Estimando Coesão Textual em Redações no Contexto do ENEM Utilizando Modelos de Aprendizado de Máquina

Hilário Oliveira[1], Péricles Miranda[2], Seiji Isotani[4,5], Jário Santo[4], Thiago Cordeiro[3], Ig Ibert Bittencourt[3,5], Rafael Ferreira Mello[2]

[1] Instituto Federal do Espírito Santo - Campus Serra

[2]Universidade Federal Rural de Pernambuco

[3]Universidade Federal de Alagoas

[4]Universidade de São Paulo

[5]Harvard Graduate School of Education

hilario.oliveira@ifes.edu.br,{pericles.miranda,rafael.mello}@ufrpe.br

thiago@ic.ufal.br,{seiji_isotani,ig_bittencourt}@gse.harvard.edu

Abstract. *Textual cohesion is a fundamental property of formal writing, as it relates to the harmonious connection of text elements. Although several works automatically analyze textual cohesion in essays, there are still few works for Portuguese. This work investigates the application of regression models to estimate the textual cohesion of essays written in Portuguese in the context of ENEM, adopting a set of 151 characteristics identified in the literature. Experiments using the Essay-BR corpus, composed of 4,570 ENEM-style essays, demonstrate that the Extremely Randomized Trees model achieved the best results with a moderate Pearson correlation (53.08%) related to cohesion grades.*

Resumo. *Coesão textual é uma propriedade fundamental da escrita formal, pois tem relação com a conexão harmoniosa dos elementos de um texto. Apesar de diversos trabalhos analisarem automaticamente a coesão textual em redações, ainda são escassos trabalhos para o português. Este trabalho investiga modelos de regressão para estimar a coesão textual de redações escritas em português no contexto do ENEM, adotando um conjunto de 151 características identificadas na literatura. Experimentos usando a base de dados do Essay-BR, composta por 4.570 redações no estilo do ENEM, demonstram que o modelo de Extremely Randomized Trees apresentou os melhores resultados com uma correlação de Pearson (53,08%) moderada com as notas relacionadas à coesão.*

1. Introdução

O domínio e uso formal da linguagem, seja ela escrita ou falada, é uma habilidade essencial para qualquer profissional. É comum, no âmbito profissional ou acadêmico, a necessidade de escrever textos que apresentem ideias, fatos e argumentos que os suportem de forma clara e objetiva [Kellogg and Raulerson 2007]. Diversas vagas de empregos e exames de admissão em universidades adotam provas escritas durante os seus processos seletivos. Por isso, estudantes são ensinados a realizarem tarefas de leitura, interpretação e produções textuais (redações) durante o período escolar [Costa et al. 2020].

DOI: 10.5753/sbie.2022.224736

No Brasil, o Exame Nacional do Ensino Médio (ENEM)[1] tem por objetivo avaliar os conhecimentos adquiridos pelos candidatos ao longo dos seus anos escolares. A partir de 2009, o ENEM passou a ser adotado como exame de admissão, sendo atualmente o principal mecanismo de entrada para a maioria das universidades e outras instituições de ensino superior no Brasil [Marinho et al. 2021]. Um dos principais componentes do ENEM é a prova de produção textual, em que o aluno deve escrever uma redação dissertativa-argumentativa sobre um determinado tópico de ordem científica, cultural, política ou social [Klein and Fontanive 2009].

Segundo a cartilha da prova de redação do ENEM[2], o candidato deve escrever um texto coeso e coerente, seguindo as regras de escrita formal da língua portuguesa do Brasil, defendendo uma tese apoiada com argumentos embasados em relação ao tema proposto. Nas avaliações são consideradas as cinco competências a seguir: (i) Domínio da escrita formal da língua portuguesa; (ii) Compreensão do tema proposto; (iii) Seleção e organização das informações; (iv) Demonstração dos conhecimentos dos mecanismos linguísticos necessários para a construção da argumentação; e (v) Elaboração de uma proposta de intervenção para o problema abordado que respeite os direitos humanos.

A Competência 4 é relacionada à coesão textual da redação e é considerada uma das que apresenta maior dificuldade para os candidatos [Klein and Fontanive 2009, Lima et al. 2018]. Essa competência avalia se o candidato consegue utilizar recursos coesivos para articular as partes do texto de maneira estruturada. Um dos possíveis motivos para essa dificuldade é que diferente da linguagem falada, na escrita formal é necessário usar um conjunto diverso de estruturas gramaticais para interligar os elementos do texto [Lima et al. 2018]. Neste contexto, é fundamental que os candidatos estejam preparados adequadamente para conseguirem alcançar boas notas na redação do ENEM.

Segundo [Antunes 2005], a coesão textual é uma propriedade fundamental de um texto bem escrito, tendo relação com o uso de mecanismos linguísticos que possibilitam a conexão entre elementos do texto como palavras, frases e parágrafos. Apesar da existência de diversos trabalhos que investigam técnicas para analisar automaticamente a coesão textual em redações, em sua maioria, eles abordam somente textos escritos em inglês, sendo ainda escassos trabalhos para o português do Brasil [Lima et al. 2018, Crossley et al. 2019].

Neste contexto, o objetivo deste trabalho é avaliar a aplicação de modelos baseados em algoritmos de Aprendizado de Máquina (AM) para estimar automaticamente a coesão textual de redações no estilo do ENEM. Para isso, 151 características foram identificadas na literatura e usadas para construir um modelo de regressão capaz de estimar a nota da Competência 4 do ENEM. As características adotadas buscam refletir diferentes aspectos relacionados à coesão textual e que são usados como critérios de avaliação na prova redação do ENEM. Experimentos foram realizados usando a base de dados do Essay-BR [Marinho et al. 2021] que é composta por 4.570 redações. Os experimentos executados avaliaram diferentes algoritmos de regressão com base nas medidas de erro quadrático médio e erro médio absoluto. Os resultados experimentais obtidos demonstram que o modelo de *Extremely Randomized Trees* apresentou os melhores resultados

[1]https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

[2]https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/outros-documentos

globais com uma correlação de Pearson (53,08%) moderada com as notas da coesão textual (Competência 4) e um erro absoluto médio de 26,97.

2. Trabalhos Relacionados

Diversas pesquisas têm investigado algoritmos de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM) para corrigir automaticamente redações escritas em exames de admissão [Lima et al. 2018, Costa et al. 2020, Filho. et al. 2021]. Essas pesquisas, em geral, buscam estimar um valor numérico (nota) representando a qualidade global da redação com base em diferentes critérios. No contexto das provas no estilo do ENEM, existem trabalhos que buscam estimar a nota geral obtida pela redação [Filho et al. 2019, Costa et al. 2020, Filho. et al. 2021] ou aqueles que focam em competências específicas [Lima et al. 2018, Filho et al. 2018, Passero et al. 2019].

Em [Costa et al. 2020] é apresentada uma revisão sistemática da literatura para analisar o estado da arte de abordagens para correção automática de redações escritas na língua portuguesa do Brasil. A pesquisa observou que a maioria dos trabalhos utiliza uma abordagem híbrida, combinando técnicas de PLN e modelos de AM supervisionados. Além disso, os autores apontam que, em geral, os trabalhos focam nas competências 1 e 2 do ENEM. Por fim, os autores evidenciam que uma das principais limitações identificadas é a inexistência de uma grande base de dados que possa ser utilizada para treinamento e validação das abordagens propostas.

O trabalho de [Palma and Atkinson 2018] propõe um método automático para analisar redações escritas em inglês, combinando modelos sintáticos e semânticos. A abordagem proposta integra atributos linguísticos com padrões de discurso para estimar a nota das redações utilizando técnicas de regressão baseadas em árvores de decisão. Em [Filho. et al. 2021] é realizada uma análise comparativa entre técnicas tradicionais baseadas em engenharia de atributos e algoritmos de Aprendizado Profundo para o problema de correção automática de redações escritas na língua portuguesa do Brasil. Para isso, os autores utilizaram uma base de dados composta por redações e analisaram as cinco competências consideradas pelos avaliadores do ENEM.

Alguns trabalhos focam em aspectos específicos que são comumente avaliados em exames de redação. Por exemplo, o trabalho desenvolvido em [Passero et al. 2019] apresenta um estudo comparativo de diversas abordagens para identificação de fuga ao tema (Competência 2 do ENEM) em redações. As abordagens identificadas na literatura, originalmente propostas para língua inglesa, foram adaptadas para a língua portuguesa e o problema foi modelado como uma tarefa de classificação binária. Para validação da abordagem proposta, os autores utilizaram um corpus com 2.164 redações.

O trabalho de [Filho et al. 2018] apresenta uma abordagem baseada em aprendizado de máquina para avaliação automática da aderência ao tema e à estrutura argumentativa de redações. Para isso, os autores exploram diversas técnicas e aplicam modelos de regressão baseados no algoritmo de máquinas de vetores de suporte. Os autores também adaptaram métodos relacionados à coesão textual para o português e verificaram que eles apresentam resultados promissores na tarefa de verificação da aderência ao tema e da estrutura argumentativa de redações.

É possível encontrar trabalhos na literatura que focam especificamente no aspecto de coesão textual, especialmente para o inglês. Tradicionalmente, as abordagens

automáticas que tratam da coesão textual analisam aspectos como a sobreposição de elementos sintáticos e/ou semânticos entre frases e/ou parágrafos adjacentes, a diversidade léxica, medidas de legibilidade, entre outros. Em [Crossley et al. 2019] é apresentada a *Tool for the Automatic Analysis of Cohesion* (TAACO 2.0), uma ferramenta de código-livre capaz de extrair diversos atributos linguísticos relacionados a coesão textual.

O trabalho de [Lima et al. 2018] introduz uma abordagem para correção automática de redações, especificamente conforme os critérios avaliados de coesão textual definidos na competência 4 do ENEM. Os autores apresentam um modelo de classificação baseado no algoritmo de máquina de vetores de suporte treinado a partir de atributos específicos de coesão textual e outras características gerais extraídas das redações. Em [Junior and Fileto 2021] é investigado a aplicação de variações do modelo linguagem *Bidirectional Encoder Representations from Transformers* (BERT) para classificar e estimar coesão textual em português. Para isso, os autores utilizaram a versão em português do modelo BERT e realizaram experimentos considerando textos de artigos de notícias e postagem em fóruns educacionais.

De maneira similar a investigação desenvolvida por [Lima et al. 2018], neste trabalho focamos na análise de atributos para estimar a coesão textual em redações escritas em português no contexto do ENEM. Contudo, ao invés de abordar a tarefa como um problema de classificação, a tratamos como um problema de regressão, cujo objetivo é estimar a nota da competência 4 do ENEM. Para isso, realizamos uma busca na literatura para identificar atributos relacionados com aspectos de coesão textual e investigamos diferentes modelos de regressão para estimar a nota da coesão textual, seguindo os mesmos critérios adotados durante as avaliações das redações do ENEM.

3. Perguntas de Pesquisa

Como mencionado nas seções anteriores, o desenvolvimento de métodos para automatizar a análise da coesão textual em textos educacionais é essencial para auxiliar os estudantes na sua preparação. Embora existam trabalhos na área de análise de redações em português, poucos focaram especificamente na análise de coesão textual. Por essa razão, a primeira pergunta de pesquisa é:

Pergunta de Pesquisa 1 (PP1):

É possível utilizar técnicas de PLN e regressão para predizer notas relacionadas a coesão textual (Competência 4 do ENEM)?

Para além da identificação das notas relacionadas à coesão textual, também é importante analisar a relevância de características para esse problema em específico. Para tanto, a segunda pergunta de pesquisa é:

Pergunta de Pesquisa 2 (PP2):

Quais atributos utilizados para construir os modelos de regressão são mais relevantes para a coesão textual das redações no contexto do ENEM?

4. Método

4.1. Corpus Essay-BR

Neste trabalho foi utilizado o corpus do Essay-BR desenvolvido por [Marinho et al. 2021]. O corpus possui 4.570 redações no estilo da prova do ENEM,

divididas em 86 tópicos. As redações foram extraídas dos portais Vestibular UOL[3] e Educação UOL[4], coletadas de dezembro de 2015 até abril de 2020. As redações disponíveis nesses portais são escritas por estudantes do ensino superior, avaliadas por especialistas seguindo os mesmos critérios de correção da prova do ENEM. Cada redação possui uma nota geral e notas individuais das cinco competências avaliadas no ENEM. A Tabela 1 apresenta algumas estatísticas descritivas do corpus Essay-BR. Como o foco deste trabalho é na Competência 4, as redações foram agrupadas pelas notas atribuídas a esse aspecto e para cada nota as seguintes estatísticas foram geradas: total de redações, média de frases e palavras por redação (desvio padrão entre parênteses).

Tabela 1. Estatísticas descritivas do corpus Essay-BR.

Notas Competência 4	#Redações	Média de Frases	Média de Palavras
0	134	8,80 (3,96)	216,16 (64,39)
40	61	9,93 (4,07)	219,62 (88,75)
80	590	10,63 (5,67)	244,32 (90,29)
120	2.000	11,90 (4,87)	285,38 (78,78)
160	1.241	13,37 (4,19)	325,44 (70,10)
200	544	13,92 (4,57)	337,53 (68,21)
Total	*4.570*	*56.019*	*1.344.757*

4.2. Extração das Características

Neste trabalho, analisamos a coesão textual de uma redação como um problema de regressão, cujo objetivo é estimar um valor que reflita a qualidade do texto da redação com base em sua coesão textual. Para isso, utilizamos como guia os critérios de avaliação adotados na Competência 4 do ENEM. No total, 151 características foram adotadas para construir o modelo de regressão responsável por atribuir a cada redação um escore representando uma estimativa de sua nota. As características adotadas neste trabalho são estruturadas em seis grupos, descritos brevemente a seguir:

Uso de Conectivos. As medidas foram computadas a partir da incidência do uso de conectivos nas frases e orações extraídas das redações. A ideia desse grupo de atributos é refletir o aspecto de coesão sequencial, ou seja, o encadeamento estrutural entre os elementos (frases e orações) que compõem a redação [Antunes 2005]. Para isso, 36 características foram calculadas considerando 17 tipos de conectivos com base na lista disponibilizada no portal Brasil Escola[5] mais uma contagem geral.

Diversidade Léxica. As medidas de diversidade léxica indicam o quão variado é o uso léxico nas redações. Essas medidas foram computadas a partir da razão entre a ocorrência de alguns tipos de palavras e englobam substantivos, verbos, adjetivos, advérbios, pronomes, entre outros. Além dessas medidas, foram computados os seguintes três índices com base no trabalho de [Palma and Atkinson 2018]: *Hapax Legomena*, *Yule's K*, *Guiraud's Index*. No total, foram geradas 15 características neste grupo.

Legibilidade. As medidas desse grupo tem por objetivo determinar o grau de facilidade de leitura de um texto em relação à diversidade lexical, complexidade das pa-

[3]https://vestibular.brasilescola.uol.com.br/banco-de-redacoes
[4]https://educacao.uol.com.br/bancoderedacoes/
[5]https://brasilescola.uol.com.br/redacao/conectores-discursivos.htm

lavras, tamanho das frases, entre outros fatores. Para isso, os seguintes 5 atributos foram computados com base no trabalho de [Palma and Atkinson 2018]: Média de sílabas por palavras e as medidas *Flesch Reading Ease*, *Gunning Fog Index*, *Word Variation Index*, *Automated Readability Index*.

Sobreposição de Frases. Foram extraídos 8 índices de sobreposição visando capturar indícios de coesão referencial no texto. Seis características foram extraídas utilizando o modelo de grade entidades que busca capturar a coesão local de um texto [Lapata and Barzilay 2005]. A intuição desse modelo é que entidades (sintagmas nominais) compartilhadas por frases subsequentes contribuem para a coesão local de um texto. Assim, quanto mais conectadas são as frases adjacentes, melhor a coesão local. Por fim, foram calculadas as médias de sobreposição de unigramas e a similaridade do cosseno usando a medida *Term Frequency–Inverse Document Frequency* (TF-IDF) entre frases adjacentes das redações.

Coh-Metrix. Oitenta e três atributos foram computados utilizando a adaptação da ferramenta Coh-Metrix [Graesser et al. 2011] desenvolvida para o português em [Camelo et al. 2020]. A ferramenta Coh-Metrix é comumente usada na literatura para extrair medidas de coesão e coerência a partir de textos escritos ou falados. Esses atributos são gerados verificando a sobreposição de elementos como, por exemplo, substantivos e pronomes entre frases adjacentes, sobreposição de verbos, adjetivos, advérbios, entre outros atributos presentes no Coh-Metrix.

Outros. Os atributos desse grupo refletem aspectos gerais das redações, sendo eles: **(i)** total de frases; **(ii)** total de palavras classificadas como *stop words*; **(iiii)** total de frases; e **(iv)** média de palavras por frase.

4.3. Seleção e Avaliação dos Modelos de Regressão

Neste trabalho, foram analisados os seguintes algoritmos de regressão[Ferreira-Mello et al. 2019], disponíveis na biblioteca *scikit-learn*[6]: *Extremely Randomized Trees*, *Gradient Boost*, Perceptron de múltiplas camadas, do inglês *Multilayer Perceptron* (MLP), regressão linear e regressão de vetores de suporte, do inglês *Support Vector Regression* (SVR). Na Tabela 2 são apresentados os algoritmos utilizados e alguns dos principais parâmetros e valores adotados nos experimentos realizados.

Tabela 2. Algoritmos de regressão e alguns dos valores dos parâmetros utilizados neste trabalho.

Algoritmos	Parâmetros
Extremely Randomized Trees	*n_estimators* = 1000
Gradient Boost	*learning_rate* = 0,1 *n_estimators* = 1000
MLP	*hidden_layer_sizes* = 100 *solver* = "adam" *learning_rate* = 0,0001 *max_iter* = 1000
SVR	*C* = 1,0 *epsilon* = 0,1
Regressão Linear	*valores padrões da biblioteca*

[6]https://scikit-learn.org/stable/

A metodologia de validação cruzada estratificada com cinco subconjuntos (5-*fold Stratified Cross Validation*) foi adotada nos experimentos realizados para avaliar o desempenho dos algoritmos de regressão, com base nas seguintes medidas [Freund et al. 2006]: **(i)** Correlação de Pearson; **(ii)** Erro quadrático médio, do inglês *Root Mean Squared Error* (RMSE); e **(iii)** Erro absoluto médio, do inglês *Mean Absolute Error* (MAE).

Como pode ser observado na Tabela 1, existe um grande desbalanceamento no número de redações por nota. Por exemplo, poucas redações são observadas para as notas 0 e 40, enquanto para as notas 120 e 160 existem muitas ocorrências. Para tentar mitigar esse problema foram utilizados os seguintes métodos de balanceamento disponíveis na biblioteca *Imbalanced-learn*[7]: Sobreamostragem aleatória, do inglês *Random Oversampling*, e a Técnica de sobreamostragem minoritária sintética, do inglês *Synthetic Minority Oversampling Technique* (SMOTE). Esses métodos foram escolhidos por serem comumente usados na literatura em problemas de aprendizado supervisionado em bases de dados desbalanceadas [Kaur et al. 2019, Filho et al. 2019]. Essas técnicas foram utilizadas somente na etapa de treinamento durante o processo de validação cruzada, resultando em um conjunto de treino balanceado contendo 1.600 redações para cada um dos grupos de notas (0-200).

5. Experimentos

5.1. PP1: Avaliação dos Algoritmos de Regressão

O objetivo deste primeiro experimento é avaliar o desempenho dos algoritmos de regressão considerados para estimar as notas da competência 4 (C4) das redações. Os valores das 151 características usadas para treinar os modelos foram normalizados para ter novos valores no intervalo entre 0 e 1. Na Tabela 3 são apresentados os resultados deste experimento considerando as medidas de correlação de Pearson, RMSE e MAE. Os valores das notas da C4 são números inteiros, enquanto as notas estimadas pelos modelos são valores contínuos. Por isso, as notas estimadas pelos modelos foram truncadas, sendo descartadas as casas decimais.

O algoritmo *Extremely Randomized Trees* sem a utilização de nenhuma técnica de balanceamento obteve a maior correlação de Pearson com 53,08% que é considerada uma correlação moderada positiva com a nota da C4. Tal resultado significa que as notas estimadas pelo algoritmo tendem moderadamente a seguir o mesmo comportamento das notas reais da C4. Além disso, essa configuração também obteve os menores valores de erro nas medidas RMSE (35,94) e MAE (26,97). Dada que a média de notas da C4 no corpus Essay-BR é de 130,63, o valor de RMSE representa um percentual de erro de 27,51% em relação à nota real.

O uso das técnicas de balanceamento analisadas resultou em uma diminuição dos valores da correlação de Pearson e em um aumento nos valores de erros em ambas as medidas (RMSE e MAE), comportamento similar a trabalhos anteriores [Barbosa et al. 2020]. Analisando as notas estimadas, observou-se que quando o corpus original (desbalanceado) foi usado durante a etapa de treinamento, os modelos apresentaram a tendência de gerar notas maiores (superestimados) durante os testes. Esse fato ocorreu principalmente com as redações com notas baixas (0, 40 e 80) devido à quantidade baixa de exemplos. Com a aplicação das técnicas de balanceamento durante o

[7]https://imbalanced-learn.org/

Tabela 3. Resultados dos experimentos comparando os modelos de regressão no corpus Essay-BR.

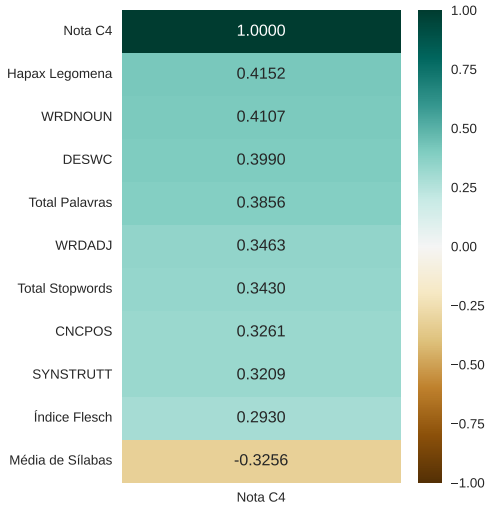
Regressores	Balanceador	RMSE	MAE	Pearson
Extremely Randomized Trees	-	**35,94**	**26,97**	**53,08**
	Random Oversampling	36,09	27,21	52,63
	SMOTE	37,28	28,51	51,58
Gradient Boost	-	37,55	28,36	48,98
	Random Oversampling	40,43	31,14	45,07
	SMOTE	40,59	31,39	45,02
MLP	-	36,44	27,45	51,64
	Random Oversampling	47,06	36,49	42,55
	SMOTE	48,03	37,36	42,86
Regressão Linear	-	36,33	27,31	51,86
	Random Oversampling	46,16	36,34	47,87
	SMOTE	48,50	38,05	46,39
SVR	-	37,46	27,81	50,67
	Random Oversampling	41,64	32,85	50,95
	SMOTE	43,35	34,39	50,07

treinamento, os modelos tendem a estimar notas com valores menores e mais próximos das notas com valores inferiores (0, 40 e 80), contudo isso impactou na geração de erros maiores nas redações com notas de 100 até 160.

5.2. PP2: Avaliação das Características

Neste segundo experimento as 151 características adotadas (ver Seção 4.2) foram analisadas visando investigar o seu impacto na estimação da nota da competência 4 do ENEM. A Figura 1 apresenta as dez características com correlações de Pearson mais significativas.

Figura 1. Top-10 atributos com correlações de Pearson mais significativas com a nota da competência 4 do ENEM.



Das dez características com correlações de Pearson mais significativas com a nota da C4 das redações, cinco foram extraídas do Coh-Metrix (*WRDNOUN*, *DESWC*, *WR-*

DADJ, *CNCPOS* e *SYNSTRUTT*). As medidas *WRDNOUN*, *WRDADJ* e *CNCPOS* verificam a incidência de substantivos, adjetivos e alguns conectivos nas redações, respectivamente. O atributo *DESWC* representa o total de palavras da redação e o *SYNSTRUTT* a média da distância entre os parágrafos da redação, medida com base na sobreposição de palavras.

A medida *Hapax Legomena* apresentou o maior valor de correlação de Pearson positivo (0,4152) com a nota da C4. Essa medida é computada verificando a quantidade de palavras que ocorrem uma única vez na redação, considerando a forma base (*lemma*) das palavras e removendo símbolos de pontuação. A ferramenta Spacy[8] foi usada para realizar o pré-processamento das redações. Já o atributo *Média de Sílabas* apresentou o menor valor de correlação negativo (-0,3256) com a nota da C4. Essa medida é computada verificando a média de sílabas usadas nas palavras da redação. Para isso foi utilizada a ferramenta Pyphen[9].

De maneira geral, os dez atributos com correlações mais significativas possuem valores considerados fracos com a nota da C4. Nenhum dos atributos investigados apresentou uma correlação moderada ou forte com a nota da C4 neste experimento. Tal comportamento é esperado dada a complexidade envolvida no processo de avaliação da coesão textual das redações. Contudo, podemos observar que alguns dos atributos adotados demonstram ser indicadores que podem ajudar na análise. Os demais atributos omitidos apresentam correlações variando de insignificante a fraca.

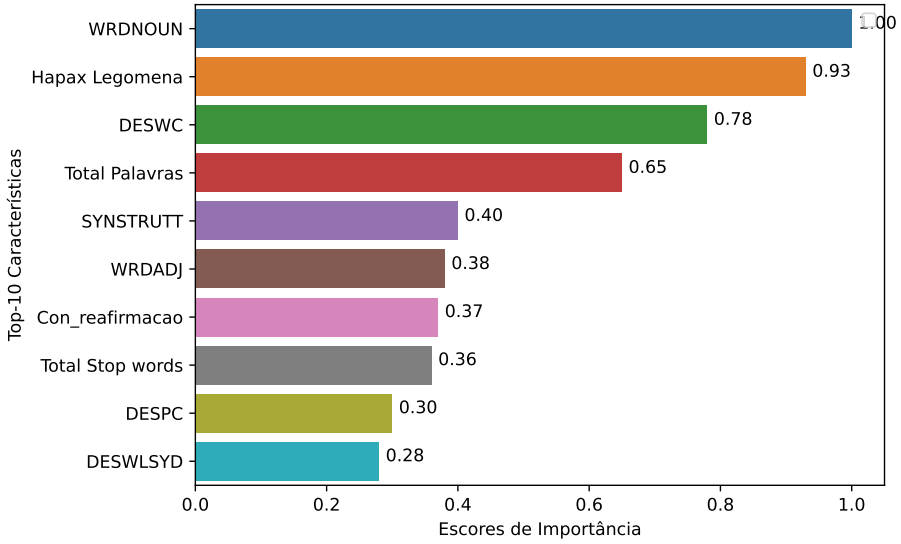
Na Figura 2 são apresentadas as dez características com maior importância baseada em impureza computadas pelo algoritmo *Extremely Randomized Trees*. Os valores de importância foram gerados usando todas as 4.570 redações do corpus Essay-BR. Algumas das medidas presentes no top-10 com base na correlação de Pearson, também foram apontadas como as mais importantes pelo algoritmo regressão, sendo elas: *WRDNOUN*, *Hapax Legomena*, total de palavras, *DESWC*, *WRDADJ*, total de *stop words* e *SYNSTRUTT*. Tal resultado reforça que essas medidas são indicadores relevantes para analisar a coesão textual em redações.

Outras características apontadas como importantes pelo algoritmo *Extremely Randomized Trees* foram: **(i)** *Con. Reafirmação* que computa a incidência de conectivos comumente usados entre frases e orações para reafirmar ideias já apresentadas anteriormente, por exemplo, as expressões "em outras palavras", "dessa forma", entre outras; **(ii)** *DESPC* que é uma medida do Coh-Metrix que verifica a quantidade parágrafos presentes na redação; e **(iii)** *DESWLSYD* que é outra medida do Coh-Metrix que representa o valor do desvio padrão do número médio de sílabas nas palavras da redação.

Os resultados obtidos em ambas as análises demonstram que as redações com maiores notas na C4 tendem a: **(i)** possuir uma maior diversidade no uso do seu vocabulário (medidas Hapax Legomena, WRDNOUN, DESWC e WRDADJ); **(ii)** adotar mais palavras e parágrafos do que as redações com notas baixas (Total de palavras e *stop words*, e a medida DESPC); **(iii)** utilizar conectivos para interligar as frases (medidas CNCPOS e Con. Reafirmação); **(iv)** apresentar um certo nível de sobreposição entre os parágrafos (medida SYNSTRUTT); **(v)** serem fáceis de ler (medida Índice Flesch); e **(vi)** adotar

[8]https://spacy.io/
[9]https://pyphen.org/

Figura 2. Top-10 características com maior importância com base no algoritmo *Extremely Randomized Trees* treinado no corpus Essay-BR completo.



palavras com poucas sílabas (medidas *DESWLSYD* e Média de Sílabas).

6. Considerações Finais e Trabalhos Futuros

Este trabalho investigou a aplicação de algoritmos de regressão para estimar a coesão textual de redações escritas no contexto do Exame Nacional do Ensino Médio (ENEM). Para isso, foram utilizadas 151 características identificadas na literatura relacionadas à coesão textual que refletem aspectos como o uso de conectivos, diversidade léxica, legibilidade, sobreposição de informações entre frases, Coh-Metrix, entre outras.

Experimentos foram realizados usando o corpus Essay-BR [Marinho et al. 2021] para avaliar as características utilizadas e comparar a utilização de diversos algoritmos de regressão. Para isso, a nota da competência 4 do ENEM relacionada com o aspecto de coesão textual da redação foi usada como atributo alvo. Os resultados experimentais obtidos indicaram que o algoritmo *Extremely Randomized Trees* obteve o melhor desempenho geral com um erro quadrático médio de 35,94 e uma correlação de Pearson moderada de 53,08. Além disso, os resultados demonstraram que as dez características mais relevantes com base na medida de correlação de Pearson foram: o índice de Hapax Legomena, a incidência de substantivos e adjetivos (WRDNOUN e WRDADJ respectivamente), total de palavras com e sem a remoção de *stop words* da redação, total *stop words*, uso de conectivos (CNCPOS), sobreposição de palavras entre os parágrafos (SYNSTRUTT), índice Flesch de legibilidade e a média de sílabas usadas nas palavras.

Como trabalhos futuros, vislumbram-se: **(i)** expandir o conjunto de características adotadas incorporando outras medidas que possam refletir mais aspectos considerados na coesão textual; **(ii)** integrar o modelo proposto com outras abordagens para analisar as outras competências consideradas durante a avaliação das redações do ENEM como, por exemplo, detecção de plágio, erros ortográficos, entre outros; e **(iii)** explorar a utilização de métodos de Inteligência Artificial explicável para geração de *feedback* personalizado.

Referências

Antunes, I. (2005). *Lutar com palavras: coesão e coerência*. Parábola Editorial, São Paulo.

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanović, V., and Gašević, D. (2020). Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 605–614.

Camelo, R., Justino, S., and Mello, R. (2020). Coh-metrix pt-br: Uma api web de análise textual para a educação. In *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*, pages 179–186, Porto Alegre, RS, Brasil. SBC.

Costa, L., Oliveira, E., and Júnior, A. C. (2020). Corretor automático de redações em língua portuguesa: um mapeamento sistemático de literatura. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1403–1412, Porto Alegre, RS, Brasil. SBC.

Crossley, S. A., Kyle, K., and Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51(1):14–27.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Filho., A., Concatto., F., Antonio do Prado., H., and Ferneda., E. (2021). Comparing feature engineering and deep learning methods for automated essay scoring of brazilian national high school examination. In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, pages 575–583. INSTICC, SciTePress.

Filho, A. H., Concatto, F., Nau, J., do Prado, H. A., Imhof, D. O., and Ferneda, E. (2019). Imbalanced learning techniques for improving the performance of statistical models in automated essay scoring. *Procedia Computer Science*, 159:764–773. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 23rd International Conference KES2019.

Filho, A. H., do Prado, H. A., Ferneda, E., and Nau, J. (2018). An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science*, 126:788–797. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia.

Freund, R. J., Wilson, W. J., and Sa, P. (2006). *Regression analysis*. Elsevier.

Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

Junior, O. B. and Fileto, R. (2021). Investigando coerência em postagens de um fórum de dúvidas em ambiente virtual de aprendizagem com o bert. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 749–759, Porto Alegre, RS, Brasil. SBC.

Kaur, H., Pannu, H. S., and Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.*, 52(4).

Kellogg, R. T. and Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, 14:237–242.

Klein, R. and Fontanive, N. (2009). Uma nova maneira de avaliar as competencias escritoras na redação do enem. *Ensaio: Avaliação e Políticas Públicas em Educação*, 17(65):585–598.

Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In *IJCAI*, pages 1085–1090.

Lima, F., Haendchen Filho, A., Prado, H., and Ferneda, E. (2018). Automatic evaluation of textual cohesion in essays. In *19th International Conference on Computational Linguistics and Intelligent Text Processing*.

Marinho, J., Anchiêta, R., and Moura, R. (2021). Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64, Porto Alegre, RS, Brasil. SBC.

Palma, D. and Atkinson, J. (2018). Coherence-based automatic essay assessment. *IEEE Intelligent Systems*, 33(5):26–36.

Passero, G., Ferreira, R., and Dazzi, R. L. S. (2019). Off-topic essay detection: A comparative study on the portuguese language. *Revista Brasileira de Informática na Educação*, 27(03):177–190.