Explorando o Aprendizado de Máquina para suporte no reconhecimento de sintomas de dislexia em crianças em processo de alfabetização

Lui Gill Aquini, Eduarda Pereira Medeiros, Tiago Duarte Mackedanz, Laura Quevedo Jurgina, Tiago Thompsen Primo, Leomar Soares da Rosa Júnior

[1]Centro de Desenvolvimento Tecnológico - Universidade Federal de Pelotas (UFPel)
R. Gomes Carneiro, 01 - Porto - CEP 96010-610, Pelotas - RS, Brasil

{lgaquini, epmedeiros, tdmackedanz, lqjurgina, tiago.primo, leomarjr}@inf.ufpel.edu.br

Abstract. *This work aims to develop and integrate a machine learning algorithm into the Alfaba device, an educational tool designed to support the literacy process, particularly for students with dyslexia. The algorithm was created using the Affinity Propagation technique to analyze and group common errors in word construction, providing feedback tailored to the students' needs. Through synthetic data, the model was tested and demonstrated the ability to identify and cluster writing errors. This study highlights the potential of the algorithm to enhance Alfaba's performance, making it an even more robust tool for supporting students with learning difficulties.*

Resumo. *Este trabalho tem como objetivo desenvolver e integrar um algoritmo de aprendizado de máquina ao dispositivo Alfaba, uma ferramenta educacional projetada para apoiar o processo de alfabetização, especialmente em estudantes com dislexia. O algoritmo foi criado utilizando a técnica de Propagação por Afinidade para analisar e agrupar erros comuns na construção de palavras, fornecendo feedback adaptado às necessidades dos alunos. Através de dados sintéticos, o modelo foi testado e demonstrou capacidade de identificar e agrupar erros de escrita. Este estudo destaca o potencial do algoritmo para melhorar o desempenho do Alfaba, tornando-o uma ferramenta ainda mais robusta para o suporte educacional de estudantes com dificuldades de aprendizado.*

1. Introdução

A alfabetização é um desafio constante no Brasil, e os efeitos da pandemia de COVID-19 agravaram essa situação, ampliando as lacunas de aprendizado e dificultando a recuperação educacional. Em 2023, apenas 56% dos estudantes do ensino fundamental estavam alfabetizados conforme o padrão nacional, evidenciando uma crise que já existia antes da pandemia [Ministério da Educação 2024]. Em 2016, menos de 50% dos alunos do 3º ano do ensino fundamental atingiram níveis satisfatórios em leitura, destacando uma fragilidade persistente [INEP 2016].

Esses desafios se tornam ainda mais complexos quando se considera o cenário dos estudantes com transtornos de aprendizagem, como a dislexia. Esse transtorno neu-

robiológico dificulta a decodificação precisa de palavras e a associação de sons, impactando diretamente o processo de alfabetização [Associação Brasileira de Dislexia 2016]. A Base Nacional Comum Curricular (BNCC) enfatiza a necessidade de incorporar Tecnologias Digitais de Informação e Comunicação (TDICs) no ambiente escolar para apoiar a diversidade de necessidades dos alunos [Ministério da Educação 2023].

Nesse contexto, o Alfaba foi construído para ser uma ferramenta educacional, desenvolvida para apoiar o processo de alfabetização de estudantes com dislexia. Concebido como um dispositivo multissensorial, o Alfaba se mostrou interessante não apenas para alunos disléxicos, mas também para todos que necessitam de reforço no aprendizado da leitura e escrita [Jurgina et al. 2023, Jurgina et al. 2024]. Além de ser de baixo custo e fácil operação, o Alfaba utiliza estímulos visuais, auditivos e táteis para engajar diferentes estilos de aprendizagem, baseados em conceitos de neuroplasticidade.

Com o crescente uso da Inteligência Artificial (IA) no campo educacional, o Alfaba pode integrar técnicas de Aprendizado de Máquina para identificar e agrupar erros comuns na escrita, oferecendo suporte para o diagnóstico de dislexia, já que outros sintomas também estão presentes em outras patologias, dificultando o diagnóstico. Essa abordagem não só auxilia na detecção de sinais de dislexia, mas também permite adaptar o ensino às necessidades individuais de cada estudante [Chen et al. 2020, Li et al. 2019, Zygouris et al. 2021].

Este trabalho propõe o aprimoramento do Alfaba, ampliando suas funcionalidades através da aplicação de técnicas como Propagação por Afinidade, Redução Dimensional e o algoritmo de Distância de Levenshtein. A hipótese é que, ao identificar e analisar erros comuns, o sistema pode prever e sinalizar casos de dislexia, contribuindo para um diagnóstico precoce e suporte educacional.

O restante deste trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados, enquanto a Seção 3 aborda a fundamentação teórica que embasou a construção do algoritmo. Na Seção 4, são discutidos os resultados obtidos, e a Seção 5 oferece as conclusões e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

Diversos estudos na literatura exploram o uso de técnicas de aprendizado de máquina no contexto educacional, utilizando abordagens que variam em termos de objetivos e métodos. Um exemplo é o trabalho de LIU e JIANG (2012), que emprega a Propagação por Afinidade para agrupar conversas orais [Liu and Jiang 2012]. O estudo de ENE e ENE (2017) utiliza a Distância de Levenshtein em um jogo educativo para o aprendizado de línguas estrangeiras. Embora o trabalho compartilhe conceitos com o Alfaba, como a repetição sonora e a manipulação de letras, ele não faz uso de aprendizado de máquina, limitando-se a técnicas de processamento linguístico [Ene and Ene 2017].

ANGGRAINI e TURSINA (2019) aplicam a Distância de Levenshtein e o algoritmo *K-Nearest Neighbors* para analisar sentimentos em textos sobre mudanças de políticas educacionais [Anggraini and Tursina 2019]. O estudo demonstrou que a combinação dessas técnicas pode proporcionar bons resultados na tarefa de análise de sentimentos. Mais recentemente, XU et al. (2023) utilizam *K-Means* para criar planos de ensino personalizados no aprendizado de chinês vocacional [Xu et al. 2023].

Os trabalhos analisados apresentam diferentes abordagens para o uso de técnicas de aprendizado de máquina e processamento de linguagem na educação. No entanto, o presente estudo avança ao focar na aplicação dessas técnicas para a detecção e agrupamento de erros na construção de palavras, contribuindo de forma específica para o diagnóstico e o suporte educacional no contexto da dislexia.

3. Fundamentação Teórica

A dislexia é um transtorno neurobiológico de aprendizagem que afeta o reconhecimento de letras palavras, a decodificação e a soletração [International Dyslexia Association 2002]. Sinais como trocas e espelhamento de letras, truncamento, repetição ou ausência de sílabas, além de dificuldades motoras são comuns na dislexia. Todavia esses indicativos podem ser confundidos com outros transtornos, tornando necessário o acompanhamento de uma equipe multidisciplinar para um diagnóstico preciso, especialmente durante a alfabetização [Gonçalves and Peixoto 2020, Figueira 2012].

No contexto de análise de dados relacionados à dislexia, técnicas de agrupamento podem ser úteis para identificar padrões subjacentes. A Propagação por Afinidade é uma técnica de agrupamento que utiliza uma medida de similaridade entre os níveis dos dados de entrada e retorna pontos centrais dos agrupamentos e a relação dos outros dados de entrada com esses pontos centrais, chamados de exemplares. Cada ponto é visto como um nó em uma rede, e o algoritmo opera através da troca interativa de mensagens entre os nós, até que as mudanças nos agrupamentos se estabilize [Frey and Dueck 2007].

Ainda em relação às ferramentas que podem ser aplicadas para análise de dados, a técnina de Redução Dimensional *t-Stochastic Neighbor Embedding* tem como objetivo facilitar a visualização de dados com alta dimensionalidade em um espaço diminuto, de normalmente duas ou três dimensões. Apresenta maior capacidade de minimizar a tendência de acúmulo excessivo de pontos no centro do mapa do que outros algoritmos similares [Van der Maaten and Hinton 2008].

Neste contexto, a Distância de Levenshtein tem como objetivo medir o quão diferentes duas *strings* são, por meio de cálculos da quantidade necessária de operações para que as duas palavras se igualem. Este algoritmo é altamente utilizado na área Processamento de Linguagem Natural, no contexto de correção de erros ortográficos [Manning et al. 2008].

Combinando essas ferramentas com Inteligência Artificial (IA), o sistema Alfaba busca auxiliar diretamente no reconhecimento de sinais de dislexia durante o processo de alfabetização. A IA pode identificar padrões de erros comuns e alertar os usuários sobre a necessidade de encaminhamento para uma equipe interdisciplinar para um diagnóstico mais aprofundado e preciso. Assim, essa abordagem visa proporcionar suporte adicional no ambiente educacional, garantindo que as dificuldades dos alunos sejam detectadas e tratadas de forma oportuna.

4. Metodologia

O desenvolvimento do modelo proposto consiste em três etapas: a construção do conjunto de dados, o pré-processamento das informações e o treinamento do modelo de inteligência

artificial. Cada uma dessas etapas é fundamental para garantir que o modelo seja capaz de identificar padrões relevantes nos dados, com o objetivo de auxiliar no diagnóstico da dislexia. A seguir, essas etapas serão detalhadas.

4.1. Conjunto de Dados

O algoritmo desenvolvido será embarcado no dispositivo Alfaba, cujo objetivo é auxiliar no diagnóstico da dislexia. No entanto, como o Alfaba ainda não dispõe de um conjunto de dados extenso com os erros específicos necessários para o desenvolvimento do algoritmo, foram utilizados dados sintéticos como *dataset* inicial. Essa abordagem permite avaliar preliminarmente a viabilidade do agrupamento de erros, preparando o modelo para futuras integrações com dados reais, que serão coletados em testes com crianças disléxicas por meio do Alfaba.

Para tornar os dados sintéticos mais representativos, a lista de palavras com alta taxa de erros gramaticais cometidos por disléxicos, conforme identificado no estudo de [Cidrim and Madeiro 2017], foi utilizada como base. Palavras compostas foram excluídas dessa lista, pois o dispositivo Alfaba, no momento, não oferece suporte para esse tipo de construção.

O *dataset* foi construído manualmente, gerando variantes errôneas das palavras selecionadas. Diferentes categorias de erros, identificadas no estudo de [Zorzi and Ciasca 2009], foram consideradas para simular as dificuldades típicas de leitura e escrita encontradas em indivíduos com dislexia. Esses erros incluem trocas de letras similares, truncamento de palavras e espelhamento de letras, características comuns em disléxicos. O processo de criação do *dataset* foi orientado para assegurar uma ampla representatividade dos tipos de erros, abrangendo desde os erros gramaticais mais comuns, observados tanto em alunos com e sem dislexia, até aqueles mais específicos, utilizados como indicadores potenciais de dislexia.

4.2. Pré Processamento de Dados

A Distância de Levenshtein foi empregada para construir uma matriz de similaridade entre todas as palavras do *dataset*, quantificando as diferenças entre as *strings*. Além disso, uma matriz similar foi gerada para as palavras corretas, que serviriam como exemplares nos agrupamentos.

Como o algoritmo de Propagação por Afinidade requer uma matriz de similaridade em vez de uma matriz de distância, os valores obtidos pela Distância de Levenshtein foram transformados. Essa transformação inverteu as distâncias, de modo que valores menores correspondessem a maiores similaridades. Esse ajuste facilita o processamento das palavras pelo algoritmo, garantindo que palavras com menores distâncias na matriz de Levenshtein fiquem mais próximas de zero na matriz de similaridade, o que, por sua vez, favorece a formação de clusters coerentes.

Esse procedimento constitui a primeira etapa do algoritmo de Propagação por Afinidade, preparando os dados de entrada ao convertê-los em níveis de similaridade adequados para o processamento pelo algoritmo.

4.3. Treinamento do Modelo

A implementação do modelo de Inteligência Artificial foi realizada utilizando a biblioteca *open source scikit-learn*, em sua versão 1.5.1[1]. Essa biblioteca fornece ferramentas e algoritmos eficientes para a solução de tarefas de aprendizado de máquina. Já para a criação de gráficos, foi utilizada a biblioteca de código aberto *matplotlib* na versão 3.8.4[2].

Os agrumantentos foram processados por um computador com processador Ryzen 5 5600 6/12 3.5GHz, 32GB de memória RAM e uma placa de vídeo AMD RX 6600 com 8GB de VRAM. Com pequenos conjuntos de dados, é possível a replicação deste trabalho com opções de *hardware* mais simples. No entanto, ao lidar com volumes maiores de informações, será necessário um poder de processamento mais elevado para realizar os agrupamentos de forma eficiente.

Durante o processo de treinamento, algumas adaptações foram necessárias para garantir que o algoritmo funcionasse corretamente com os dados pré-processados.

O algoritmo de Propagação de por Afinidade foi escolhido como agente ao invés de outras técnicas de agrupamento, como o *K-Means*, graças a sua sinergia com a matriz de similaridade gerada pela Distância de Levenshtein e pela fácil definição dos valores centrais de cada agrupamento, que obrigatoriamente devem ser as palavras presentes no conjunto de dados construídas corretamente.

Embora o cálculo padrão da similaridade entre as entradas da Propagação por Afinidade utilize a Distância Euclidiana, essa similaridade já havia sido calculada na etapa de pré-processamento. Para adequar o modelo a essa característica, o hiperparâmetro de afinidade foi ajustado para *precomputed*. Além disso, o hiperparâmetro *preferences* foi configurado para selecionar os exemplares centrais nos agrupamentos, definindo a probabilidade de cada ponto ser escolhido como exemplar. A matriz de preferência foi gerada multiplicando-se a mediana da matriz de similaridade das palavras corretas por uma matriz de 1s. Para as palavras corretas, as preferências foram ajustadas ao valor máximo da matriz de similaridade, normalizando as entradas para essas palavras.

Outro ajuste importante envolveu o hiperparâmetro *damping*, configurado com o valor padrão de 0,5. Este valor controla a influência dos valores atuais nas atualizações dos nós durante a troca de mensagens, o que contribui para a convergência do algoritmo. Para garantir a reprodutibilidade dos experimentos, o hiperparâmetro *random state* foi fixado em 42, assegurando que os resultados possam ser replicados em futuras execuções.

Com todos os ajustes realizados, o modelo foi treinado e realizou os agrupamentos, organizando as construções errôneas ao redor dos exemplares definidos previamente como palavras corretas. Assim, cada palavra correta em um determinado *cluster* serviu como ponto central para as variantes errôneas correspondentes, demonstrando a eficácia do processo de agrupamento implementado.

5. Resultados

O modelo desenvolvido utilizou a técnica *t-Stochastic Neighbor Embedding* (*t-SNE*) para reduzir a dimensionalidade da saída gerada para duas dimensões, o que possibilitou a

[1] https://github.com/scikit-learn/scikit-learn
[2] https://github.com/matplotlib/matplotlib

construção de um gráfico interpretável dos agrupamentos. O algoritmo de Propagação de Afinidade, aplicado ao conjunto de dados, buscou identificar padrões relevantes, como inversão de letras e truncamento de palavras. Esses padrões são indicativos de possíveis sinais de dislexia, e a correta inserção dos casos nos agrupamentos esperados demonstra a eficácia do modelo.

No entanto, o modelo apresentou algumas anomalias nos *clusters*, especialmente em casos específicos de truncamento de palavras. Isso sugere a necessidade de ajustes futuros para melhorar a precisão na detecção desses padrões. Para avaliar o desempenho do modelo, utilizou-se a *Silhouette Score*, que mede o nível de coesão e separação dos agrupamentos, e o *Índice Calinski-Harabasz*, que avalia a dispersão entre agrupamentos em relação à dispersão interna. Através da Figura 1, é possível observar os agrupamentos realizados pelo modelo, e a Tabela 1 apresenta os resultados obtidos nas métricas de avaliação.

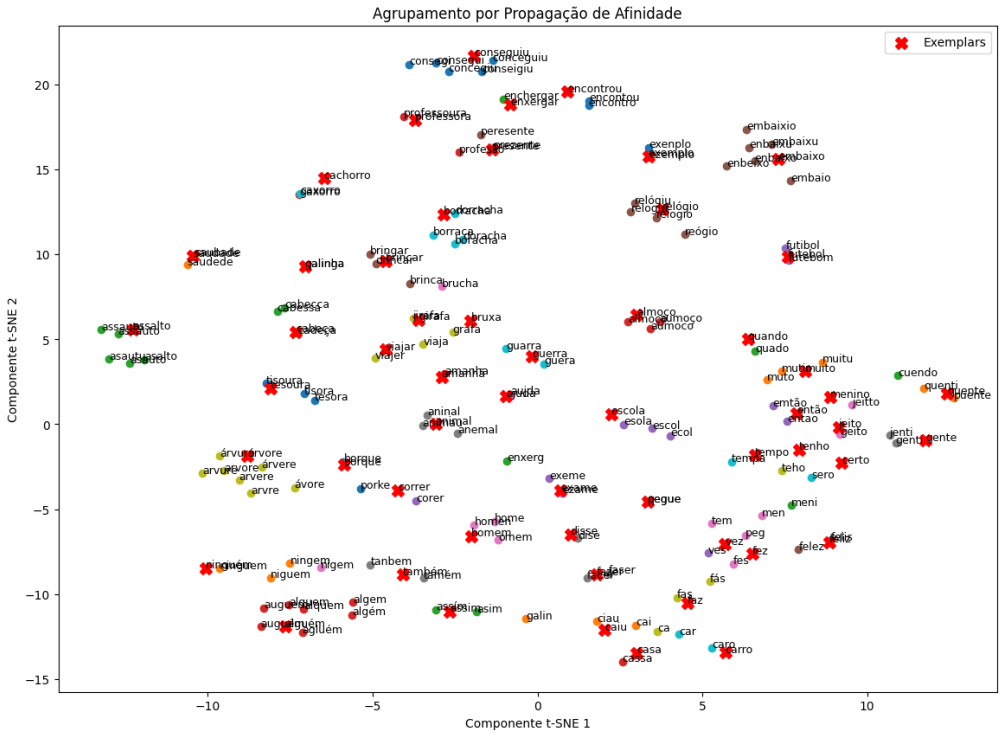

Figura 1. Redução dimensional *t-SNE* em um espaço 2D.
Fonte: Elaborada pelo autor.

Métrica	Valor
Silhouette Score	0.32
Calinski-Harabasz Index	15.5

Tabela 1. Resultados das métricas de avaliação do agrupamento.

Os resultados indicam um desempenho satisfatório, com os *clusters* apresentando um balanço positivo em termos de coesão e separação. No entanto, as anomalias identificadas, particularmente no tratamento de truncamentos de palavras, apontam para a neces-

sidade de melhorias adicionais no modelo para aumentar sua eficácia na identificação de padrões associados à dislexia.

5.1. Caso de Uso

Para ilustrar a eficácia do modelo, foi desenvolvido um caso de uso que simula o funcionamento do algoritmo em um contexto hipotético de aplicação do Alfaba. Considerando um cenário educacional simulado, foi solicitado a um estudante fictício que construísse três palavras utilizando o dispositivo: ”caiu”, ”pegue”e ”borracha”.

O estudante realizou a tarefa utilizando todos os recursos multissensoriais do Alfaba. Após a conclusão do exercício, o modelo de aprendizado de máquina embarcado no dispositivo gerou informações novas que exemplificam as construções realizadas e indicam a proximidade dessas construções em relação às palavras corretas. As palavras formadas por esse estudante hipotético, juntamente com os erros cometidos, estão apresentadas na Tabela 2.

Palavra Correta	Erros Realizados
caiu	ciau, cai
pegue	qegue, peg
borracha	dorracha, doracha, borraca, boracha

Tabela 2. Tabela de palavras corretas e erros comuns

Esses dados gerados pelo modelo de Inteligência Artificial permitem ao educador, em um cenário simulado, observar e identificar possíveis sinais de transtornos de aprendizagem, como a dislexia, constatando a presença de espelhamento de letras e truncamento de palavras.

6. Conclusão

O uso combinado de TDICs e técnicas de Inteligência Artificial demonstra ser uma abordagem potencial para o desenvolvimento de ferramentas educacionais personalizadas que atendam às necessidades específicas dos estudantes de maneira acessível, rompendo barreiras socioeconômicas. Essa combinação promove a democratização do acesso à educação, contribuindo para a redução de desigualdades e lacunas educacionais.

O modelo de Propagação por Afinidade mostrou-se eficiente na identificação e agrupamento de erros comuns na construção de palavras, reconhecendo padrões relevantes para o suporte ao diagnóstico da dislexia, como truncamento de palavras e inversão de letras. Além disso, o uso de algoritmos de Redução Dimensional possibilitou a criação de representações dessas informações, oferecendo uma fonte adicional de estímulo visual para o sistema de *feedback* do Alfaba e proporcionando aos educadores uma ferramenta gráfica para o acompanhamento do desempenho dos alunos.

Apesar dos resultados promissores, é importante reconhecer as limitações do uso de dados sintéticos, que, embora viabilizem a validação inicial do modelo, não capturam completamente a complexidade dos dados educacionais reais sobre dislexia. Adicionalmente, a presença de anomalias em alguns agrupamentos, possivelmente decorrentes do cálculo de similaridade pela Distância de Levenshtein, ressalta a necessidade de aprimoramentos.

Para superar essas limitações, será fundamental a coleta e utilização de dados reais de usuários com dislexia, o que permitirá ao modelo uma representação mais fiel da realidade. Parcerias com instituições educacionais serão essenciais para viabilizar essa coleta. A exploração e implementação de algoritmos de similaridade mais robustos, como técnicas baseadas em *embeddings* de palavras, devem ser consideradas para reduzir as anomalias nos agrupamentos.

Como trabalhos futuros, planeja-se o desenvolvimento de sistemas tutores inteligentes que ofereçam feedback personalizado e estratégias de intervenção adaptativas, utilizando técnicas avançadas de aprendizado de máquina. Outro ponto importante é o desenvolvimento de um sistema de alertas automatizado, em que sinaliza-se que os cometidos pelos educandos são similares a erros de pessoas com dislexia. Também é fundamental realizar avaliações contínuas do modelo em diferentes contextos e com diversos perfis de estudantes, garantindo assim sua eficácia e aplicabilidade em ambientes educacionais variados.

7. Agradecimentos

Agradecemos à FAPERGS, ao CNPq e à CAPES pelo fomento à pesquisa que viabilizou este trabalho.

Referências

Anggraini, N. and Tursina, M. J. (2019). Sentiment analysis of school zoning system on youtube social media using the k-nearest neighbor with levenshtein distance algorithm. In *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, volume 7, pages 1–4. IEEE.

Associação Brasileira de Dislexia (2016). O que e dislexia?

Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278.

Cidrim, L. and Madeiro, F. (2017). Tecnologias da informação e da comunicação (tic) aplicadas à dislexia: revisão de literatura. *Revista CEFAC*, 19:99–108.

Ene, A. and Ene, A. (2017). An application of levenshtein algorithm in vocabulary learning. In *2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE.

Figueira, G. L. M. (2012). Um olhar psicopedagógico sobre a dislexia. *Especialização em Psicopedagogia. Universidade Cândido Mendes*.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

Gonçalves, P. and Peixoto, A. (2020). *10 perguntas e respostas para compreender a Dislexia*, volume 81320. Editora Dialética e Realidade, Curitiba.

INEP (2016). Resultados da ana 2016 por estados e municípios estão disponíveis no painel educacional do inep.

International Dyslexia Association, I. (2002). Definition of dyslexia.

Jurgina, L. Q., Aquini, L. G., de Aguiar, M. S., da Rosa, L. S., Lopes, J. P., Mackedanz, T. D., Klein, A. I., Primo, T. T., and Iankowski, R. S. (2024). Neuroplasticity-based literacy rescue: A multisensory and tangible learning methodology for children at risk. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–10. IEEE.

Jurgina, L. Q., Aquini, L. G., Iankowski, R. S., da Rosa, L. S., de Aguiar, M. S., and Primo, T. T. (2023). Alfaba: A tangible solution to support brazilian dyslexic students in their literacy process. In *2023 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9. IEEE.

Li, G., Wang, Y., and Zhang, H. (2019). A deep learning approach for detecting dyslexia in children. *Journal of Medical Systems*, 43(8):1–9.

Liu, D. and Jiang, M. (2012). Affinity propagation clustering on oral conversation texts. In *2012 IEEE 11th International Conference on Signal Processing*, volume 3, pages 2279–2282. IEEE.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Ministério da Educação (2023). Estudo Internacional de Progresso em Leitura (PIRLS). Acesso em 16 de maio de 2023.

Ministério da Educação (2024). Programa criança alfabetizada. Acesso em: 14 ago. 2024.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Xu, M., Ao, Y., Chang, C., and Wu, Y. (2023). The application of improved k-means clustering algorithm in chinese language education in vocational colleges. In *2023 International Conference on Data Science and Network Security (ICDSNS)*, pages 01–06. IEEE.

Zorzi, J. L. and Ciasca, S. M. (2009). Análise de erros ortográficos em diferentes problemas de aprendizagem. *Revista Cefac*, 11:406–416.

Zygouris, N., Tsolaki, M., and Giakoumakis, E. (2021). Artificial intelligence in the early diagnosis of dyslexia. *Frontiers in Neuroinformatics*, 15:704150.