Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 02/Sep/2022; 1st round notif.: 04/Nov/2022; New version: 12/Dec/2022; 2nd round notif.: 07/Mar/2023;
Camera ready: 02/Apr/2023; Edition review: 20/Apr/2023; Available online: 14/May/2023; Published: 14/May/2023;

Avaliação Automática de Redação: Uma revisão sistemática

Title: Automatic Essay Evaluation in Portuguese: A Systematic Review

Tiago Barbosa de Lima
Universidade Federal Rural de Pernambuco
ORCID: 0000-0002-0707-522X
tiago.blima@ufrpe.br

Ingrid Luana Almeida da Silva
Universidade Federal Rural de Pernambuco
ORCID: 0009-0000-9197-7535
ingrid.luana@ufrpe.br

Elyda Laisa Soares Xavier Freitas
Universidade de Pernambuco
ORCID: 0000-0001-7439-9040
elyda.freitas@upe.br

Rafael Ferreira Mello
Universidade Federal Rural de Pernambuco
ORCID: 0000-0003-3548-9670
rafael.mello@ufrpe.br

Resumo

A Avaliação Automática de Redação (do inglês, Automatic Essay Scoring - AES) tem sido tema amplamente explorado na literatura. Ela permite dispensar o esforço humano aplicado na correção de um grande número de redações em um curto espaço de tempo. A maior parte dos trabalhos disponíveis na literatura se concentra no esforço de desenvolver algoritmos que sejam capazes de corrigir automaticamente textos em inglês. No entanto, para a língua portuguesa, essa ainda é uma área que está em desenvolvimento. Neste contexto, este artigo apresenta um Mapeamento Sistemático da Literatura que busca identificar as abordagens de Inteligência Artificial que estão sendo utilizadas para oferecer suporte à avaliação de redações escritas na língua portuguesa. Os principais achados deste artigo incluem os seguintes fatos: (i) as abordagens dos trabalhos selecionados costumam focar no uso de atributos extraídos do texto em vez do uso de modelos pré-treinados baseados em Deep Learning; (ii) existe prevalência de métricas tradicionais, como Precisão, Cobertura e F-Measure na validação dos resultados; (iii) os feedbacks gerados pelas abordagens possuem um baixo detalhamento; e (iv) os artigos selecionados não analisam o impacto prático em aplicações do mundo real.
Palavras-chave: *Correção de Redação; Análise de Conteúdo; Processamento de Linguagem Natural*

Abstract

The literature has vastly explored Automatic Essay Scoring (AES) in the last few years. The critical motivation is the possibility of reducing the human effort in scoring a large number of essays in a short period. In literature, most of the work concentrates on the English language; there is still a need for progress in Brazilian Portuguese. Thus, this works provides a Systematic Mapping Study aiming to identify Artificial Intelligence methods that support Automatic Essay Correction in Brazilian Portuguese. Furthermore, the main facts this paper brings are: (i) the methods focus on feature engineering methods instead of deep learning models; (ii) there is a prevalence of traditional metrics such as precision, coverage, and f-measure to evaluate the results; (iii) feedbacks provided by the tools have low-level of details; and (iv) there is no practical evaluation of the advancement in real-world applications.
Keywords: *Essay Scoring; Content Analysis; Natural Language processing*

Cite as: Lima, T. B., Silva, I. L. A., Freitas, E. L. S. X. Mello, R. F. (2023). Avaliação Automática de Redação: Uma revisão sistemática. Revista Brasileira de Informática na Educação – RBIE), 31, 205-221. DOI: 10.5753/rbie.2023.2869.

1 Introdução

No Brasil, milhões de estudantes participam anualmente do Exame Nacional do Ensino Médio (ENEM) como parte da sua trajetória acadêmica. Esse exame mede diferentes competências dos estudantes em áreas diversas, como português, matemática e ciências. Uma dessas competências é a de escrever uma redação no formato discursivo argumentativo, seguindo um tema proposto por um texto motivador. A redação deve seguir alguns critérios[1], de acordo com as seguintes competências: (i) aderência à escrita formal do português; (ii) escrita de acordo com o estilo argumentativo discursivo; (iii) a defesa de um ponto de vista; (iv) estrutura argumentativa; e (v) a elaboração de uma proposta de intervenção no problema debatido ao longo do texto (Marinho et al., 2021). Cada competência é pontuada entre 0 e 200, onde 0 é a pior nota e 200 a melhor.

No ano de 2022, o número de inscritos no ENEM atingiu a marca de 5.3 milhões de estudantes. A vasta quantidade de redações a serem avaliadas gera uma demanda por professores habilitados, ocasionando um custo excessivo (Mello et al., 2021). Por outro lado, a investigação de métodos de correção automática proporcionou um avanço significativo na forma como são tratados os processos de correção de redação. No entanto, essa área apresenta muitos desafios, pois é preciso garantir que a correção das redações esteja sendo realizada de forma precisa. Dessa forma, diversos métodos já foram propostos - desde aqueles que utilizam *deep learning* àqueles que são baseados na extração de atributos capazes de serem usados em classificadores como Máquina de Vetor de Suporte (MVS), Árvore de Decisão e XGBoost (Chen et al., 2015; Safavian e Landgrebe, 1991; Cortes e Vapnik, 1995; Fonseca et al., 2018; Ferreira Mello et al., 2022). Além disso, também costumam ser considerados desafiadores os aspectos relacionados às avaliações dos modelos, à detecção de possíveis vieses e à apresentação de *feedbacks* assertivos e compreensíveis.

A avaliação automática de redações é uma área que apresenta bastantes estudos aplicados à língua inglesa. Na língua portuguesa, essa é uma área que vem se expandindo ao longo dos últimos anos (Costa et al., 2020). Dessa forma, investigar a literatura permite esclarecer as principais questões relacionadas, métodos utilizados, desafios e propostas apresentadas. Neste contexto, este artigo propõe um Mapeamento Sistemático da Literatura a fim de investigar propostas, desafios e limitações da avaliação automática de redações em português. O processo de pesquisa foi dividido em três etapas: (i) Busca; (ii) Seleção e (iii) Extração. Ao final da aplicação das etapas planejadas, foram selecionados 6 artigos que foram analisados levando em consideração 6 questões de pesquisa, contribuindo assim para um entendimento mais aprofundado da área em questão.

2 Trabalhos Relacionados

Para tornar as avaliações das redações cada vez mais precisas, a comunidade científica tem trabalhado os diversos aspectos que envolvem essa questão. E, bem como neste trabalho, outros artigos buscaram revisar a literatura a fim de entender o estado da arte no que se refere à avaliação automática de redações. Desse modo, esta seção apresenta um conjunto de revisões da literatura já realizadas sobre o tema e discute quais as principais diferenças para a proposta desta pesquisa.

[1] A consulta detalhada aos critérios está disponível em: http://portal.mec.gov.br/ultimas-noticias/418-enem-946573306/81381-conheca-as-cinco-competencias-cobradas-na-redacao-do-enem

Em Nau et al. (2019) é realizada uma Revisão Sistemática da Literatura que busca analisar o estado da arte na área de avaliação automática de redações. Esse trabalho realizou buscas de artigos de acordo com um conjunto de palavras-chave em três diferentes repositórios: ACL, Scopus e Science Direct. Os critérios de seleção consideram tanto título como resumo e palavras-chave, organizadas de acordo com a seguinte expressão:

essay AND (scoring OR identifying) AND ("discourse element"OR "discourse analysis"OR "discourse structure"OR "conclusion statements") AND (nlp OR "natural language processing")

No referido trabalho, artigos que propõem algum tipo de intervenção também foram considerados. Foram incluídos artigos escritos na língua inglesa ou portuguesa que foram publicados entre 01/2012 e 12/2017. Após as buscas, foram retornados 87 trabalhos, sendo selecionados 5 trabalhos para a avaliação final. Esses trabalhos foram avaliados levando em consideração as características dos *corpus* de redações utilizados que possuem variados temas, a proposta de avaliação e as técnicas utilizadas nessa proposta.

No trabalho de Costa et al. (2020) foi realizado um Mapeamento Sistemático da Literatura cujo objetivo é apresentar um panorama do estado da arte na avaliação de corretores automáticos para a língua portuguesa. Diferentemente do trabalho de Nau et al. (2019), este trabalho foca apenas em artigos que fazem a aplicação da avaliação automática de redações em textos escritos na língua portuguesa. Os critérios de inclusão se concentram em artigos que são relacionados à correção automática em língua portuguesa - seja explicitamente seja relacionado. Também são incluídos artigos de *surveys*. Os autores buscaram definir as principais estratégias utilizadas na identificação de elementos textuais, os aspectos linguísticos utilizados, as métricas e as bases de dados utilizadas nas soluções. As buscas foram realizadas nas fontes digitais Scopus e IEEE e foram retornados 787 artigos no total. Foram selecionados 6 artigos através de critérios de inclusão e exclusão, e foram inseridos 4 artigos de forma manual, totalizando 10 artigos ao final do processo - que vão desde o ano 2013 ao ano 2018. Os autores não informam quais artigos foram selecionados pelos critérios da pesquisa e quais foram inseridos manualmente.

O trabalho de Costa et al. (2020) apresenta um diferencial, pois foca em trabalhos que apresentam propostas para a avaliação de textos escritos em português. Além disso, o referido trabalho avalia aspectos textuais que incluem o domínio do uso da escrita formal, utilização de conhecimento de diversas áreas, saber selecionar e organizar as informações; construir a argumentação através de mecanismos linguísticos; e a elaboração de uma proposta de intervenção. No entanto, nenhum dos dois trabalhos avalia a existência de validação das propostas em ambiente real, além de não explorar outras fontes digitais como a *Engineering Village* e a *Web of Science* na etapa de busca dos artigos, as quais foram consideradas na presente pesquisa por meio do protocolo de mapeamento sistemático apresentado na seção 3.

3 Metodologia

Um Mapeamento Sistemático da Literatura (MSL) deve realizar uma avaliação crítica das pesquisas que abordam um determinado assunto e deve ter uma estrutura bem definida para que os resultados não sejam enviesados. Além disto, o rigor de um mapeamento da literatura precisa ser reforçado, reduzindo os efeitos aleatórios e garantindo a reprodutibilidade (Becheikh et al.,

2006). De acordo com Kitchenham e Charters (2007), os critérios de seleção podem ser aplicados liberalmente, considerando a avaliação da conclusão, o que foi realizado em uma etapa posterior. O mesmo ocorreu na revisão sistemática realizada por Nunes et al. (2022), que segue as diretrizes estabelecidas por Moher et al. (2009).

O MSL proposto neste artigo seguiu as diretrizes e o modelo de protocolo de mapeamento sistemático proposto por Kitchenham e Charters (2007), e incluiu três etapas principais:

1. **Etapa de planejamento**: Nessa etapa foram definidos os objetivos do mapeamento e o protocolo que foi seguido, bem como, as questões de pesquisas;
2. **Etapa de Execução**: os artigos foram buscados e selecionados e, por último, foi realizada a extração das informações e síntese dos resultados;
3. **Etapa de Relatório**: Refere-se à apresentação e discussão dos resultados.

A seção seguinte apresenta o protocolo definido e aplicado no MSL realizado neste trabalho.

3.1 Questões de Pesquisa

A avaliação automática de redações em língua portuguesa está relacionada a diversas questões como correção ortográfica, pontuação automática das redações em diferentes competências, entre outros aspectos. A seguinte questão de pesquisa foi, então, elaborada:

Questão de Pesquisa: Como a Inteligência Artificial (IA) tem sido utilizada para oferecer suporte à avaliação automática de redações?

Levando em consideração essa pergunta de pesquisa, o trabalho foi dividido nas seguintes subquestões:

Questão de Pesquisa 1 (Q1): *Quais são os principais objetivos da utilização de inteligência artificial na avaliação de redações?*

Questão de Pesquisa 2 (Q2): *Quais os principais algoritmos de inteligência artificial que são utilizados para a avaliação de redações?*

Questão de Pesquisa 3 (Q3): *Quais são as métricas mais utilizadas para validação?*

Questão de Pesquisa 4 (Q4): *Quais são os bancos de dados mais utilizados para validação?*

Questão de Pesquisa 5 (Q5): *Existe alguma evidência de que a inteligência artificial auxilia na avaliação de redações?*

Questão de Pesquisa 6 (Q6): *Quais os critérios utilizados na avaliação das redações?*

3.2 Estratégia de Busca

As palavras-chave deste trabalho foram definidas levando em consideração o idioma inglês e o português e quatro domínios: Educacional, Inteligência Artificial, Aplicabilidade e Idioma de Aplicação. As palavras-chave definidas e utilizadas são as seguintes:

- Educacional - redação (*essay*), tema (*prompt*), gramática (*grammar*);
- Inteligência Artificial - aprendizado de máquina (*machine learning*), *deep learning*, processamento de linguagem natural (*natural language processing*);
- Aplicabilidade - avaliação (*evaluation/assessment*), pontuação (*scoring*), correção (*correction*), classificação (*grading*);
- Idioma de Aplicação - português (*portuguese*).

A *string* de busca foi construída com o auxílio dos operadores lógicos OR e AND, sendo o operador OR utilizado entre as palavras-chave do mesmo domínio e o operador AND entre os diferentes domínios. A *string* de busca foi utilizada nos dois idiomas (Inglês e Português) - e abaixo tem-se um exemplo da *string* de busca final no idioma português.

(“redação” OR “tema” OR “grámatica”)
AND
(“aprendizado de máquina” OR “deep learning” OR “processamento de linguagem natural”)
AND
(“avaliação” OR “pontuação” OR "correção"OR "classificação")
AND
(“português”)

As *strings* de busca nos dois idiomas foram aplicadas nas seguintes bases de dados de artigos científicos: ACM[2], IEEExplore[3], Engineering Village[4], Science Direct[5], SpringerLink[6], Scopus[7], Web of Science[8], SBC OpenLib[9]. As bases de dados escolhidas são amplamente usadas no processo de revisão bibliográfica, sendo algumas delas já utilizadas em trabalhos anteriores, como Nau et al. (2019) que utilizaram a Science Direct e Scopus. Além disso, a base SBC-OpenLib provê uma ampla variedade de artigos em português, sendo ideal para buscar artigos relacionados à área de educação no idioma. Ademais, as bases ACM e Springer retornaram resultados relevantes para a nosso mapeamento, como os artigos Mello et al. (2021) e Ferreira Mello et al. (2022). Por fim, a base de artigos IEEE também foi utilizada por Costa et al. (2020).

3.3 Processo de Seleção e Extração

Foram definidos alguns critérios de inclusão e exclusão para selecionar os artigos que fariam parte do Mapeamento Sistemático da Literatura. O Quadro 1 apresenta os critérios de seleção utilizados neste trabalho.

[2] https://dl.acm.org/
[3] https://ieeexplore.ieee.org/
[4] https://www.engineeringvillage.com/
[5] https://www.sciencedirect.com
[6] https://link.springer.com/
[7] https://www.scopus.com/
[8] https://www.webofscience.com
[9] https://sol.sbc.org.br/

Nº	Tipo	Descrição
1	Inclusão	Estudos primários
2	Inclusão	Estudos que propõem abordagens de inteligência artificial na avaliação de redações
3	Inclusão	Estudos que analisam redações em português
4	Exclusão	Estudos secundários ou terciários
5	Exclusão	Estudos duplicados ou re-indexados
6	Exclusão	Artigos escritos em idioma diferente do português/inglês
7	Exclusão	Artigos publicados em literatura cinza
8	Exclusão	Estudos incompletos
9	Exclusão	Veículos de publicação diferentes de conferência ou *journals*

Quadro 1: Critérios de Seleção. Fonte: os autores (2022).

A etapa de seleção manual se iniciou com a leitura dos títulos e dos resumos de todos os artigos retornados da etapa de busca, com o objetivo de avaliar os artigos de maneira geral quanto à importância da aplicação para o mapeamento realizado. Os artigos que obedeceram os critérios de inclusão e os artigos que não apresentaram informações suficientes para exclusão passaram para a próxima etapa do processo de seleção. Nessa próxima etapa, os autores realizaram a leitura da introdução e das considerações finais dos artigos, com o objetivo de incluir ou excluir os artigos com base nos critérios de seleção.

Na etapa da extração, os autores leram os textos completos dos artigos retornados da etapa de seleção com o objetivo de extrair dados relevantes para responder às perguntas de pesquisa. O Quadro 2 apresenta todas as categorias dos dados extraídos dos artigos.

#	Tipo	Descrição
1	ID	Identificador único do artigo
2	Título	Título do artigo
3	Autores	Autores do artigo
4	Ano	Ano de publicação do artigo
5	Países	País do primeiro autor do artigo
6	Tipo de publicação	Conferência ou *Journal*
7	Tipo de estudo	Experimental, Estudo de Caso, Aplicação
8	Ferramentas	Ferramentas utilizadas no estudo
9	Validação em ambiente real	Se a abordagem foi validada em um ambiente real
10	Banco de dados	Informações sobre o banco de dados utilizado
11	Principais resultados	Quais são os principais resultados do artigo?
12	Limitações/Trabalhos Futuros	Quais são as limitações apontadas pelos autores?
13	Principais objetivos (Q1)	Qual o principal objetivo do artigo?
14	Algoritmos (Q2)	Quais os principais algoritmos utilizados?
15	Métricas de validação (Q3)	Quais as métricas de validação utilizadas?
16	Banco de dados (Q4)	Quais os bancos de dados utilizados na validação?
17	Evidência de melhora (Q5)	Há evidência de impacto positivo ou negativo na aplicação de avaliação automática de redação?
18	Critérios (Q6)	Quais os critérios utilizados na avaliação?

Quadro 2: Categorias dos dados extraídos. Fonte: os autores (2022).

4 Execução e Relatório

A primeira etapa da execução é a etapa de busca, onde a *string* de busca de cada idioma foi aplicada nas bases de dados de artigos científicos e em seguida foi realizado o *download* das referências dos artigos retornados. A busca final dos artigos foi realizada no mês de maio de 2022 e a Tabela 1 apresenta a quantidade de artigos retornados em cada uma das bases. Foram considerados em nossa busca artigos no período de Janeiro/2012 até Março/2022.

Tabela 1: Número de artigos retornados pela *string* de busca em cada base científica. Os artigos estão divididos em relação ao idioma a qual o artigo está escrito. O símbolo '-' significa que nenhum artigo foi encontrado. Fonte: os autores 2022.

Base Científica	**Número de Artigos**		
	Inglês	*Português*	Total
ACM	282	33	315
IEEExplore	6	-	6
Engineering Village	25	-	25
Science Direct	321	-	321
SpringerLink	494	-	494
Scopus	8	4	12
Web of Science	11	-	11
SBC-OpenLib	2	2	4
Total	1149	39	1188

A próxima etapa do processo de execução é a seleção. Nessa etapa foi utilizada a ferramenta Rayyan[10], dos autores Johnson e Phillips (2018), que oferece apoio ao desenvolvimento de revisões sistemáticas. A ferramenta Rayyan foi utilizada em trabalhos anteriores de revisão sistemática da literatura como o de Papadopoulos et al. (2020) e de Nunes et al. (2022). Foi também avaliada por diferentes trabalhos em relação à condução de revisões sintemáticas da literatura apresentando um desempenho similar a outras plataformas com o propósito análogo, sendo uma das mais precisas na detecção de artigos duplicados (Guimarães et al., 2022; McKeown e Mir, 2021).

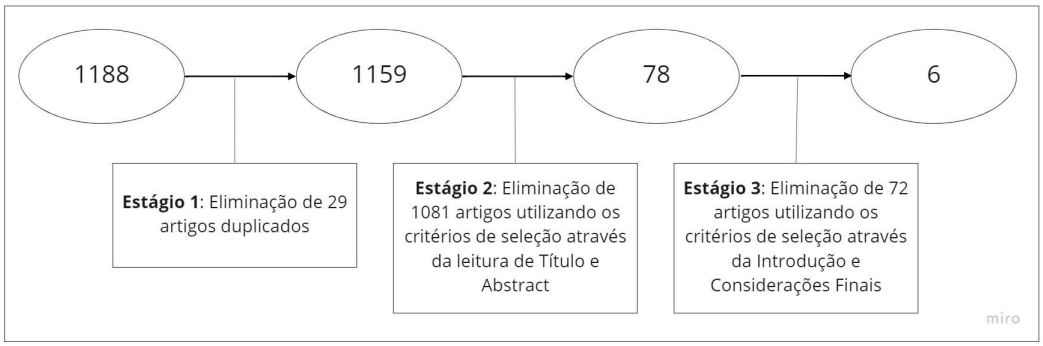

Figura 1: Fases de Seleção. Fonte: os autores (2022).

No nosso trabalho, a ferramenta auxiliou no processo de seleção em três estágios: (1) Remoção de artigos duplicados, visto que a ferramenta detecta artigos possivelmente duplicados e,

[10]https://www.rayyan.ai/

em seguida, é possível confirmar ou retificar essa informação de forma manual; (2) Identificação dos critérios de seleção através da leitura dos resumos e dos títulos; e (3) Aplicação dos critérios de seleção através da leitura da Introdução e das Considerações Finais. A Figura 4 apresenta o número de artigos selecionados em cada um desses estágios.

Dessa forma, no **Estágio 1**, 29 artigos foram detectados como duplicados pela ferramenta e conferidos e removidos da revisão pelos autores. No **Estágio 2**, os autores aplicaram os critérios de inclusão e exclusão a fim de selecionar apenas artigos de estudos primários de conferências ou revistas e que tratavam diretamente de Inteligência Artificial aplicada ao contexto de correção de redação.

Nesse estágio foram removidos cerca de 93% dos artigos e 78 artigos passaram para o próximo estágio. Os artigos que não continham informações suficientes nos resumos e nos títulos para a seleção através dos critérios de inclusão e exclusão passaram para o Estágio 3. No **Estágio 3**, os autores revisaram os artigos através da leitura da Introdução e das Considerações Finais. Nesse estágio foram selecionados 6 artigos que atenderam a todos os critérios previamente definidos para serem incluídos de forma definitiva e participarem da pesquisa.

Os trabalhos selecionados foram publicados entre os anos de 2018 e 2022. O Brasil aparece como País de publicação em 5 dos artigos selecionados, sendo que em dois deles a pesquisa foi realizada em parceria com outros países (Reino Unido, Austrália e Arábia Saudita). O país de publicação do sexto artigo é Portugal. Os 6 artigos selecionados foram publicados em conferências. O Quadro 3 apresenta uma síntese dos principais resultados encontrados relacionados às perguntas de pesquisa estudadas nesse artigo.

4.1 Principais Objetivos (Q1)

No que se refere à primeira pergunta **"Quais são os principais objetivos da utilização de inteligência artificial na avaliação de redações?"**, tais objetivos estão ligados a diferentes fatores entre eles: (i) Algoritmos de IA são mais eficientes que avaliadores humanos; (ii) não são influenciados por questões subjetivas, evitando inconsistências; e (iii) criação de *feedbacks* importantes para educadores e a geração de dados analíticos referentes ao desempenho dos alunos em sala de aula na atividade de produção textual. A partir desses objetivos, outros objetivos mais específicos são traçados, como a avaliação de diferentes abordagens, o foco em diferentes aspectos de avaliação e a melhoria nas performances.

Um dos principais objetivos identificados neste mapeamento foi a avaliação da estrutura argumentativa e a coerência de uma redação durante a defesa de um ponto de vista relacionado ao tema proposto, aparecendo como objetivo específico em 4 dos 6 trabalhos relacionados: Ferreira Mello et al. (2022), Sousa et al. (2021), Filho et al. (2018), Mello et al. (2021). Quanto aos artigos Mello et al. (2021) e Filho et al. (2019) , apesar de similares, o artigo de Ferreira Mello et al. (2022) é mais abrangente e aprofundado no assunto, explorando diversos algoritmos e *datasets* sendo, portanto, uma extensão do artigo anterior de Mello et al. (2021). Já o trabalho de Filho et al. (2019) busca resolver o problema do desbalanceamento entre classes na avaliação de uma redação quanto ao domínio da escrita formal da língua portuguesa. Por último, o trabalho de Fonseca et al. (2018) faz uma comparação entre o uso de redes neurais profundas e um sistema baseado em engenharia de atributos na avaliação de uma redação levando em consideração aspectos gerais.

Nesse sentido, pode-se concluir que os trabalhos atuais da literatura buscam por uma forma determinística e eficiente de gerar pontuações relacionadas às redações do ENEM. Sendo assim, os trabalhos propõem objetivos gerais importantes para o uso da IA no contexto de correção de redação, embora tais objetivos não sejam devidamente assegurados (como a questão de performance entre classificadores e o viés na correção de redação), tais objetivos podem servir de norte para trabalhos futuros na área.

4.2 Principais Métodos (Q2)

A segunda pergunta de pesquisa **"Quais os principais algoritmos de inteligência artificial que são utilizados para a avaliação de redações"** se concentra em desvendar os métodos utilizados na literatura. Os principais métodos de IA avaliados em português são: (i) baseados na extração de atributos, seguidos da classificação através de algoritmos de Aprendizado de Máquina (AM) supervisionado; (ii) baseados em *deep learning* em uma abordagem de classificação supervisionada.

Na primeira categoria, (i), podemos incluir os trabalhos de Ferreira Mello et al. (2022) e Mello et al. (2021), onde são explorados métodos independentes do conteúdo, que são obtidos através de ferramentas como *Coh-Metrix* e *Linguistic Inquiry Word Count (LIWC)*, e métodos dependentes do conteúdo, como Frequência do Termo e Frequência Inversa do Documento (TF-IDF, na sigla em inglês). O *Coh-Metrix* é uma ferramenta que permite a extração de diferentes atributos relacionados a diversos aspectos linguísticos como legibilidade, coesão, entre outros. Atualmente o *Coh-Metrix* possui 48 métricas e recebeu uma versão *web* desenvolvida por Camelo et al. (2020). Já a LIWC é uma ferramenta utilizada para a detecção de sentimentos no texto, que é importante nesse tipo de trabalho, pois com ela é possível detectar expressões verbais que podem conter valor semântico significativo no texto Kahn et al. (2007). Os algoritmos como *Support Vector Machine* (SVM), Árvore de Decisão e Adaboost obtiveram melhores resultados utilizando TF-IDF, ao passo que para os atributos independentes de conteúdo os resultados foram melhores usando XGBoost e *Conditional Random Fields* (CRF).

A abordagem se repete para o trabalho de Filho et al. (2018), que utiliza o classificador e o regressor do algortimo SVM para a avaliação de aderência ao tema em um texto dissertativo-argumentativo. O processo utilizado é semelhante ao de Ferreira Mello et al. (2022), onde atributos relacionados à contagem e repetições de palavras, além de outras métricas relacionadas ao domínio argumentativo, são utilizadas, totalizando 89 métricas. Apesar do estudo adotar um processo similar ao de Ferreira Mello et al. (2022), os resultados não foram satisfatórios em completude pelo fato do algoritmo falhar ao classificar pontuações intermediárias. Uma das possíveis razões talvez seja a tentativa falha de balancear as classes da base de dados com o método de *oversampling* Synthetic Minority Oversampling TEchnique (SMOTE). O trabalho de Filho et al. (2019) explora mais a fundo o problema do desbalanceamento de classes na área da avaliação automática de redações. As técnicas de balanceamento utilizadas são o SMOTE, o Adaptive Synthetic (ADASYN), o *Random Oversampling* e o *Random Undersampling*. Os algoritmos de regressão utilizados na avaliação são o *Least Absolute Shrinkage and Selection Operator (LASSO)* e o regressor SVM e os algoritmos de classificação utilizados foram o *Gradient Boosted Trees* e o classificador SVM. Os resultados do estudo apontam que as técnicas SMOTE e ADASYN são menos efetivas quando comparadas às outras técnicas de abordagem aleatória.

Na segunda categoria, (ii), temos o trabalho de Sousa et al. (2021) que considera o uso de modelo pré-treinado baseado em modelos recentes como *Transformer* e CRF. Inicialmente, esse trabalho propõe uma atividade de mineração de argumentos utilizando para isso o conjunto de dados de *Persuasive Essay corpus* contendo uma série de textos com diferentes tópicos. Após a realização de um processo de tradução dos textos, os dados foram treinados em uma abordagem de *token-level tagging* usando bidirectional *Long Short Term Memory* (LSTM) e *multilingual Bidirectional Encoder Representation Transformer* (mBERT). Nos experimentos de classificação realizados por Sousa et al. (2021), utilizando etiquetagem automática, o modelo mBERT obteve os melhores resultados, com 70.12 f1-macro para a versão traduzida em português em comparação com o algoritmo BLSTMCRF+Char (character), que obteve 68.59 f1-macro.

Por último, o trabalho de Fonseca et al. (2018) explora as duas categorias ao comparar uma abordagem baseada em *deep learning*, que utiliza camadas de LSTM bidirecionais e vetores de palavras treinados com o Global Vectors for Word Representation (GLOVE), e uma abordagem baseada em engenharia de atributos. Os algoritmos utilizados na abordagem baseada em engenharia de atributos foram o Gradient Boosting e a Regressão Linear. Nesse estudo, o Gradient Boosting apresentou o melhor resultado na avaliação das competências 1, 2, 3 e 4, enquanto a rede neural apresentou melhor performance para a competência 5.

Sendo assim, é possível concluir que a literatura explora algoritmos baseados na extração de características pré-definidas, sejam elas baseadas no conteúdo ou independentes do conteúdo, bem como, algoritmos baseados em *deep learning*. Além disso, algoritmos de AM são explorados de duas maneiras principais, sejam por meio de regressão ou classificação, com destaque para os algoritmos **Gradient Boost** e **XGBoost**.

4.3 Métricas de Avaliação (Q3)

Na terceira questão de pesquisa **"Quais as métricas utilizadas?"**, entre os estudos selecionados neste mapeamento sistemático da literatura grande parte empregou técnicas de validação dos ensaios combinando diferentes métricas. Estas foram: Precisão, Cobertura e *F-Measure*, que são métricas amplamente utilizadas no campo da aprendizagem de máquina e foram utilizadas nos trabalhos de Ferreira Mello et al. (2022), Sousa et al. (2021) e Mello et al. (2021) para avaliar os algoritmos em relação à verificação da estrutura argumentativa e a coerência das redações. Nos trabalhos de Ferreira Mello et al. (2022) e Mello et al. (2021), além das métricas mencionadas, também foi utilizado o *Kappa* de *Coehn* que é uma métrica bastante utilizada na área de mineração de dados educacionais.

Já os trabalhos de Filho et al. (2018) e de Filho et al. (2019) utilizaram matriz de confusão para fazer validações relacionadas à aderência ao tema e ao domínio formal da língua portuguesa, respectivamente. A matriz de confusão facilita a visualização dos erros produzidos pelos modelos em termos de verdadeiro positivo e falso negativo. Além da matriz de confusão, o estudo de Filho et al. (2019) também utilizou as métricas de Precisão e Cobertura e a Correlação de Pearson nas suas validações. Por último, o estudo de Fonseca et al. (2018) utilizou as métricas de Erro médio quadrático (RMSE) e o *Quadratic Weight Kappa* (QWK) para avaliar as redações de maneira geral. Essas são métricas que costumam ser utilizadas na literatura relacionada à avaliação automática de redações.

4.4 Banco de Dados (Q4)

Tabela 2: A tabela mostra o número de redações em cada um das base de dados citadas..

ID	Fonte do Banco de Dados	Número de Redações
1	Santos et al. (2018)	271
2	Haendchen Filho et al. (2018)	50
3	Sousa et al. (2021)	402
4	Redações da UOL e do Brasil Escola	1983
5	Fonseca et al. (2018)	56.644

A Tabela 2 mostra todos os conjuntos de bases de dados utilizados nos trabalhos presentes na revisão relacionados a questão de pesquisa **"Quais os bancos de dados mais utilizados para validação?"**. Na base de dados de Santos et al. (2018) e Nau et al. (2019) encontramos 271 e 50 redações, respectivamente, utilizadas em ambos trabalhos de Ferreira Mello et al. (2022) e Mello et al. (2021), que utilizou extração de características para análise de estrutura retórica. A segunda ainda foi utilizada no trabalho de Mello et al. (2021). A base de dados 3 foi utilizada no trabalho Sousa et al. (2021) para análise de estrutura argumentativa e persuasiva. A base de dados era originalmente escrita em inglês, mas foi traduzida para Português a fim de realizar-se as análises pretendidas em relação à estrutura argumentativa do texto. As bases 4 e 5 são redações no formato do ENEM. No caso da base 4, os dados foram extraídos de redações disponíveis na internet através de *web crawling*, enquanto a base de dados 5 foi obtida através da escrita de redações avaliadas por profissionais de educação em uma plataforma e pontuadas de acordo com as competências do ENEM (Fonseca et al., 2018).

4.5 Evidência de Melhora (Q5)

A quinta questão de pesquisa, **"Existe alguma evidência de que a aprendizagem de máquina auxilia na avaliação de redações?"**, proposta neste mapeamento, tinha como objetivo responder se há evidência de que o uso de estratégias de IA trazem melhorias ao campo da avaliação automática de redações. Apesar dos resultados promissores apresentados por todos os estudos incluídos, o que se pôde perceber foi uma lacuna no que tange à validação em ambientes reais, já que nenhum dos artigos selecionados apresentou esse tipo de validação. Inúmeros fatores podem justificar essa falta de comunicação entre a comunidade científica e o mundo real, como a falta de investimentos, falta de padronização e/ou estruturação nas bases disponíveis e até a própria magnitude de possibilidades de aplicação de estratégias de IA, que pode acabar motivando os cientistas a experimentar novas variações e/ou abordagens em contextos distintos em vez do aprofundamento de uma pesquisa que já obteve resultados preliminares.

Não obstante, o trabalho realizado por Nunes et al. (2022) investigou a existência de trabalhos que avaliam empiricamente a efetividade dos algoritmos de AES. Afinal, apenas 8 artigos foram selecionados considerando um espaço temporal de 20 anos (2000-2020), apontando apenas dois trabalhos fora dos Estados Unidos da América (EUA) e com a maioria deles avaliando estudantes que estão entre a educação primária e secundária. O período das avaliações realizadas nos trabalhos varia desde alguns dias a meses, obtendo-se avaliação positiva do uso da correção automática em 7 de 8 deles.

O trabalho Wilson e Roscoe (2020), por exemplo, avaliou alunos que tem o inglês como

língua nativa com idade média de aproximadamente 11 anos aplicando um pré e pós testes em uma análise quantitativa. Os estudantes foram separados em dois grupos, Experimental (E) e de Controle ativo (C), com 56 (64% meninas) e 58 (62% meninas) alunos cada um, respectivamente. O grupo experimental recebeu *feedback* do sistema de correção automática *Project Essay Grade* (PGE, sigla em inglês) em relação à gramática, desenvolvimento da ideia, organização, estrutura da sentença, escolha da palavra, convenções e estilo. Por fim, os alunos poderia receber *feedbacks* adicionais dos professores através de um *chat* na plataforma. Os resultados foram comparados com alunos que usaram o Google Docs e receberam o *feedback* dos professores sem utilizarem um sistema automático. Os resultados mostraram que apesar da inexistência de diferença significativa em relação à nota final de ambos os grupos no pós-teste, os alunos que usaram o sistema de correção automática se tornaram mais confiantes. Por fim, 3 educadores também tiveram uma impressão positiva do sistema em relação à usabilidade, e qualidade e quantidade dos *feedbacks* dados aos alunos pela plataforma.

Assim sendo, há uma percepção positiva tanto para alunos, através da melhora do processo de escrita ao gerar maior confiança, quanto para professores ao permitir focá-los apenas em ensinar e diminuindo o esforço em corrigir os textos. Apesar disso, os professores perceberam que estudantes com mais dificuldades em relação à escrita podem não ser capazes de utilizar as ferramentas pelo fato de produzirem textos curtos demais e com uma quantidade de erros excessiva (Palermo e Thomson, 2018; Tang e Rich, 2017; Ware, 2014; Nunes et al., 2022). Portanto, no geral, o uso de um sistema de AES apresenta benefícios significativos quando aplicados no mundo real, embora hajam poucos artigos que avaliam tal aspecto. Alguns pontos são destacados em relação a esses trabalhos, como a não consideração do processo de integração e aprendizado da plataforma como sendo um fator a influenciar, bem como, a ausência de um processo rigoroso para detectar explicações mais diversificadas.

4.6 Critérios Avaliados (Q6)

Na última questão de pesquisa, **"Quais os critérios utilizados na avaliação das redações?"**, buscou-se avaliar quais são os aspectos linguísticos que são utilizados nos trabalhos relacionados à avaliação automática de redações.

Nos artigos de Filho et al. (2018), Fonseca et al. (2018) e Filho et al. (2019) foram utilizados aspectos relacionados às competências exigidas no Exame Nacional do Ensino Médio (ENEM). No estudo de Fonseca et al. (2018) foram utilizadas as 5 competências do ENEM. No estudo de Filho et al. (2018) foi utilizada a competência 2, que investiga se a redação escrita está relacionada ao tema proposto. Já no estudo de Filho et al. (2019) foi explorada a competência 1, que está relacionada ao domínio da escrita formal da língua portuguesa.

Os estudos de Ferreira Mello et al. (2022), Sousa et al. (2021) e Mello et al. (2021) exploram aspectos linguísticos que são exigidos no ENEM, mais especificamente as competência 3 e 4. O estudo de Sousa et al. (2021) explora o aspecto relacionado à argumentação, que pode ser relacionado às competências 3 e 4 do ENEM, onde se avalia a organização de fatos e opiniões em defesa de um ponto de vista e a demonstração do conhecimento linguístico necessário para construir a argumentação. Os estudos de Ferreira Mello et al. (2022) e Mello et al. (2021) expandiram a avaliação para aspectos para além da coesão, coerência, incluindo também legibilidade e relações semânticas presentes nos textos, que são aspectos esperados em uma redação desenvolvida para

esse exame.

5 Considerações Finais

Este artigo apresenta uma visão geral dos estudos relacionados à correção automática de redação que foram coletados por meio de um Mapeamento Sistemático da Literatura. Fazendo um pequeno recorte sobre as abordagens de inteligência artificial que são utilizadas para realizar a avaliação automática de redações, conclui-se que o principal objetivo da utilização da inteligência artificial nessa área é a busca por uma forma determinística e eficiente de gerar pontuações associadas a redações escritas por alunos que buscam ingressar no ensino superior.

Um dos aspectos identificados nos trabalhos selecionados é o baixo uso de atributos extraídos de forma automatizada através de técnicas de *deep learning*. Apenas dois trabalhos exploram essa abordagem, apesar de ser uma estratégia bastante utilizada em outras áreas de Processamento de Linguagem Natural (PLN). Levando em consideração as abordagens utilizadas na validação das estratégias propostas, foi identificado que os trabalhos nem sempre utilizam análises estatísticas, como testes de hipótese, para assegurar a eficiência de um modelo quando comparado com outro. Outro fato identificado é que não há uma avaliação de diferentes abordagens em relação à performance computacional, embora esse aspecto seja citado como uma das motivações por trás do uso de algoritmos de IA.

Outra lacuna identificada entre os trabalhos selecionados é o baixo detalhamento nos *feedbacks* retornados pelos modelos de avaliação. A geração de *feedbacks* a partir das análises dos algoritmos é algo de extrema importância quando leva-se em consideração o contexto educacional e o propósito didático que algumas abordagens podem assumir em trabalhos futuros. Além disso, não foram encontradas investigações relacionadas ao viés que pode ser introduzido na correção automática de redações, ainda que esse seja um problema encontrado em outras tarefas ligadas ao PLN.

Por último, foi identificado que nenhum dos trabalhos fez a validação dos seus resultados em um ambiente real e nenhum trabalho utilizou a abordagem proposta no desenvolvimento de uma plataforma que realizasse a avaliação automática de redações. Portanto, os objetivos traçados pelos artigos selecionados acabam por não serem totalmente atingidos ou verificados em sua completude. Assim sendo, existem amplas oportunidades que podem ser exploradas em trabalhos futuros nessa área.

Referências

Becheikh, N., Landry, R., & Amara, N. (2006). Lessons from innovation empirical studies in the manufacturing sector: A systematic review of the literature from 1993–2003. *Technovation*, *26*(5-6), 644–664. https://doi.org/10.1016/j.technovation.2005.06.016. [GS Search]

Camelo, R., Justino, S., & Mello, R. (2020). Coh-Metrix PT-BR: Uma API web de análise textual para a educação. *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*, 179–186. https://doi.org/10.5753/cbie.wcbie.2020.179. [GS Search]

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1–4. https://doi.org/10.1145/2939672.2939785. [GS Search]

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018. [GS Search]

Costa, L., Oliveira, E. H. T. d., & Castro Júnior, A. (2020). Corretor Automático de Redações em Língua Portuguesa: um mapeamento sistemático de literatura. *Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE 2020)*, 1403–1412. https://doi.org/10.5753/cbie.sbie.2020.1403. [GS Search]

Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., & Gasevic, D. (2022). Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese. *LAK22: 12th International Learning Analytics and Knowledge Conference*, 404–414. https://doi.org/10.1145/3506860.3506977. [GS Search]

Filho, A. H., Concatto, F., Nau, J., Prado, H. A. d., Imhof, D. O., & Ferneda, E. (2019). Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. *Procedia Computer Science*, *159*, 764–773. https://doi.org/10.1016/j.procs.2019.09.235. [GS Search]

Filho, A. H., do Prado, H. A., Ferneda, E., & Nau, J. (2018). An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science*, *126*, 788–797. https://doi.org/10.1016/j.procs.2018.08.013. [GS Search]

Fonseca, E., Medeiros, I., Kamikawachi, D., & Bokan, A. (2018). Automatically Grading Brazilian Student Essays [Series Title: Lecture Notes in Computer Science]. Em A. Villavicencio, V. Moreira, A. Abad, H. Caseli, P. Gamallo, C. Ramisch, H. Gonçalo Oliveira & G. H. Paetzold (Ed.), *Computational Processing of the Portuguese Language* (pp. 170–179). Springer International Publishing. https://doi.org/10.1007/978-3-319-99722-3_18. [GS Search]

Guimarães, N. S., Ferreira, A. J., Silva, R. d. C. R., de Paula, A. A., Lisboa, C. S., Magno, L., Ichiara, M. Y., & Barreto, M. L. (2022). Deduplicating records in systematic reviews: there are free, accurate automated ways to do so. *Journal of Clinical Epidemiology*, *152*, 110–115. https://doi.org/j.jclinepi.2022.10.009. [GS Search]

Haendchen Filho, A., do Prado, H. A., Ferneda, E., & Nau, J. (2018). An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science*, *126*, 788–797. https://doi.org/10.1016/j.procs.2018.08.013. [GS Search]

Johnson, N., & Phillips, M. (2018). Rayyan for systematic reviews. *Journal of Electronic Resources Librarianship*, *30*(1), 46–48. https://doi.org/10.1080/1941126X.2018.1444339. [GS Search]

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, *120*(2), 263–286. [GS Search].

Kitchenham, B. A., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering* (rel. técn. EBSE 2007-001). Keele University e Durham University Joint Report. [GS Search].

Marinho, J., Anchiêta, R., & Moura, R. (2021). Essay-BR: a Brazilian Corpus of Essays. *Anais do III Dataset Showcase Workshop*, 53–64. https://doi.org/10.5753/dsw.2021.17414. [GS Search]

McKeown, S., & Mir, Z. M. (2021). Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. *Systematic reviews*, *10*, 1–8. https://doi.org/10.1186/s13643-021-01583-y. [GS Search]

Mello, R. F., Fiorentino, G., Miranda, P., Oliveira, H., Raković, M., & Gašević, D. (2021). Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays [Series Title: Lecture Notes in Computer Science]. Em I. Roll, D. McNamara, S. Sosnovsky, R. Luckin & V. Dimitrova (Ed.), *Artificial Intelligence in Education* (pp. 162–167). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2_29. [GS Search]

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*, t. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135. [GS Search]

Nau, J., Haendchen Filho, A., & Dazzi, R. L. S. (2019). Identificação e Avaliação Automática da Proposta de Intervenção em Textos Dissertativos-Argumentativos: Uma Revisão Sistemática da Literatura. *Anais do Computer on the Beach*, 493–501. https://doi.org/10.4013/cld.2017.153.08. [GS Search]

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, *38*(2), 599–620. https://doi.org/10.1111/jcal.12635. [GS Search]

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, *54*, 255–270. https://doi.org/10.1016/j.cedpsych.2018.07.002. [GS Search]

Papadopoulos, I., Koulouglioti, C., Lazzarino, R., & Ali, S. (2020). Enablers and barriers to the implementation of socially assistive humanoid robots in health and social care: a systematic review. *BMJ open*, *10*(1), e033096. https://doi.org/10.1136/bmjopen-2019-033096. [GS Search]

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660–674. https://doi.org/10.1109/21.97458. [GS Search]

Santos, K. S., Soder, M., Marques, B. S. B., & Feltrim, V. D. (2018). Analyzing the rhetorical structure of opinion articles in the context of a Brazilian college entrance examination. *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, 3–12. https://doi.org/10.1007/978-3-319-99722-3_1. [GS Search]

Sousa, A., Leite, B., Rocha, G., & Lopes Cardoso, H. (2021). Cross-Lingual Annotation Projection for Argument Mining in Portuguese [Series Title: Lecture Notes in Computer Science]. Em G. Marreiros, F. S. Melo, N. Lau, H. Lopes Cardoso & L. P. Reis (Ed.), *Progress in Artificial Intelligence* (pp. 752–765). Springer International Publishing. https://doi.org/10.1007/978-3-030-86230-5_59. [GS Search]

Tang, J., & Rich, C. S. (2017). Automated writing evaluation in an EFL setting: Lessons from China. *JALT CALL Journal*, *13*(2), 117–146. https://doi.org/10.29140/jaltcall.v13n2.215. [GS Search]

Ware, P. (2014). Feedback for Adolescent Writers in the English Classroom. *Writing & Pedagogy*, *6*(2). https://doi.org/10.1558/wap.v6i2.223. [GS Search]

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87–125. https://doi.org/10.1177/0735633119830764. [GS Search]

Artigo	Objetivos(Q1)	Algoritmos(Q2)	Validação(Q3)	Banco de Dados(Q4)	Evidência (Q5)	Critérios *(Q6)
Ferreira Mello et al. (2022)	Estrutura retórica	SVM, Random Forest, Adaboost, XGBoost e CRF	Kappa, Cobertura, Precisão e F-Measure	Banco de dados de Santos et al. (2018) e Haendchen Filho et al. (2018)	Não	Competências 3 e 4
Sousa et al. (2021)	Análise de argumentação	BLSTM e CNN	Cobertura, Precisão e F-Measure	Banco próprio	Não	Competência 3 e 4
Filho et al. (2018)	Fuga ao tema	R-SVM e C-SVM	Precisão e correlação	Redações da UOL e do Brasil Escola	Não	Competência 2
Fonseca et al. (2018)	Avaliação das competências do ENEM	BLSTM e CNN	QWK e RMSE	Banco próprio	Não	Todas as competências
Filho et al. (2019)	Domínio formal do português	SVM e GBT	Relação de verdadeiro positivo	Banco da UOL	Não	Competência 1
Mello et al. (2021)	Estrutura retórica	SVM, Random Forest, Adaboost, XGBoost e CRF	Kappa, Cobertura, Precisão e F-Measure	Santos et al. (2018).	Não	Competências 3 e 4

Quadro 3: Resumo das cinco perguntas de pesquisa.* As competências citas são do ENEM. Fonte: os autores (2022).