SME0878 - Mineração Estatística de Dados

Projeto ENEM Grupo A

Lucas Garzeri de Melo - 13731344
Lua Nardi Quito - 11371270
Lucas Schimidt Coelho - 11913019
Lucca Baptista Silva Ferraz - 13688134

11 de junho de 2025

1 Introdução

1.a Sobre o projeto Desde 2009, o Exame Nacional do Ensino Médio (ENEM) tornou-se a principal porta de entrada para o ensino superior no Brasil, assumindo o papel de um vestibular unificado. Com isso, passou a ter um impacto social significativo, exigindo um processo rigoroso e criterioso de elaboração das questões para garantir justiça, qualidade e confiabilidade na avaliação. O processo de criação dos itens do ENEM envolve diversas etapas: seleção e capacitação de professores, elaboração e múltiplas revisões das questões, aplicação de pré-testes com estudantes, análise psicométrica e, por fim, a escolha dos itens que comporão a prova. Entre essas etapas, o pré-teste é fundamental para estimar a dificuldade dos itens antes da aplicação oficial. No entanto, ele também representa um dos maiores gargalos do processo — tanto em termos logísticos quanto financeiros. Em 2009, o pré-teste envolveu apenas uma etapa com cerca de 48 mil estudantes e teve custo estimado de R$ 939,5 mil. Já em 2010, esse número saltou para quatro etapas, envolvendo 100 mil estudantes em 40 municípios, com custo estimado de R$ 6,1 milhões. Além dos altos custos, o processo ainda expõe os itens a possíveis vazamentos e limitações operacionais. Diante desses desafios, este projeto propõe uma abordagem inovadora: utilizar técnicas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina para prever a dificuldade dos itens de Ciências da Natureza com base em suas características textuais, eliminando — ou ao menos reduzindo — a necessidade de pré-testes presenciais.

2 Sobre os dados

O conjunto de dados utilizado neste estudo é composto por 675 questões de Ciências da Natureza aplicadas no ENEM entre 2009 e 2023. As questões foram enriquecidas com 23 variáveis provenientes dos dados públicos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), incluindo:

- Texto do enunciado e alternativas;
- Identificadores de área e tema;

- Parâmetros da Teoria de Resposta ao Item (TRI): discriminação (a), dificuldade (b) e acerto ao acaso (c);
- Características técnicas como cor da prova, tipo de item, adaptação, entre outros.

Uma análise descritiva revelou que os enunciados possuem, em média, 688 caracteres, com um desvio padrão de 338. A questão mais longa possui mais de 2.400 caracteres. O tamanho total da questão (enunciado + alternativas) tem uma média de cerca de 893 caracteres.

Em relação ao parâmetro de dificuldade `NU_PARAM_B`, a distribuição dos dados não é normal, conforme verificado pelo teste de Shapiro-Wilk (estatística = 0,805, p-valor $< 0,001$). Além disso, a correlação entre o tamanho da questão e sua dificuldade é praticamente nula (correlação de Pearson $r = 0,027$), indicando que o comprimento do item não é um bom preditor da dificuldade por si só.

2.a Sobre os microdados Embora os microdados do ENEM contenham informações extremamente ricas — como desempenho individual, escola, localização, redação e perfil socioeconômico — eles não foram utilizados diretamente nesta análise. A decisão de excluí-los foi baseada em dois fatores principais:

- **Foco no texto das questões**: O objetivo deste estudo é avaliar se a dificuldade pode ser prevista apenas a partir das características linguísticas dos itens, sem a necessidade de dados sensíveis ou contextuais dos participantes.
- **Privacidade e generalização**: Evitar dependência de dados que envolvam informações pessoais permite que o modelo seja mais facilmente generalizado para outros exames ou contextos avaliativos.

Ainda assim, é importante reconhecer a relevância desses microdados. Eles contêm informações de milhões de participantes por ano — por exemplo, em 2023 houve mais de **3,9 milhões** de inscritos confirmados. Essa base pode, no futuro, ser explorada para:

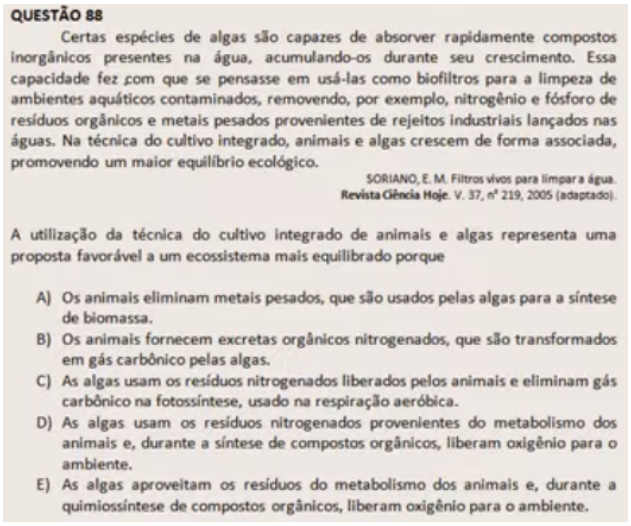
- Validação indireta dos modelos desenvolvidos
- Análises complementares de viés
- Estudos de impacto educacional e regionalidade

2.b O Parâmetro `NU_PARAM_B` O `NU_PARAM_B` é um dos parâmetros da **Teoria de Resposta ao Item (TRI)** utilizado no ENEM para quantificar a dificuldade de uma questão. Corresponde ao parâmetro b da TRI e representa o ponto na escala de proficiência em que a probabilidade de acerto é de 50%, considerando os demais parâmetros do item.

2.b.1 Interpretação

- **Escala**:
  - Valores negativos: questões **mais fáceis** (ex.: `NU_PARAM_B = -1.80` no item mais fácil da amostra).
  - Valores positivos: questões **mais difíceis** (ex.: `NU_PARAM_B = 2.99` no item mais difícil).
  - $b = 0$: dificuldade média (proficiência mediana para 50% de acerto).
- **Relação com outros parâmetros**:
  - Parâmetro a: discriminação (sensibilidade do item a diferentes níveis de proficiência).
  - Parâmetro c: probabilidade de acerto ao acaso ("chute").

2.c Questão mais fácil e mais difícil A questão mais fácil tem `NU_PARAM_B` igual -1.80092 é do ENEM 2011 Caderno Amarelo



QUESTÃO 88

Certas espécies de algas são capazes de absorver rapidamente compostos inorgânicos presentes na água, acumulando-os durante seu crescimento. Essa capacidade fez com que se pensasse em usá-las como biofiltros para a limpeza de ambientes aquáticos contaminados, removendo, por exemplo, nitrogênio e fósforo de resíduos orgânicos e metais pesados provenientes de rejeitos industriais lançados nas águas. Na técnica do cultivo integrado, animais e algas crescem de forma associada, promovendo um maior equilíbrio ecológico.

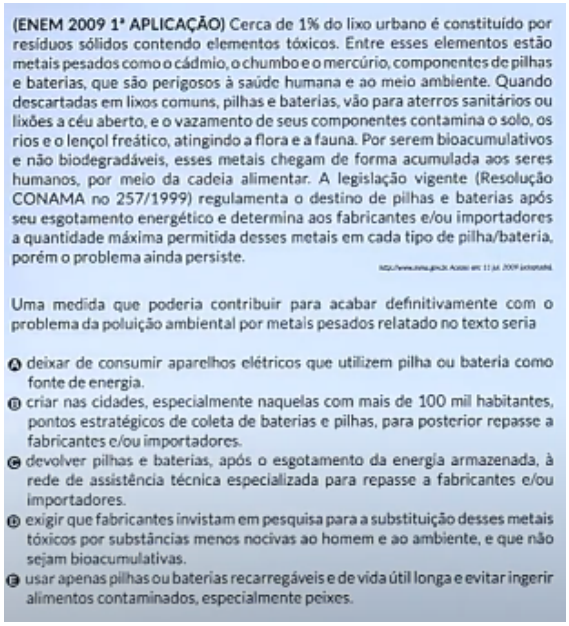
SORIANO, E. M. Filtros vivos para limpar a água. **Revista Ciência Hoje**. V. 37, nº 219, 2005 (adaptado).

A utilização da técnica do cultivo integrado de animais e algas representa uma proposta favorável a um ecossistema mais equilibrado porque

A) Os animais eliminam metais pesados, que são usados pelas algas para a síntese de biomassa.
B) Os animais fornecem excretas orgânicos nitrogenados, que são transformados em gás carbônico pelas algas.
C) As algas usam os resíduos nitrogenados liberados pelos animais e eliminam gás carbônico na fotossíntese, usado na respiração aeróbica.
D) As algas usam os resíduos nitrogenados provenientes do metabolismo dos animais e, durante a síntese de compostos orgânicos, liberam oxigênio para o ambiente.
E) As algas aproveitam os resíduos do metabolismo dos animais e, durante a quimiossíntese de compostos orgânicos, liberam oxigênio para o ambiente.

Figura 1: Questão mais fácil da nossa amostra

Já a questão mais difícil tem `NU_PARAM_B` igual 2.99298 é do ENEM 2009 Caderno Azul



(ENEM 2009 1ª APLICAÇÃO) 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 no 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.

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 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.
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 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.
E usar apenas pilhas ou baterias recarregáveis e de vida útil longa e evitar ingerir alimentos contaminados, especialmente peixes.

Figura 2: Questão mais difícil da nossa amostra

2.d Provas LEDOR O ENEM LEDOR é um recurso de acessibilidade que permite a participantes com deficiência visual, dislexia ou autismo ter as questões lidas em voz alta por um profissional. Esse ledor segue regras rígidas: não pode interpretar o conteúdo, apenas repetir o texto exatamente como está.

Em algumas edições do ENEM, como 2013, 2015 e 2018, houve falhas na implementação desse recurso. Em 2013, muitos locais de prova não tinham ledores disponíveis. Já em 2015, candidatos que solicitaram o auxílio não o receberam. Em 2018, ocorreram erros na distribuição, com alguns participantes recebendo ledores sem necessidade, enquanto outros ficaram sem.

Para este projeto, testamos modelos de IA como ChatGPT, Gemini e Claude para adaptar as questões do ENEM que não tiveram LEDOR. Primeiro, enviamos questões no formato LEDOR e depois enviamos a questão não LEDOR, pedimos para transformá-las. Os resultados mostraram que os modelos conseguem fazer essa conversão, mas ainda precisam de ajustes para evitar simplificações excessivas ou mudanças no significado original.

3 BERTopic

3.a Oque é o BERTopic O BERTopic é uma técnica avançada de modelagem de tópicos que integra modelos de linguagem modernos, como o BERT, com algoritmos de agrupamento (clustering) para identificar e organizar temas em grandes volumes de textos. Seu funcionamento ocorre em quatro etapas principais:

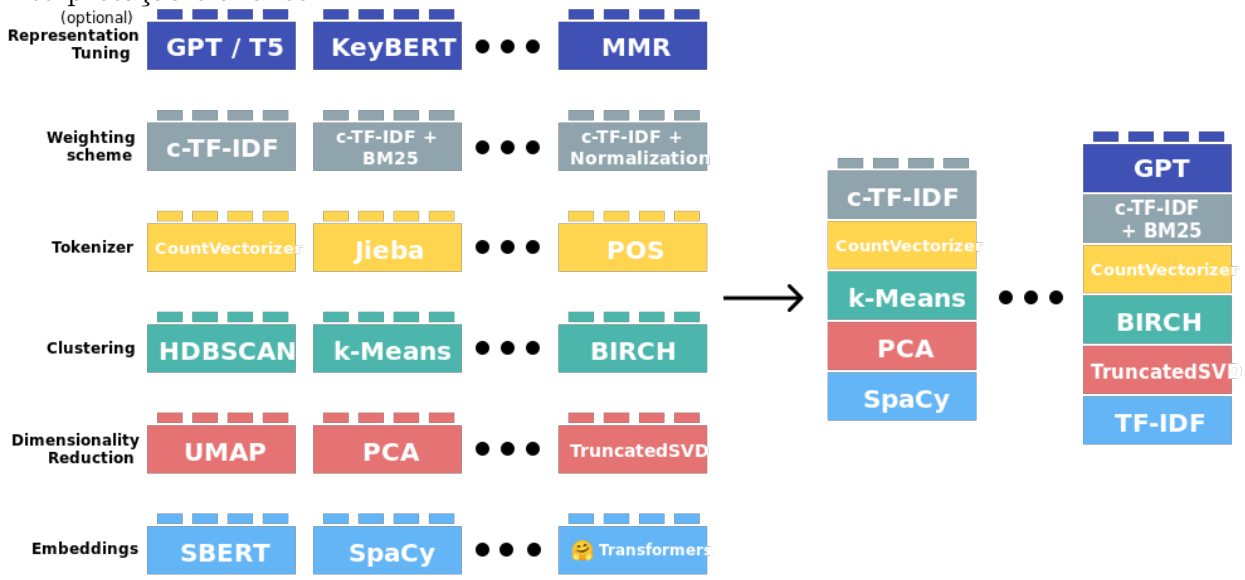
Primeiramente, na fase de pré-processamento e geração de embeddings, o BERTopic utiliza modelos como BERT ou Sentence-BERT para transformar textos em vetores numéricos (embeddings) que capturam o significado semântico do conteúdo. Esses embeddings servem

como representações numéricas dos documentos, preservando suas características linguísticas essenciais.

Em seguida, como esses vetores possuem alta dimensionalidade, o BERTopic aplica técnicas de redução dimensional, como o UMAP, para simplificar a estrutura dos dados sem perder a relação semântica entre os textos. Este passo é crucial para otimizar o processamento posterior.

Na terceira etapa, o algoritmo HDBSCAN é empregado para agrupar os documentos em clusters naturais, identificando os tópicos presentes no corpus. O HDBSCAN é particularmente eficaz por sua capacidade de lidar com documentos que não se encaixam claramente em nenhum tópico específico (outliers), oferecendo maior flexibilidade na organização do conteúdo.

Finalmente, para caracterizar cada tópico identificado, o BERTopic utiliza variações do TF-IDF (como o c-TF-IDF) para extrair as palavras-chave mais representativas de cada agrupamento. Este processo gera descrições concisas e significativas dos tópicos, facilitando sua interpretação e análise.



3.b Otimização de Hiperparâmetros no BERTopic Em nosso projeto de modelagem de tópicos utilizando o *BERTopic*, realizamos uma busca abrangente pelos melhores parâmetros para os algoritmos *UMAP* e *HDBSCAN*. Esta investigação sistemática nos permitiu explorar diversas combinações de configurações com o objetivo de otimizar a qualidade da nossa análise textual.

Para o algoritmo *UMAP*, responsável pela redução dimensional, testamos diferentes valores para seus principais parâmetros. Avaliamos o `n_neighbors` com valores 5, 10, 15 e 30, que controla o balanço entre a estrutura local e global dos dados. Também examinamos o `min_dist` com valores 0.0, 0.1 e 0.3, que regula a densidade dos agrupamentos no espaço reduzido.

No caso do *HDBSCAN*, algoritmo encarregado do agrupamento, investigamos o `min_cluster_size` com valores 5, 8 e 10, que define o tamanho mínimo dos agrupamentos. Além disso, testamos

o `min_samples` com valores 1, 2 e 5, que influencia diretamente na sensibilidade da formação de *clusters*.

Para avaliar o desempenho de cada combinação de parâmetros, estabelecemos critérios rigorosos. Analisamos cuidadosamente a coerência semântica dos tópicos gerados, a distribuição de documentos entre os *clusters* formados, o número ideal de agrupamentos naturais e a estabilidade dos resultados obtidos em diferentes execuções.

Essa abordagem metodológica trouxe benefícios significativos para nossa pesquisa. Permitiu-nos compreender em profundidade como cada parâmetro afeta a qualidade final dos tópicos, encontrar o ponto ótimo entre granularidade e generalização, reduzir consideravelmente a ocorrência de ruídos e tópicos irrelevantes, e maximizar a interpretabilidade dos resultados obtidos.

A combinação ideal desses parâmetros revelou-se fundamental para produzir tópicos semanticamente coerentes que representassem adequadamente os temas presentes em nosso *corpus* textual. Conseguimos assim garantir a especificidade necessária para análises precisas, mantendo ao mesmo tempo a abrangência suficiente para capturar os padrões mais significativos em nossos dados.

Na Tabela a seguir (Figura 3), apresentamos os quatro tópicos com maior frequência identificados após a aplicação do BERTopic. Os nomes e palavras representativas indicam a coerência temática de cada grupo, com destaque para tópicos relacionados a água, organismos vivos e reações químicas.

Topic	Count	Name	Representation	Representative_Docs
-1	59	-1_água_anticorpos_membrana_célula	[água, anticorpos, membrana, célula, glicose, ...	[moedas despertam interesse colecionadores num...
0	45	0_temperatura_água_vazão_umidade	[temperatura, água, vazão, umidade, relativa, ...	[pessoa lendo manual ducha acabou adquirir cas...
1	38	1_machos_fêmeas_peixes_animais	[machos, fêmeas, peixes, animais, espécies, vi...	[hermaphroditic demasculinized frogs after exp...
2	36	2_carga_positiva_chumbo_reação	[carga, positiva, chumbo, reação, concreto, ca...	[aplicações ambientais persulfato remediação á...

Figura 3: Tabela de tópicos com palavras representativas

O dendrograma de *Hierarchical Clustering* (Figura 4) mostra a similaridade entre os tópicos obtidos, agrupando os que compartilham maior proximidade semântica. A visualização evidencia agrupamentos naturais e bem definidos, reforçando a validade dos clusters formados.

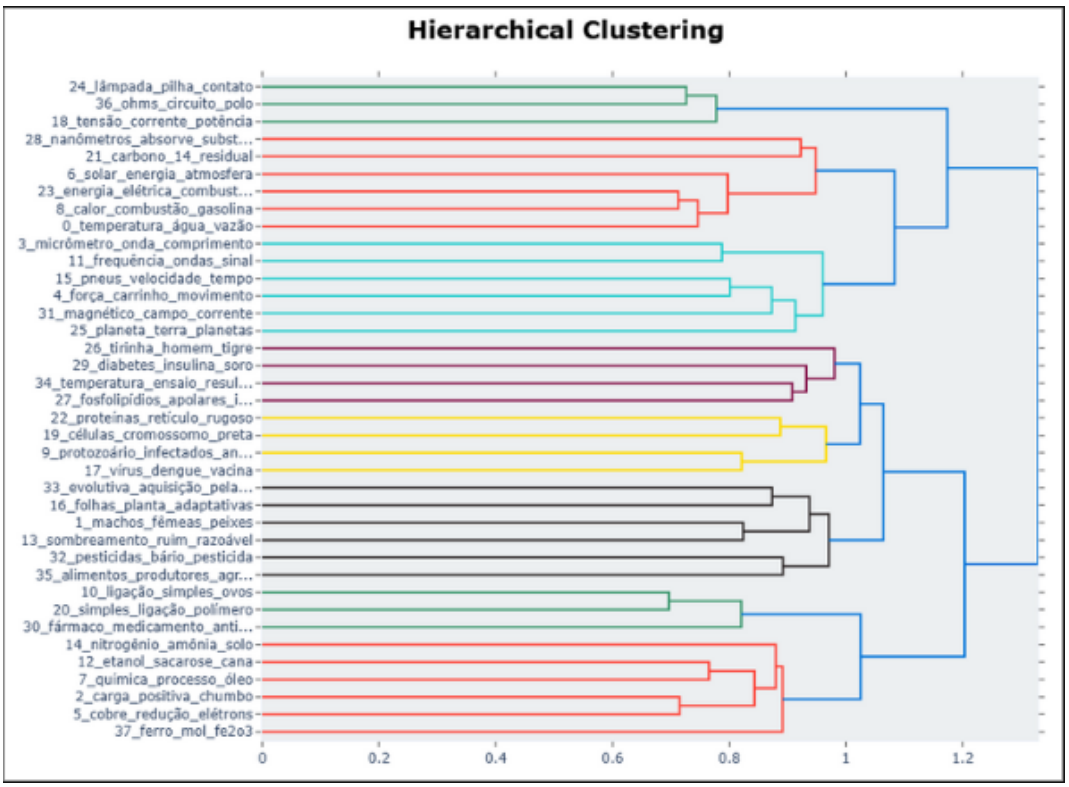


Figura 4: Dendrograma de tópicos com agrupamentos hierárquicos

Por fim, o *Intertopic Distance Map* (Figura 5) apresenta os tópicos em um espaço bidimensional reduzido, permitindo visualizar a distribuição e sobreposição entre os grupos. A dispersão equilibrada e a distância entre tópicos indicam que a segmentação atingiu um bom nível de granularidade sem redundâncias excessivas.

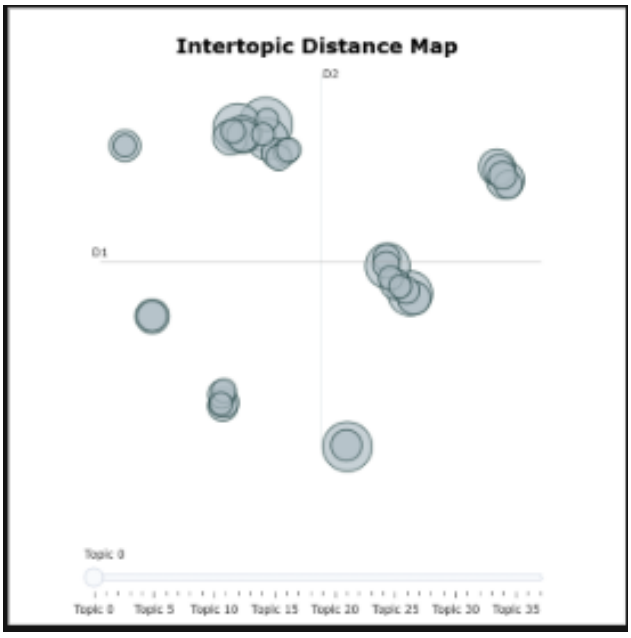


Figura 5: Mapa de distância entre tópicos (Intertopic Distance Map)

4 Metodologias

A metodologia adotada neste projeto envolveu múltiplas abordagens baseadas em Processamento de Linguagem Natural (PLN), com foco na análise linguística das questões do ENEM e na aplicação de modelos semânticos avançados para prever sua dificuldade.

4.a Análise de Atributos Linguísticos Como etapa inicial, foi realizada a extração de atributos linguísticos dos enunciados das questões com base no artigo do NILC-Metrix (**?**), adaptado para o português brasileiro. Utilizando a biblioteca `spaCy`, foram implementadas métricas relacionadas à inteligibilidade, coesão textual, complexidade sintática e lexical.

NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese

Sidney Evaldo Leal [1] sidleal@gmail.com
Magali Sanches Duran [1] magali.duran@gmail.com
Carolina Evaristo Scarton [2] c.scarton@sheffield.ac.uk
Nathan Siegle Hartmann [3] nathanshartmann@gmail.com
Sandra Maria Aluísio [1] sandra@icmc.usp.br

[1] Instituto de Ciências Matemáticas e de Computação - University of São Paulo, São Paulo, Brazil
[2] The University of Sheffield, Sheffield, UK
[3] Itaú Unibanco, São Paulo, Brazil

arXiv:2201.03445v1 [cs.CL] 17 Dec 2021

Abstract

This paper presents and makes publicly available the NILC-Metrix, a computational system comprising 200 metrics proposed in studies on discourse, psycholinguistics, cognitive and computational linguistics, to assess textual complexity in Brazilian Portuguese (BP). These metrics are relevant for descriptive analysis and the creation of computational models and can be used to extract information from various linguistic levels of written and spoken language. The metrics in NILC-Metrix were developed during the last 13 years, starting in 2008 with Coh-Metrix-Port, a tool developed within the scope of the PorSimples project. Coh-Metrix-Port adapted some metrics to BP from the Coh-Metrix tool that computes metrics related to cohesion and coherence of texts in English. After the end of PorSimples in 2010, new metrics were added to the initial 48 metrics of Coh-Metrix-Port. Given the large number of metrics, we present them following an organisation similar to the metrics of Coh-Metrix v3.0 to facilitate comparisons made with metrics in Portuguese and English. In this paper, we illustrate the potential of NILC-Metrix by presenting three applications: (i) a descriptive analysis of the differences between children's film subtitles and texts written for Elementary School I[1] and II (Final Years)[2]; (ii) a new predictor of textual complexity for the corpus of original and simplified texts of the PorSimples project; (iii) a complexity prediction model for school grades, using transcripts of children's story narratives told by teenagers. For each application, we evaluate which groups of metrics are more discriminative, showing their contribution for each task.

1 Introduction

A set of metrics called NILC-Metrix was developed both in funded projects, involving multiple re-

Figura 6: Referência teórica e extração de atributos com base no NILC-Metrix

As métricas extraídas foram organizadas em quatro categorias principais. A primeira delas, **Distribuição e estrutura lexical**, contempla medidas como *Type-Token Ratio* (diversidade vocabular), número de sentenças (fragmentação do texto) e o tamanho médio das sentenças (indicador de complexidade sintática).

A segunda categoria, **Distribuição morfossintática**, considera a proporção de diferentes classes gramaticais, como verbos (*Verb-Token Ratio*), substantivos (*Noun-Token Ratio*), adjetivos (*Adjective-Token Ratio*) e pronomes (*Pronoun-Token Ratio*), fornecendo um retrato morfológico do texto.

A terceira, voltada à **fluidez e formalidade**, inclui a proporção de palavras funcionais sem conteúdo semântico (*Stopword-Token Ratio*), pausas inferidas por pontuação (*Pausality*) e uma estimativa de informalidade baseada na ocorrência de erros ortográficos.

Por fim, a dimensão de **complexidade e semântica textual** agrega o *Índice de Brunet* (diversidade lexical ajustada ao tamanho), o *Gunning-Fog Index* (indicador de inteligibilidade, sendo valores acima de 12 considerados complexos) e a *especificidade semântica*, medida pela proporção de entidades nomeadas (como tempo e espaço).

4.b Modelos Contextuais e Similaridade de Texto Além da análise linguística tradicional, utilizamos modelos de linguagem para converter os textos das questões em representações vetoriais, ou *embeddings*, que pudessem ser usados em modelos preditivos.

Duas arquiteturas principais foram consideradas: os **modelos estáticos**, como Word2Vec e GloVe, que geram representações fixas para cada palavra, independentemente do contexto; e os **modelos contextuais**, como BERT e Sentence-BERT (SBERT), que utilizam mecanismos de autoatenção para atribuir vetores que variam conforme o significado contextual da palavra na frase. Essa última abordagem foi priorizada por capturar melhor nuances semânticas e relações implícitas, características importantes para entender a dificuldade de questões mais abstratas ou interdisciplinares.

4.c Fine-Tuned BERT Para maximizar o desempenho preditivo, realizamos o *fine-tuning* do modelo BERT em dados educacionais específicos, ajustando seus pesos para a tarefa de regressão do grau de dificuldade dos itens. Essa adaptação foi realizada com técnicas de validação cruzada e regularização, permitindo maior sensibilidade às estruturas próprias dos itens de Ciências da Natureza, sem perda de generalização.

5 Word2Vec e BERT Embeddings

Nesta etapa do projeto, buscamos compreender o comportamento semântico dos enunciados e alternativas por meio da geração de embeddings com duas abordagens distintas: Word2Vec e BERT. A intenção foi analisar como as diferentes arquiteturas representam semanticamente os itens e se há relação entre essas representações e a dificuldade das questões, mensurada pelo parâmetro `NU_PARAM_B`.

5.a Arquitetura Word2Vec O Word2Vec é uma técnica de word embedding estática baseada em janelas de contexto. Constituída por redes neurais rasas, os modelos deste tipo criam representações vetoriais estáticas, isto é, uma única representação vetorial de alta dimensão para cada palavra em seu treinamento, compensando essa limitação com um custo computacional baixo quando contraposta aos modelos baseados em Transformers.

As duas principais arquiteturas Word2Vec são o *CBOW* (Continuous Bag of Words), treinada sobre a tarefa prever uma palavra central mascarada a partir do contexto; e o *Skip-Gram*, que faz o oposto: tenta prever o contexto a partir da palavra central. A Figura 7 ilustra visualmente essas duas abordagens, demonstrando como o modelo aprende representações vetoriais úteis para capturar relações semânticas simples.

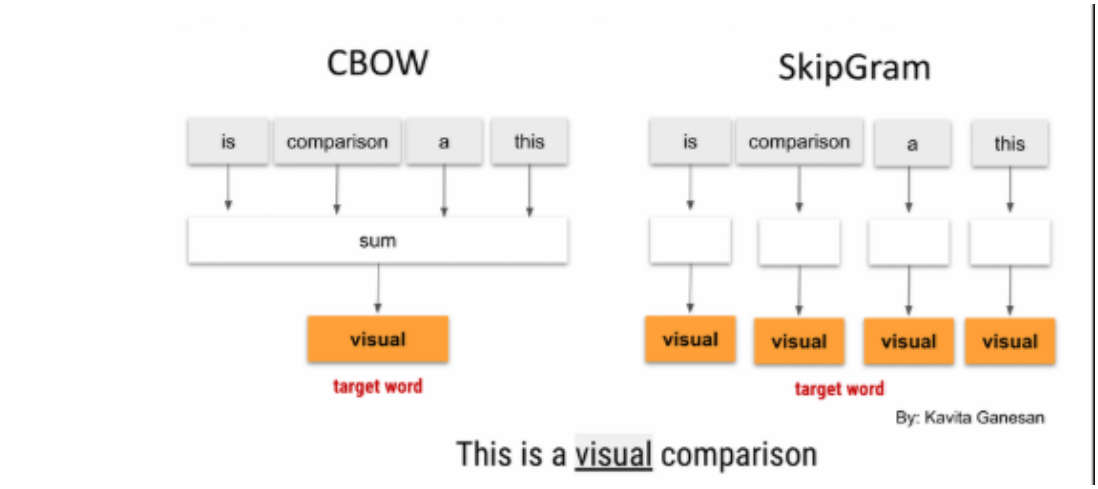


Figura 7: Arquitetura Word2Vec: CBOW e Skip-Gram
Fonte: Kavita Ganesan

5.b BERT (Bidirectional Encoder Reoresentations from Transformers Baseada na arquitetura transformers (7), BERT é uma família de modelos construídos sobre os mecanismos de autoatenção bidirecional. Dessa forma, estes modelos são capazes de capturar relações semânticas mais profundas e dependências de longo prazo ao considerar o contexto inteiro da sentença, tanto à esquerda quanto à direita da palavra-alvo. Vale o destaque, ainda, ao fato de que os modelos BERT são treinados com duas tarefas: a predição de um token mascarado -análogo às arquiteturas Word2Vec- e a predição de sentanças, a qual gera um senso maior de encadeamento lógico ao modelo.

Para este estudo, utilizamos principalmente do BERTimbau (8), especificamente desenvolvida para o português brasileiro, de forma que sua performance em tarefas downstream seja superior quando comparados aos modelos multilíngues.

A Figura 8 apresenta a arquitetura geral do BERT (Bidirectional encoding representations from transformers), com destaque para os blocos de atenção e normalização aplicados em camadas empilhadas.

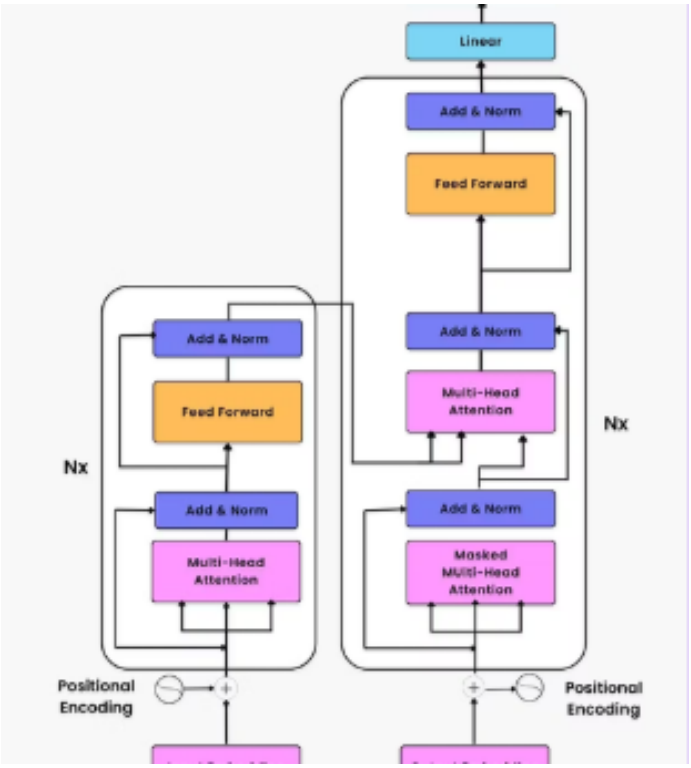


Figura 8: Arquitetura geral do BERT baseada em transformadores

5.c Projeções UMAP dos Embeddings A Figura 9 apresenta a projeção bidimensional dos embeddings gerados a partir do modelo Word2Vec, utilizando a técnica de redução de dimensionalidade UMAP (12). Cada ponto no gráfico representa uma questão do ENEM, posicionada no espaço de acordo com a representação semântica vetorial dos textos das questões. As cores dos pontos correspondem ao valor do parâmetro de dificuldade `NU_PARAM_B`, em uma escala que varia do vermelho (itens mais fáceis) ao azul escuro (itens mais difíceis). Observa-se que há uma organização semântica coerente entre os itens, com formações de agrupamentos locais — reflexo da capacidade do Word2Vec em capturar padrões de vocabulário e contexto limitado. No entanto, a dispersão das cores dentro desses grupos indica que a representação vetorial obtida por esse modelo pode não ser suficientemente sensível para discriminar níveis de dificuldade de maneira clara. Essa limitação se deve ao fato de o Word2Vec gerar embeddings independentes do contexto completo da sentença, o que compromete a captura de nuances linguísticas mais sutis. Esse resultado reforça a hipótese de que modelos contextuais, como o BERT, podem oferecer maior poder explicativo na análise da dificuldade dos itens, ao considerar relações semânticas mais profundas entre os termos. A Figura 10 exibe a projeção UMAP dos embeddings produzidos pelo modelo BERT, aplicado sobre os textos das questões de Ciências da Natureza do ENEM. Assim como no gráfico anterior, cada ponto representa uma questão e está posicionado no plano de acordo com sua similaridade semântica, agora capturada por um modelo contextual. A escala de cores reflete os valores do parâmetro de dificuldade `NU_PARAM_B`, indo do vermelho (questões mais fáceis) ao azul escuro (questões mais difíceis). Visualmente, nota-se uma organização

mais densa e coesa dos pontos, com regiões semânticas mais bem definidas em comparação à projeção obtida com Word2Vec. Além disso, há indícios de padrões estruturais mais alinhados à escala de dificuldade, com maior consistência cromática em certas áreas do gráfico. Isso sugere que o BERT, ao considerar o contexto completo em que cada palavra está inserida, foi mais eficaz em gerar representações que refletem aspectos linguísticos associados à complexidade dos itens. Portanto, os embeddings contextuais revelaram-se mais adequados para tarefas que exigem maior sensibilidade semântica, como a predição da dificuldade de questões educacionais.

Figura 9: Projeção UMAP dos embeddings gerados com Word2Vec

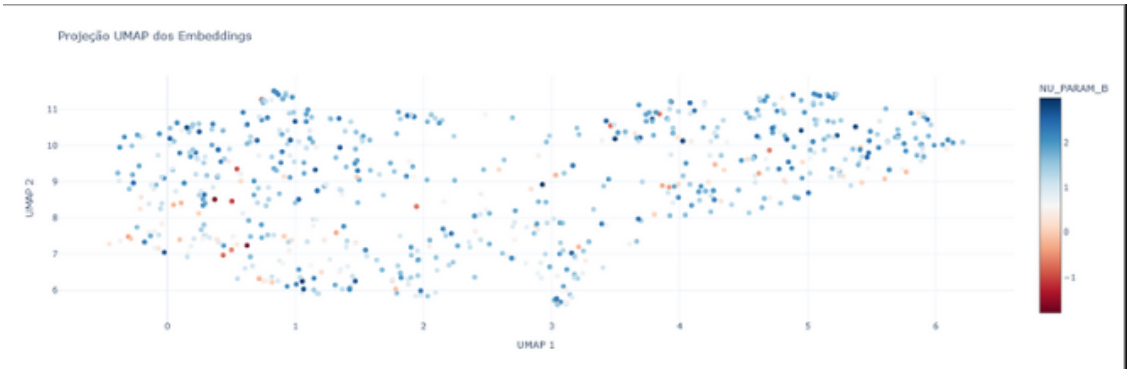


Figura 10: Projeção UMAP dos embeddings gerados com BERT

As projeções revelam que, embora ambas as abordagens apresentem estrutura, os embeddings do BERT mostraram uma distribuição semântica mais densa e consistente com o parâmetro de dificuldade, evidenciando sua superioridade em capturar nuances do texto.

5.d Similaridade entre Embeddings A Figura 11 apresenta a relação entre a média de similaridade dos embeddings das alternativas de cada item e o respectivo parâmetro de dificuldade `NU_PARAM_B`. Cada ponto do gráfico representa um item da prova, em que o eixo horizontal corresponde à dificuldade estimada e o eixo vertical expressa a média das similaridades entre as alternativas, calculada com base nos embeddings extraídos por um modelo contextual. Observa-se uma tendência sutil de que, à medida que as alternativas se tornam semanticamente mais semelhantes entre si, a dificuldade do item tende a aumentar.

Essa observação é coerente com a hipótese de que alternativas muito parecidas podem gerar maior ambiguidade para o candidato, exigindo maior capacidade de discriminação conceitual. A correlação positiva identificada — embora baixa — reforça o potencial da análise de similaridade semântica como um indicador indireto da dificuldade de questões, oferecendo uma métrica quantitativa adicional para calibragem automatizada de itens.

Figura 11: Correlação entre similaridade média entre alternativas e a dificuldade

A Figura 12 ilustra a correlação entre a média de similaridade semântica entre o gabarito (alternativa correta) e o enunciado, e o parâmetro de dificuldade `NU_PARAM_B`. Cada ponto representa um item, com a dificuldade posicionada no eixo horizontal e a similaridade dos embeddings no eixo vertical. Nota-se uma leve tendência decrescente: itens cuja alternativa correta é semanticamente mais distante do enunciado tendem a apresentar maior dificuldade. Essa evidência sugere que, quanto mais implícita ou sutil for a relação entre o enunciado e a resposta correta, maior será o desafio cognitivo imposto ao candidato. A correlação negativa observada reforça essa hipótese e destaca o valor da análise de embeddings contextuais como ferramenta para antecipar a dificuldade de um item com base em sua estrutura linguística e semântica.

Figura 12: Similaridade entre gabarito e enunciado versus dificuldade

A Figura 13 apresenta a correlação entre a similaridade média dos embeddings das alternativas incorretas com o enunciado e o parâmetro de dificuldade `NU_PARAM_B`. Cada ponto

representa um item da prova, sendo o eixo horizontal o valor da dificuldade estimada e o eixo vertical a média de similaridade semântica entre o enunciado e as opções erradas, calculada a partir de embeddings contextuais. Observa-se uma leve correlação positiva: à medida que as alternativas erradas se tornam semanticamente mais próximas do enunciado, a dificuldade do item tende a aumentar. Esse comportamento é consistente com a ideia de que, quanto mais plausíveis as alternativas incorretas parecerem em relação ao contexto apresentado, maior será o esforço de discriminação necessário por parte do candidato. Portanto, esse tipo de análise pode ser útil na elaboração de itens mais desafiadores e na identificação de questões potencialmente ambíguas.

Figura 13: Similaridade entre respostas erradas e enunciado versus dificuldade

A Figura **??** mostra o histograma da métrica `diff_similarity`, definida como a diferença entre a similaridade do enunciado com a alternativa correta (gabarito) e a média de similaridade do enunciado com as alternativas incorretas. Valores negativos indicam que o gabarito está semanticamente mais distante do enunciado do que as alternativas erradas — um cenário potencialmente mais desafiador para o candidato, já que a resposta correta se destaca menos em relação às demais. A distribuição é assimétrica e concentrada em valores negativos, sugerindo que, na maioria das questões, o gabarito tende a ser menos semanticamente próximo do enunciado do que as respostas incorretas.

Complementando essa análise, a Figura 15 apresenta a relação entre essa diferença de similaridade e o parâmetro de dificuldade `NU_PARAM_B`. Cada ponto representa um item, posicionando a dificuldade no eixo horizontal e a métrica `diff_similarity` no eixo vertical. A tendência observada indica que, quanto menor essa diferença (ou seja, quanto mais o gabarito se confunde semanticamente com as alternativas incorretas), maior tende a ser a dificuldade do item. Isso confirma a hipótese de que o distanciamento semântico do gabarito em relação ao enunciado — comparado às alternativas erradas — pode ser um fator determinante para tornar uma questão mais difícil, servindo como possível preditor adicional da complexidade de itens avaliativos.

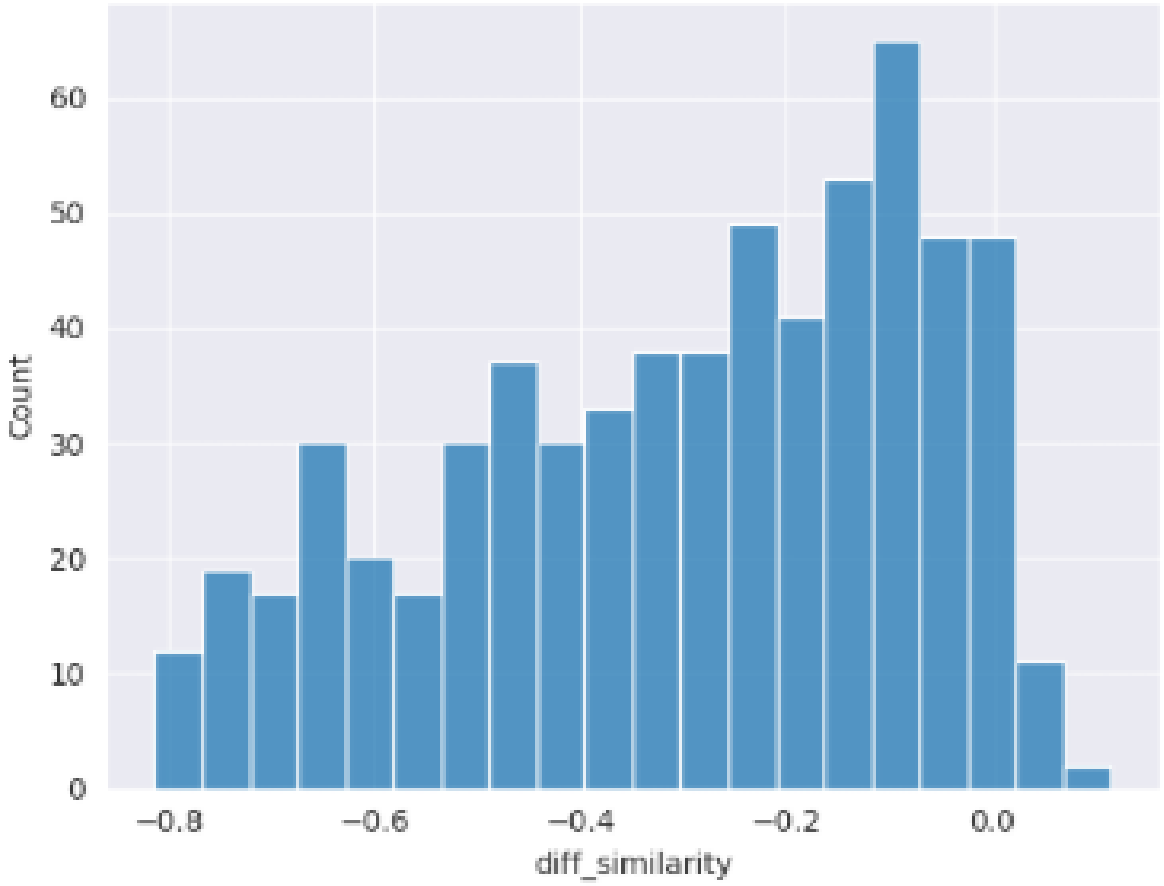


Figura 14: Diferença de similaridade entre gabarito/enunciado e erradas/enunciado

Figura 15: Diferença de similaridade entre gabarito/enunciado e erradas/enunciado

6 Regressores

Nesta etapa, buscamos avaliar o potencial preditivo das representações vetoriais geradas pelos modelos de linguagem aplicadas aos enunciados das questões. A metodologia consistiu em extrair os embeddings das palavras dos enunciados e utilizar das dimensões destes como variáveis explicativas em modelos de regressão, com o parâmetro de dificuldade `NU_PARAM_B` como variável dependente. O objetivo era verificar se os embeddings, ao capturarem aspectos semânticos dos textos, poderiam ser utilizados para prever a dificuldade das questões.

Para isso, foram utilizados dois tipos principais de modelos de linguagem. O primeiro grupo abrange os modelos Word2Vec nas variantes CBOW, Skip-Gram, e GloVe com diferentes configurações de dimensionalidade (50, 100 e 300), conforme disponibilizado no repositório de embeddings do NILC (10). O segundo grupo incluiu os embeddings extraídos a partir do BERT, utilizados com o intuito de comparar o desempenho de representações contextuais frente às representações estáticas. Mais adiante, uma arquitetura BERT será ajustada à tarefa de predição, compondo a metodologia mais adequada para este tipo de modelo.

Quanto aos modelos de regressão, adotamos os métodos Linear e Lasso, que permitiram avaliar não apenas a capacidade preditiva, mas também a relevância das variáveis utilizadas. O Lasso, em particular, foi útil por incorporar regularização, promovendo seleção de variáveis e combatendo o sobreajuste. Aqui, destaca-se que o parâmetro de regularização α foi estimado e validado por meio de um GridSearch nos dados de treino para todos os experimentos com Lasso.

O processo de pré-processamento incluiu o ajuste dimensional dos embeddings para garantir a consistência entre os vetores de entrada dos modelos. Em seguida, os dados foram divididos em conjuntos de treino e teste por meio de divisões padronizadas, assegurando reprodutibilidade e comparabilidade entre os experimentos. Com isso, foi possível monitorar o desempenho das regressões com base em métricas como o erro quadrático médio e o coeficiente de determinação (R^2), entre outras.

Para as arquiteturas do tipo Word2Vec, seguiu-se o protocolo sugerido em (1), que prevê algumas medidas de pré-processamento, dentre as quais: a remoção de stopwords (feita a partir da biblioteca SpaCy, a conversão de todas as letras para caixa baixa e a normalização dos numerais - uma vez que a grande quantidade de numerais aumentaria a esparsidade da matriz que gera as representações vetoriais.

7 Resultados Parciais

A Tabela a seguir resume os principais resultados obtidos nos experimentos de regressão aplicados aos embeddings das questões. Foram comparadas três abordagens distintas: CBOW com 300 dimensões seguindo um protocolo específico de vetorização, CBOW sem protocolo

de vetorização, e embeddings contextuais extraídos do BERT.

Tabela 1: Desempenho (MSE e Correlação no conjunto de teste) para diferentes embeddings e regressões

Embedding	**Dimensão**	**Regressão**	**MSE (teste)**	**Correlação (teste)**
CBOW	300	Linear	1.2082	0.2638
		Lasso	0.5820	0.4192
CBOW	100	Linear	0.7507	0.1845
		Lasso	0.6040	0.3185
CBOW	50	Linear	0.6435	0.2601
		Lasso	0.5819	0.3977
GloVe	300	Linear	0.9399	0.2794
		Lasso	0.5801	0.4639
GloVe	100	Linear	0.7239	0.2078
		Lasso	0.6166	0.3279
GloVe	50	Linear	0.5959	0.3367
		Lasso	0.5957	0.3673
Skip-Gram	300	Linear	1.1411	0.2039
		Lasso	0.5895	0.4292
BERT	—	Linear	1.9000	0.1261
		Lasso	0.5948	0.4107

No caso do **CBOW com 300 dimensões e protocolo de vetorização**, os resultados indicaram um desempenho superior do modelo Lasso em comparação ao modelo de regressão linear, com erro quadrático médio (MSE) de 0,581 e correlação de 0,4192, frente a um MSE de 1,208 e correlação de 0,265 obtidos pela regressão linear. Além disso, para todos os modelos aqui testados, os embeddings com 300 dimensões obtiveram melhores resultados, sendo seguidos pelos modelos com 50 e 100 dimensões respectivamente.

As arquiteturas **GloVe** e **SKIP-GRAM**, analogamente, também apresentam os melhores resultados em seus modelos maiores (300 dimensões). Aqui, vale o destaque para o GloVe, cujas métricas obtidas foram as melhores dentre os modelos estáticos.

Quando removido o protocolo de vetorização, ainda com CBOW e 300 dimensões, observou-se uma leve melhora no modelo linear (MSE de 1,101 e correlação de 0,295), mas uma pequena queda no desempenho do Lasso (MSE de 0,610 e correlação de 0,389), sugerindo que o processo de vetorização influencia diretamente na estrutura da informação utilizada para previsão.

Por fim, os resultados com **embeddings gerados pelo BERT** apresentaram o maior erro médio quadrático no modelo linear (MSE de 1,899), com correlação bastante reduzida (0,12). O modelo Lasso, embora com desempenho um pouco melhor (MSE de 0,594 e correlação de 0,41), ainda ficou abaixo dos resultados obtidos com CBOW.

Esses resultados preliminares sugerem que, neste cenário específico, os embeddings estáticos do Word2Vec (GloVe) foram mais eficazes na tarefa de predição da dificuldade em relação aos embeddings contextuais do BERT, especialmente quando processados com vetorização padronizada e com um maior número de dimensões. O modelo Lasso, por sua vez, demonstrou-se mais robusto e informativo do que a regressão linear simples em todos os cenários analisados.

8 Fine-tuned BERT

Na etapa final dos experimentos, realizamos o *fine-tuning* do modelo BERT com foco na predição do parâmetro de dificuldade dos itens. O objetivo foi adaptar os embeddings contextuais gerados por essa arquitetura às especificidades do domínio educacional e, mais especificamente, às características linguísticas das questões de Ciências da Natureza.

O processo de ajuste fino envolveu três estratégias principais:

- **Reinicialização das últimas camadas:** conforme sugerido em (9) as camadas superiores do modelo foram reinicializadas e treinadas com os dados do nosso corpus, permitindo que o modelo aprendesse representações mais especializadas sem comprometer os conhecimentos linguísticos adquiridos nas fases anteriores de pré-treinamento.

- **Otimização de hiperparâmetros:** realizamos uma busca sistemática pelos melhores hiperparâmetros através da biblioteca Optuna (11), incluindo taxa de aprendizagem, tamanho do *batch*, número de épocas e estratégias de regularização, com o intuito de maximizar o desempenho preditivo e evitar *overfitting*.

- **Avaliação de diferentes variantes do BERT:** testamos múltiplas variantes do modelo, como o BERTimbau e o MiniLM, com o objetivo de comparar desempenho, tempo de treinamento e generalização dos modelos no contexto da tarefa proposta.

Essas abordagens permitiram personalizar o BERT para a tarefa de regressão do parâmetro de dificuldade, explorando ao máximo sua capacidade de capturar nuances semânticas complexas. Os resultados obtidos com os modelos ajustados foram posteriormente comparados aos dos embeddings estáticos e embeddings não ajustados, possibilitando uma análise crítica do impacto do fine-tuning no desempenho final.

9 Resultados do Fine-tuning

Após o processo de fine-tuning dos modelos BERT, avaliamos o desempenho das versões adaptadas ao domínio educacional por meio de métricas de validação em regressão.

O modelo **BERTimbau**, treinado especificamente para a língua portuguesa, apresentou os melhores resultados dentre os modelos testados. Obteve um erro quadrático médio

(MSE) de 0,537 e uma correlação de 0,473 com os valores reais do parâmetro de dificuldade `NU_PARAM_B`. Tais resultados indicam que o modelo conseguiu capturar, de forma eficiente, padrões linguísticos relevantes para estimar a dificuldade dos itens.

Já o **Multilingual BERT**, embora ligeiramente inferior ao BERTimbau, também demonstrou bom desempenho. Seu MSE foi de 0,552, com correlação de 0,447, sugerindo que, apesar de não ser especializado em português, o modelo ainda é capaz de aprender representações úteis para a tarefa, especialmente após o processo de ajuste fino com dados específicos da prova.

Esses resultados confirmam a eficácia do fine-tuning como estratégia para adaptar modelos pré-treinados ao contexto da avaliação educacional, com destaque para modelos linguísticos treinados nativamente em português.

10 Análise das Dimensões da Regressão

Com o objetivo de interpretar semanticamente os resultados dos modelos de regressão, realizamos uma análise das dimensões mais relevantes nos embeddings utilizados como variáveis independentes. A ideia central foi identificar quais palavras estavam mais associadas às dimensões que mais contribuíram — positiva ou negativamente — para a predição do parâmetro de dificuldade `NU_PARAM_B`.

Para isso, selecionamos as dimensões com os maiores pesos (em módulo) atribuídos pelos coeficientes do modelo Lasso e analisamos as palavras que apresentavam maior valor nessas dimensões, com base nos embeddings do modelo Word2Vec. A seguir, construímos nuvens de palavras para visualizar de forma intuitiva essas contribuições.

Figura 16: Nuvens de palavras das dimensões de regressão: à esquerda, termos com contribuição positiva para a predição da dificuldade; à direita, termos com contribuição negativa

A Figura 16 ilustra duas nuvens de palavras. A nuvem à esquerda representa as palavras associadas às dimensões com **contribuição positiva** para o valor predito de dificuldade. Entre os termos mais proeminentes estão "herschell", "celsius" e "inclinação", indicando que tópicos relacionados à física, medições e fenômenos naturais tendem a estar associados a questões mais difíceis.

Já a nuvem à direita mostra palavras associadas às dimensões com **contribuição negativa**, ou seja, que puxam o valor de dificuldade para baixo. Termos como "tomate", "sentidos"

e “pílula” aparecem em destaque, sugerindo a presença de conteúdo mais cotidiano, concreto ou de maior familiaridade para o aluno, o que pode estar relacionado a questões menos complexas cognitivamente.

Essa análise qualitativa complementa os resultados quantitativos e reforça o potencial explicativo dos embeddings como ferramentas para capturar nuances semânticas relevantes na predição da dificuldade de itens.

Referências

[1] JALOTO, Alexandre; PERES, Alexandre José de Souza; ZUANAZZI, Ana Carolina; CAINÃ, Araê; PRIMI, Ricardo. **É possível calibrar os itens do Enem sem pré-teste?** [S.l.]: [s.n.], [2023].

[2] Andrade, D. F.; Soares, T. M.; Soares, J. F. (2013). **Introdução à Teoria de Resposta ao Item: Aplicações no ENEM**. Avaliação Psicológica, 12(2), pp. 139-152.

[3] Ha, L. A.; Yacef, K. (2019). **Predicting Item Difficulty from Text Features in Educational Assessments**. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), pp. 9547-9554.

[4] Belov, D.; Wollack, J. A. (2016). **Comparing the Efficiency of CAT and Pre-Testing Designs**. Journal of Educational Measurement, 53(3), pp. 300-318.

[5] Polak, S.; Łupkowski, P. (2022). **Linguistic Features for Automatic Difficulty Prediction in Science Tests**. Natural Language Engineering, 28(5), pp. 601-623.

[6] Ribeiro, C.; Guimarães, M. H. (2018). **Operational Challenges in Large-Scale Educational Testing: The ENEM Case**. Estudos em Avaliação Educacional, 29(72), pp. 634-658.

[7] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). **BERT: Pre-training of deep bidirectional transformers for language understanding.** In Burstein, J., Doran, C., and Solorio, T., editors, Proceedings of the 2019 Conference of the NorthAmericanChapteroftheAssociationforCom putational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota.

[8] Souza, F., Nogueira, R., and Lotufo, R. (2020). **BERTimbau: pretrained BERT models for Brazilian Portuguese.** In Cerri, R. and Prati, R. C., editors, Proceedings of the 9th Brazilian Conference on Intelligent Systems, BRACIS,RioGrandedoSul, Brazil, October20-23, pages 403–417, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-61377-8_28.

[9] SU, Yu; SONG, Yujia; ZHANG, Zhenyu; XU, Hu; WANG, Yichong; ZHANG, Xi Victoria. **Revisiting Few-sample BERT Fine-tuning.** In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, p. 2069–2082.

[10] HARTMANN, Nathan; FONSECA, Erick; SHULBY, Christopher; TREVISO, Marcos; RODRIGUES, Jessica; ALUISIO, Sandra. **Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks.** 2017. Preprint. Disponível em: https://arxiv.org/abs/1708.06025.

[11] AKIBA, Takuya; SANO, Shotaro; YANASE, Toshihiko; OHTA, Takeru; KOYAMA, Masanori. **Optuna: A Next-generation Hyperparameter Optimization Framework.** Preprint, 2019. Disponível em: https://arxiv.org/abs/1907.10902

[12] McInnes et al., (2018). **UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software**, 3(29), 861, https://doi.org/10.21105/joss.00861