AVALIAÇÃO AUTOMÁTICA DE REDAÇÕES POR MEIO DE APRENDIZADO DE MÁQUINA

Pâmela Camilo Chalegre (PIC/UEM), Valéria Delisandra Feltrim (Orientadora). E-mail: vdfeltrim@uem.br.

Universidade Estadual de Maringá, Centro de Tecnologia, Maringá, PR.

Área e subárea do conhecimento: Ciências Exatas e da Terra/Metodologia e Técnicas da Computação.

Palavras-chave: Processamento de linguagem natural; NILC-Metrix; redações ENEM.

RESUMO

A predição automática de notas de redações (AES - *Automatic Essay Scoring*) permite a otimização de processos avaliativos em larga escala, como o Exame Nacional do Ensino Médio (ENEM). Este trabalho investigou a eficácia das métricas de complexidade textual NILC-Metrix na AES de um *corpus* de redações do ENEM. As métricas NILC-Metrix foram utilizadas como representação para as redações e foram treinados modelos de aprendizado de máquina para regressão, regressão ordinal e classificação para prever as notas por competência e também a nota final. Os resultados foram comparados a uma abordagem tradicional baseada em TF-IDF (*Term Frequency-Inverse Document Frequency*) e a uma abordagem híbrida, que combina representações. As análises mostraram que a combinação de TF-IDF e NILC-Metrix gerou modelos mais robustos, alcançando melhor desempenho em métricas de concordância e acurácia, evidenciando o potencial da integração de múltiplas representações textuais para o avanço da AES no contexto estudado.

INTRODUÇÃO

A correção manual de redações, especialmente em avaliações de larga escala como o ENEM (SILVEIRA et al., 2024), é um processo lento, trabalhoso e subjetivo. Como alternativa, a área de Processamento de Linguagem Natural (PLN) tem desenvolvido soluções de Avaliação Automática de Redações (AES), sendo que alguns deles são amplamente utilizados em exames internacionais como TOEFL e GRE (SHERMIS; BURSTEIN, 2013).

No Brasil, a pesquisa em AES é limitada pela escassez de *corpus* públicos anotados. Diante da recente publicação do NILC-Metrix (LEAL et al., 2023), uma ferramenta que calcula cerca de 200 métricas de complexidade para o português, este trabalho se propôs a investigar o potencial dessas métricas para a tarefa de AES no contexto do ENEM.

MATERIAIS E MÉTODOS

O estudo foi desenvolvido com um subconjunto do *corpus* de redações criado por Silveira et al. (2024), chamado *Source A*. Tais redações foram escritas entre os anos de 2015 e 2020 e extraídas do site Educação UOL, sendo que a maioria delas conta com três avaliações: uma realizada por avaliadores do site e duas realizadas por novos avaliadores, totalizando 1.164 redações manualmente avaliadas.

Neste projeto foram avaliadas três abordagens distintas de representação textual. A primeira, utilizada como *baseline*, empregou a técnica TF-IDF, que pondera a relevância das palavras (JURAFSKY; MARTIN, 2024). A segunda abordagem se baseou nas 200 métricas de complexidade textual extraídas pela ferramenta NILC-Metrix (LEAL et al., 2023). Tal processo contou com a colaboração da equipe do NILC-USP/São Carlos para viabilizar o cálculo de características que dependiam de licença comercial. A terceira abordagem, de natureza híbrida, foi criada para investigar a complementaridade das representações. Nela, as 50 características mais relevantes de cada uma das abordagens anteriores (TF-IDF e NILC-Metrix) foram selecionadas por meio de testes estatísticos de correlação com as notas e concatenadas, resultando em um vetor de 100 características por redação. Na etapa de modelagem, as representações NILC-Metrix e Híbrida foram padronizadas utilizando a função *StandardScaler* da biblioteca *scikit-learn.*

Seguindo a metodologia de Silveira et al. (2024), as notas das competências (0-200) foram mapeadas para uma escala ordinal de 0 a 5. Foram avaliados diversos algoritmos da biblioteca *scikit-learn*, incluindo modelos lineares, vizinhos mais próximos, *ensembles*, árvores de decisão, redes neurais e SVM tanto para regressão quanto para classificação (otimizado com *GridSearchCV*). Também foi utilizada a biblioteca *mord* para regressão ordinal. As partições de treino e teste foram as mesmas utilizadas por Silveira et al. (2024), que consistiu de 720 redações para treino e 200 redações para teste. Segundo o autor, tais partições preservaram o equilíbrio entre os temas das redações, garantindo que textos referentes a um mesmo comando permanecessem no mesmo subconjunto, evitando vazamento de informação entre as partições. No total, foram avaliados 42 modelos de regressão e 3 modelos de classificação. O desempenho dos regressores foi avaliado por métricas de regressão (R^2, MSE, RMSE, MAE), pela acurácia no padrão ENEM (diferença até 80 pontos) e pelo Kappa Ponderado Quadrático (QWK). O desempenho dos classificadores foi avaliado em termos de acurácia, precisão, revocação e *f1-score*.

RESULTADOS E DISCUSSÃO

A Tabela 1 apresenta os melhores resultados obtidos pelos regressores para cada competência e nota final com as diferentes representações avaliadas. Foram selecionados os modelos com os maiores valores de QWK e acurácia no padrão ENEM, e, ao mesmo tempo, as menores taxas de erro (RMSE, MSE e MAE). Para maior clareza, apenas um subconjunto das métricas é apresentado na tabela.

Tabela 1: Principais resultados das abordagens TF-IDF, NILC-Metrix e Híbrida

Comp.	Repres.	Melhor Modelo	RMSE	QWK	ENEM Acc.
C1	TF-IDF	Gradient Boosting Regressor	1,20	0,05	0,90
	NILC-Metrix	Ordinal Regression LogisticAT	1,29	0,00	0,92
	Híbrida	Hist Gradient Boosting Regressor	1,17	0,17	0,91
C2	TF-IDF	Least Absolute Deviation	1,39	0,24	0,93
	NILC-Metrix	Least Absolute Deviation	1,49	0,03	0,92
	Híbrida	Least Absolute Deviation	1,49	0,11	0,91
C3	TF-IDF	Least Absolute Deviation	1,42	0,19	0,92
	NILC-Metrix	Ordinal Regression LogisticIT	1,40	0,00	0,92
	Híbrida	MLP	1,39	0,21	0,83
C4	TF-IDF	Least Absolute Deviation	1,20	0,23	0,94
	NILC-Metrix	Least Absolute Deviation	1,18	0,00	0,92
	Híbrida	Least Absolute Deviation	1,15	0,24	0,95
C5	TF-IDF	Least Absolute Deviation	1,46	0,22	0,93
	NILC-Metrix	KNeighborsRegressor	1,56	0,02	0,82
	Híbrida	Least Absolute Deviation	1,43	0,32	0,93
Nota Final	TF-IDF	MLP	199,80	0,41	0,33
	NILC-Metrix	MLP	233,74	0,01	0,30
	Híbrida	MLP	215,28	0,25	0,28

Em termos de QWK, os resultados foram baixos, evidenciando a dificuldade da tarefa e a limitação das representações. Considerando-se as três abordagens avaliadas, a híbrida se mostrou a mais vantajosa, pois alcançou os melhores índices de concordância na predição das competências C1 (0,17), C3 (0,21), C4 (0,24) e C5 (0,32). A abordagem TF-IDF se mostrou mais eficaz para a C2 (0,24) e para a Nota

Final (0,41). Em contrapartida, o uso isolado das métricas do NILC-Metrix apresentou os resultados mais baixos em todas as frentes, com QWK próximo ou abaixo de zero, indicando que, sozinhas, essas características são insuficientes para a tarefa.

Ao modelar o problema como uma tarefa de classificação, os resultados de acurácia se mantiveram modestos, porém próximos dos reportados por Silveira et al. (2024) com o uso de modelos baseados em redes neurais profundas. Na classificação, a representação TF-IDF atingiu o melhor valor para C1 (0,43) e C4 (0,46), a híbrida para C2 (0,35) e a NILC-Metrix obteve o maior valor para C3 (0,30) e C5 (0,19). Esses resultados estão entre 3%-10% abaixo dos obtidos por Silveira et al. (2024) e reforçam que as competências C2, C3 e C5 são mais difíceis de serem avaliadas automaticamente do que as competências C1 e C4.

CONCLUSÕES

Este trabalho avaliou as métricas de complexidade textual do NILC-Metrix na tarefa de Avaliação Automática de Redações no padrão ENEM. Os resultados indicam que, embora relevantes, tais métricas são insuficientes quando usadas de forma isolada. A principal contribuição do estudo foi mostrar que ao combinar características lexicais e de complexidade, em uma representação híbrida, é possível obter resultados próximos aos obtidos por modelos mais complexos e computacionalmente caros, como os usados por Silveira et al. (2024). Conclui-se, portanto, que a integração de múltiplas representações textuais é um caminho promissor para o avanço dos sistemas de AES no contexto brasileiro, alinhando a avaliação automática à complexidade da grade de correção do exame.

REFERÊNCIAS

JURAFSKY, D.; MARTIN, J.H. **Speech and Language Processing**, 3rd Ed., Online manuscript, 2024. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 25 ago. 2025.

LEAL, S. E.; DURAN, M. S.; SCARTON, C. E.; HARTMANN, N. S.; ALUÍSIO, S. M. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **Language Resources & Evaluation**, v. 58, p 73-110, 2023.

SHERMIS, M.D.; BURSTEIN, J.; BURSKY, S.A. Introduction to Automated Essay Evaluation. In: SHERMIS, M.D.; BURSTEIN, J. (Org.). **Handbook of Automated Essay Evaluation: Current Applications and New Directions**. 1st ed. New York: Taylor & Francis, 2013, p. 1-15.

SILVEIRA, I. C.; BARBOSA, A.; MAUÁ, D. D. A New Benchmark for Automatic Essay Scoring in Portuguese. In: 16th INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE (PROPOR), 1., 2024, Galiza. **Proceedings**... Association for Computational Linguistics, 2024.