas, melhorando o processo de ensino-aprendizagem como um todo.

Nos últimos anos, o avanço dos modelos de linguagem de larga escala (*LLMs)* revolucionou a área de Processamento de Linguagem Natural (PLN), com impactos diretos sobre aplicações educacionais. Além dos *LLM*s, modelos como o *BERT* (*Bidirectional Encoder Representations from Transformers*), voltado para tarefas de entendimento de linguagem, e o T5 [2] (*Text-To-Text Transfer Transformer)*, com arquitetura *encoder-decoder*, são amplamente utilizados em tarefas supervisionadas e de geração textual. A ampla disponibilidade desses modelos e suas variantes, aliados a conjuntos de dados públicos, tem impulsionado a pesquisa em aplicações educacionais baseadas em IA.

Modelos de linguagem pré-treinados podem ser refinados (adaptados) para tarefas específicas como classificação de texto, geração de questões e criação de distratores. Entretanto, ainda há desafios significativos no uso de modelos de linguagem para avaliar a dificuldade de questões. Ao desenvolver um modelo capaz de classificar uma questão conforme a dificuldade, ele pode ser usado como juiz para o processo de geração de questões com diferentes níveis de dificuldade. A geração de cada questão em si é uma tarefa mais simples, que pode ser refinada via *prompt.* O desafio reside em desenvolver o modelo juiz que vai segmentar as questões geradas, pensando em um *pipeline* de geração de questões rotuladas com dificuldade. Estudos como o de Scaria et al. [3] demonstraram que abordagens baseadas exclusivamente em *prompting* não foram satisfatórias quando comparadas à avaliação humana. Isso reforça a relevância de se investigar métodos supervisionados para treinar classificadores capazes de prever o nível de dificuldade com maior precisão. Em um cenário educacional cada vez mais pautado por dados, automatizar e qualificar a análise da dificuldade das questões torna-se essencial para garantir que os instrumentos avaliativos sejam adequados aos perfis dos estudantes, promovendo a equidade no acesso ao conhecimento e o aprimoramento contínuo.

METODOLOGIA

O conjunto de dados utilizado neste trabalho é composto por questões extraídas de edições anteriores do Exame Nacional do Ensino Médio (ENEM), com foco na área de Ciências Humanas e Linguagens. A cada ano, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) divulga os microdados da aplicação, incluindo os parâmetros da Teoria de Resposta ao Item (TRI), que são empregados para estimar a proficiência dos candidatos e a dificuldade das questões.

Os parâmetros da TRI [4], especialmente o parâmetro b (dificuldade), não estão calibrados em uma escala padronizada entre diferentes anos. Isso ocorre porque a modelagem é realizada separadamente a cada edição da prova, resultando em escalas distintas de proficiência. Embora o INEP possa aplicar técnicas de equiparação, esses procedimentos não são publicados.

Devido à ausência de uma escala comum entre os anos, a comparação direta entre valores de b de diferentes edições não é apropriada. Para viabilizar a unificação dos dados em um único modelo de classificação, foi adotada uma estratégia de rotulagem relativa dentro de cada ano. Em cada edição, as questões foram ordenadas segundo o valor do parâmetro b e divididas em três faixas, correspondentes às classes fácil, médio e difícil.

Após a rotulagem individual por ano, todas as questões foram agrupadas em um único conjunto de dados, agora com rótulos de dificuldade consistentes entre si. Essa abordagem permite a utilização de questões de diferentes edições do ENEM em uma tarefa supervisionada de classificação, mantendo a coerência relativa entre os níveis de dificuldade considerados.

Para classificar as questões segundo seu nível de dificuldade, foram geradas duas versões do conjunto de dados, cada uma utilizando uma abordagem distinta de rotulagem com base na métrica *parB*. A primeira versão (DS1) adotou uma estratégia simples de ordenação. As questões foram agrupadas por prova e ordenadas de forma crescente segundo o valor de *parB*, sendo então divididas em três faixas de mesmo tamanho (terços). As questões no primeiro terço, com menor *parB*, foram rotuladas como de dificuldade 0 (fáceis); as do terço intermediário, como dificuldade 1 (médias); e as do terço superior, com maior *parB,* como dificuldade 2 (difíceis). Essa abordagem garante uma distribuição equilibrada entre os níveis de dificuldade dentro de cada prova.

Já a segunda versão (DS2) explorou uma abordagem não supervisionada por meio do algoritmo *K-Means*, com o objetivo de identificar agrupamentos naturais nos dados. Novamente, as questões foram separadas por prova, mas, em vez de usar a ordenação direta, aplicou-se o *K-Means* com três *clusters* sobre os valores de *parB*. Essa versão tende a ser mais flexível em relação à distribuição dos dados.

O *dataset* de ciências humanas, utilizado na maioria dos experimentos, possui 985 questões. Em um dos experimentos, foram utilizadas também as questões de Linguagens, totalizando cerca de 1400 questões.

Modelos *BERT*

O principal modelo utilizado neste estudo foi o *BERTimbau*[1] [5], uma versão do *BERT* pré-treinada especificamente para o português brasileiro. Baseado na arquitetura *Transformer*, o *BERTimbau* emprega mecanismos de atenção bidirecional que permitem a captura contextualizada das palavras, considerando simultaneamente o contexto à esquerda e à direita de cada termo no texto.

O pré-treinamento do *BERTimbau* foi realizado em um grande *corpus* textual em língua portuguesa, o que favorece sua capacidade de representação semântica e sintática adequada para tarefas de processamento de linguagem natural (PLN) em português. Esta especialização linguística resulta em desempenho superior em comparação a modelos multilíngues ou treinados exclusivamente em inglês, especialmente em aplicações como classificação de texto e análise semântica.

A escolha do *BERTimbau* justifica-se pela necessidade de representações contextuais precisas para a análise e classificação das questões, garantindo maior acurácia e robustez nos resultados.

Além disso, realizou-se um experimento adicional com o *SQUAD*[2], um modelo similar, com as mesmas condições do último experimento com o *BERTimbau* (fornecendo o maior número de detalhes possível ao modelo), a fim de realizar uma comparação entre os resultados.

Os experimentos foram com o *dataset* dividido em treinamento (80%) e teste (20%). As métricas avaliadas foram:

- **Acurácia (*Accuracy*)**: mede a proporção de predições corretas sobre o total de exemplos.
- **Precisão (*Precision*)**: indica a proporção de predições corretas dentre todas as predições feitas para uma classe (micro ou macro, dependendo do uso).
- ***Recall**:* mede a proporção de exemplos relevantes que foram corretamente identificados.
- ***F1-score***: média harmônica entre precisão e revocação, ponderada (*average="weighted*") para levar em conta o desbalanceamento entre classes.

A tabela 1 sintetiza os experimentos que foram realizados. Foram realizados 6 experimentos com o modelo *BERTimbau* (linhas 1 a 6 da tabela) e 1 com o modelo *SQUAD* (linha 7 da tabela), a fim de avaliar como a variação de parâmetros, métodos de separação dos níveis de dificuldade do dataset e número de questões influenciam no resultado da predição por parte do modelo.

Tabela 1. Experimentos dos modelos *BERT*

Exp	Áreas	*Input*	*Split*	*Learning Rate*
1	Humanas	Enunciado	DS1	1e-5
2	Humanas	Enunciado	DS1	2e-5
3	Humanas + Linguagens	Enunciado	DS1	1e-5
4	Humanas	Enunciado + Alternativas	DS1	1e-5
5	Humanas	Enunciado +	DS2	1e-5

[1] https://huggingface.co/neuralmind/bert-base-portuguese-cased
[2] https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese

		Alternativas		
6	Humanas	Enunciado + Alternativas	DS2	2e-5
7	Humanas	Enunciado + Alternativas	DS2	2e-5

Modelos *LLM*

Em relação aos modelos *LLM,* a estratégia adotada foi a realização de um *prompt* com as instruções de avaliação e com um número variável de exemplos escolhidos do *dataset,* por meio de um algoritmo que busca balancear o número de questões de cada classe de dificuldade. O *prompt* utilizado foi bem simples: *"Você é um especialista em educação que classifica a dificuldade de questões como 'fácil', 'médio' ou 'difícil'.",* seguido dos exemplos. Em seguida, um outro conjunto de questões era selecionado, com o mesmo algoritmo de balanceamento, para testar o modelo, e obter sua acurácia. Alguns modelos testados incluem o *Deepseek, Llama e o Mistral.*

RESULTADOS E DISCUSSÃO

Modelos *BERT*

Tabela 2. Resultados dos modelos *BERT*

Exp	Melhor acurácia	Melhor *F1-score*	Melhor *Loss*
1	0.327	0.301	1.102
2	0.400	0.392	1.088
3	0.379	0.372	1.098
4	0.441	0.435	1.085
5	0.580	0.426	0.972
6	0.591	0.459	0.970
7	0.580	0.470	0.9797

A tabela 2 apresenta os resultados dos experimentos, que mostram uma variação considerável de desempenho entre as diferentes configurações testadas, com foco nas métricas de acurácia, *F1-score* ponderado e *loss* de validação. A maioria das configurações, especialmente os experimentos 1, 2, 3 e 4, apresentou acurácias entre 0,32 e 0,44. Embora esses valores sejam ligeiramente superiores à acurácia esperada de uma classificação aleatória (aproximadamente 0,33 em um cenário com três classes), eles ainda indicam que o modelo não está capturando adequadamente as nuances envolvidas na tarefa de classificação da dificuldade das questões.

Um destaque relevante é o experimento 5, que alcançou uma acurácia de 0,580, superando significativamente todas as configurações anteriores. Esse resultado sugere que a combinação do segundo método de rotulagem (baseado em *K-Means)* com a inclusão das alternativas das questões como parte da entrada possui um potencial promissor para a tarefa. No entanto, ao analisar as demais

métricas, nota-se uma discrepância importante: o *F1-score* ponderado foi consideravelmente mais baixo, atingindo apenas 0,426. Essa diferença sugere a presença de um desequilíbrio entre as classes, em que o modelo tende a predizer com maior frequência uma única classe, provavelmente a mais representada no conjunto, o que inflaciona a acurácia geral, mas compromete a capacidade do modelo de classificar corretamente as instâncias das classes menos frequentes. A baixa precisão (0,337) e o recall relativamente elevado (0,580) reforçam essa hipótese, indicando que o modelo acerta muitos exemplos da classe majoritária, mas erra frequentemente ao tentar identificar as demais.

Ao comparar o Experimento 4 com o Experimento 2, ambos com LRs similares, o Experimento 4 apresentou uma melhoria clara na acurácia (0.441 vs. 0.400) e no *F1-score* (0.435 vs. 0.392). Isso sugere que a inclusão das alternativas como parte da entrada do modelo fornece informações contextuais valiosas que auxiliam na classificação da dificuldade.

Em relação ao experimento com o modelo *SQUAD,* é possível perceber que o *BERTimbau* apresentou acurácia e precisão mais consistentes.

Modelos *LLM*

Em relação aos experimentos com *LLM*s, não foi possível obter um resultado relevante, pois para todos os modelos testados, a acurácia raramente ultrapassava a classificação aleatória de 33%. Nesse caso, imaginamos que alguns fatores poderiam ser adotados futuramente para garantir que os modelos tenham sucesso na tarefa de classificar a dificuldade das questões. Primeiramente, um *prompt* mais direcionado, com instruções detalhadas de como o agente deve classificar as questões, passo a passo. Além disso, realizar testes com as questões do *prompt* sendo passadas em uma ordem aleatória. Atualmente, elas são passadas na ordem: fácil, médio e difícil. Percebemos que o modelo está classificando muito mais questões como difíceis, possivelmente por ser a última classe que ele recebe pelo prompt como exemplo.

CONCLUSÃO

Em resumo, embora os modelos *BERT* tenham demonstrado algum grau de aprendizado, seu desempenho geral ainda é insatisfatório, o que pode ser atribuído, em parte, à quantidade limitada de questões disponíveis no *dataset.* As *LLM*s, por sua vez, também não apresentaram resultados expressivos, indicando que a tarefa proposta continua sendo um desafio. Apesar disso, conseguimos produzir um *dataset* que poderá servir de base para futuras pesquisas na área. Além disso, o aprendizado parcial observado nos modelos *BERT* oferece uma validação preliminar da consistência e relevância do *dataset* construído.

BIBLIOGRAFIA

[1] Joëls M, Pu Z, Wiegert O, Oitzl MS, Krugers HJ. Learning Under Stress: How does it work? Trends Cogn Sci. 2006;10(4):152-8.

[2] Carmo, Diedre and Piau, Marcos and Campiotti, Israel and Nogueira, Rodrigo and Lotufo, Roberto (2020). PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. arXiv preprint arXiv:2008.09144

[3] Nicy Scaria, Suma Dharani Chenna, Deepak Subramani. Automated Educational Question Generation at Different Bloom's Skill Levels using Large Language Models: Strategies and Evaluation. https://arxiv.org/abs/2408.04394

[4] Costa, Sidney Tadeu Santiago. Teoria De Resposta Ao Item Aplicada No ENEM. 2017. https://bdtd.ibict.br/vufind/Record/UFG-2_e37686eaddeec7b02a9de89141d91232

[5] Souza, F., Nogueira, R., Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds) Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science(), vol 12319. Springer, Cham. https://doi.org/10.1007/978-3-030-61377-8_28