CONIC SEMESP

25º Congresso Nacional de Iniciação Científica

TÍTULO: SYNAPSPEECH - ESCUTANDO OS PRIMEIROS SINAIS DO ALZHEIMER

CATEGORIA: EM ANDAMENTO

ÁREA: CIÊNCIAS EXATAS, DA TERRA E AGRÁRIAS

SUBÁREA: Computação e Informática

INSTITUIÇÃO: UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ - UTFPR

AUTOR(ES): JOÃO PEDRO MADUREIRA SALES

ORIENTADOR(ES): ROBSON PARMEZAN BONIDIA, ANDRE CARLOS PONCE DE LEON FERREIRA DE CARVALHO

CATEGORIA EM ANDAMENTO

1. RESUMO

No enfrentamento da Doença de Alzheimer, há um desafio silencioso: o da detecção precoce. Para muitas pessoas, especialmente em regiões com recursos limitados, esse passo inicial nem sempre acontece, seja por falta de acesso, de infraestrutura, ou mesmo de informação. É nesse cenário que o SynapSpeech propõe uma alternativa. Em vez de exames caros e restritivos, ele utiliza a fala espontânea como base para análise. Com o auxílio de técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM), o sistema busca identificar padrões linguísticos que possam sinalizar alterações cognitivas associadas aos estágios iniciais da doença. Ou seja, a partir de um celular comum, uma gravação, pode-se iniciar um processo de triagem. A proposta não propõe a substituição do profissional especializado, e sim, uma potencialização na descoberta do diagnóstico antecipado.

2. INTRODUÇÃO

A Doença de Alzheimer (DA) está entre as principais causas de demência no mundo (SAÚDE, 2024). No Brasil, seu impacto é especialmente evidente entre os idosos. O diagnóstico precoce pode fazer uma grande diferença na vida de quem convive com a condição, mas, infelizmente, esse cuidado ainda não chega para todos. Há muitos motivos: exames caros, serviços concentrados em poucos centros, escassez de profissionais fora das grandes cidades (Dr.Consulta, 2024). E, além disso, uma certa invisibilidade do problema, sobretudo em regiões periféricas e afastadas dos grandes centros urbanos (YANG et al., 2022).

Considerando isso, o presente projeto propõe uma solução chamada, SynapSpeech, que visa reduzir a dependência de estruturas clínicas sofisticadas para triagem, desenvolvendo uma tecnologia simples e acessível: a análise da fala espontânea como estratégia inicial de triagem. A proposta utiliza Inteligência Artificial (IA) para detectar, na forma de falar, sinais de possíveis alterações cognitivas onde

padrões linguísticos podem passar por uma análise computacional (FRASER; MELTZER; RUDZICZ, 2015). Ao adotar esse caminho, o projeto busca ampliar o acesso à saúde digital de forma inclusiva (TÓTH et al., 2018).

3. OBJETIVOS

Este estudo busca desenvolver uma solução tecnológica baseada em IA para auxiliar na triagem clínica e na tomada de decisão no diagnóstico precoce do Comprometimento Cognitivo Leve (MCI) e da Doença de Alzheimer. A ideia é criar um sistema que escute e interprete narrativas orais espontâneas, captando nuances linguísticas e sinais cognitivos que possam apontar para alterações neurológicas, ainda que imperceptíveis. Para isso, o projeto propõe o uso de técnicas de PLN e AM, fortalecendo a sensibilidade e a precisão na identificação precoce com menos margem para erros.

Mas não se trata somente de tecnologia: o projeto visa democratizar o acesso a diagnósticos especializados, especialmente onde faltam profissionais, valorizando uma abordagem que una inovação e empatia. Assim, a IA assume a função de aliada ética e humana para promover saúde, contribuindo diretamente para a qualidade de vida dos familiares e parentes.

4. METODOLOGIA

Para o desenvolvimento da IA proposta neste estudo, adotaram-se classificadores supervisionados baseados em Aprendizado de Máquina (AM), visando à identificação de padrões linguísticos presentes em narrativas orais espontâneas. A construção do modelo foi conduzida em ambiente **Visual Studio Code** com suporte a GPU, utilizando a linguagem **Python** e bibliotecas amplamente reconhecidas, como scikit-learn, pandas, nltk, gensim e transformers.

A base de dados utilizada faz parte do repositório DNLT-BP (Discurso Narrativo Longitudinal em Português Brasileiro), desenvolvido sob coordenação da Profa. Dra. Lilian Cristine Hubner e colaboradores, especialista em Linguística e

responsável pela organização científica desses corpora. O repositório reúne diferentes conjuntos de dados de narrativas orais: o *Cinderella* (60 amostras distribuídas equilibradamente entre as classes), o *Dog* (106 amostras, sendo a maioria pertencente ao grupo controle), o *Lucia* (89 registros, sendo 9 de Alzheimer e 80 controles) e o *Wallet* (70 amostras, incluindo 23 de MCI e 12 de indivíduos saudáveis). Esses conjuntos trazem transcrições de histórias narradas por participantes diagnosticados como saudáveis, MCI ou AD, constituindo a base para o treinamento e avaliação dos modelos propostos (NILC NLP, 2023).

Antes da etapa de modelagem, os dados textuais passaram por um processo cuidadoso de pré-processamento, envolvendo limpeza textual, lematização, remoção de *stopwords* e padronização, gerando versões com e sem exclusão de termos semanticamente irrelevantes. Para a representação linguística das transcrições, foram explorados diferentes tipos de *embeddings*, como *Bag-of-Words, TF-IDF, Word2Vec, FastText, GloVe, BERTimbau e MiniLM*, além da combinação de matrizes vetoriais por meio de técnicas de fusão. A etapa de classificação foi estruturada em dois níveis hierárquicos: inicialmente, o modelo distingue entre indivíduos saudáveis e não saudáveis (MCI + AD); em seguida, nos casos positivos, procede-se à diferenciação entre MCI e Alzheimer.

Entre os algoritmos utilizados, destacam-se *CatBoost, LightGBM e XGBoost*, aplicados com validação cruzada estratificada e avaliados com métricas como acurácia, F1-score, precisão, revocação e acurácia balanceada. Para lidar com o desbalanceamento entre as classes, foram empregadas técnicas como *SMOTE, ADASYN, TomekLinks e SMOTETomek*. Como desdobramento aplicado da pesquisa, está em andamento a prototipagem de uma interface de triagem acessível, com potencial de integração a assistentes virtuais como o Telegram. Essa estratégia foi pensada para ampliar o alcance diagnóstico em comunidades com baixo acesso a serviços clínicos especializados, reforçando o papel da tecnologia como instrumento de inclusão e cuidado em saúde.

5. DESENVOLVIMENTO

O projeto encontra-se na etapa de prototipagem, dedicada ao desenvolvimento e à validação das abordagens propostas para a classificação dos distúrbios cognitivos com base na fala espontânea. A equipe tem explorado diferentes caminhos metodológicos, combinando diversas representações textuais, visando ajustar a melhor configuração para cada fase do diagnóstico.

Na primeira, o sistema diferencia indivíduos saudáveis daqueles que apresentam algum grau de comprometimento; na segunda, busca distinguir entre casos de MCI e AD. Os testes são conduzidos com rigor estatístico, utilizando validação cruzada estratificada e análise de diversas métricas para garantir resultados confiáveis.

Além disso, o projeto mantém um diálogo constante com a produção científica recente e se apoia em estudos de referência nas áreas de neurociência computacional e IA aplicada à saúde, reforçando a consistência dos achados. Ética e transparência também orientam cada etapa do processo, com especial atenção à mitigação de vieses e à clareza das decisões algorítmicas.

6. RESULTADOS PRELIMINARES

Entre os melhores resultados dos experimentos, foi selecionada uma combinação de representações vetoriais com o modelo LightGBM e a técnica de balanceamento ADASYN. Essa configuração alcançou um macro F1-score de 85%, demonstrando equilíbrio entre precisão e revocação nas classes avaliadas. A acurácia geral foi de 86%, com precisão de 88% para indivíduos saudáveis e 82% para indivíduos não saudáveis. Em termos de recall, os valores foram de 90% para saudáveis e 79% para não saudáveis, evidenciando a elevada sensibilidade do modelo para identificar corretamente os casos sem comprometimento cognitivo.

Esses resultados indicam que o modelo tende a apresentar maior confiança ao classificar indivíduos como saudáveis, o que pode ser vantajoso no contexto de triagens, desde que complementado por uma segunda etapa analítica mais sensível aos casos de risco. O desempenho observado também reforça a viabilidade da

proposta como uma solução acessível, precisa e de fácil integração em fluxos de triagem digital voltados à saúde cognitiva.

7. FONTES CONSULTADAS

NILC NLP. DNLT-BP: **Deep Neural Language Tool for Brazilian Portuguese**. 2023. [Online; acessado em 24-fev-2025]. Disponível em: <https://github.com/nilc-nlp/DNLT-BP>.

SAÚDE, M. da. **Relatório Nacional sobre a Demência: Epidemiologia, (re)conhecimento e projeções futuras**. 2024. Disponível online em: https://www.gov.br/saude/.

FRASER, K. C.; MELTZER, J. A.; RUDZICZ, F. Linguistic features identify alzheimer's disease in narrative speech. Journal of Alzheimer's disease, SAGE Publications Sage UK: London, England, v. 49, n. 2, p. 407–422, 2015.

Dr. Consulta. Doença de Alzheimer: sintomas, causas e tratamento. 2024. Acesso em: 27 maio 2025. Disponível em: https://www.drconsulta.com/artigos/doenca-de-alzheimer-sintomas-causas-e-tratamento.

YANG, Q. et al. Deep learning-based speech analysis for alzheimer's disease detection: a literature review. Alzheimer's Research & Therapy, Springer, v. 14, n. 1, p. 186, 2022.

TÓTH, L. et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Current Alzheimer Research, Bentham Science Publishers direct, v. 15, n. 2, p. 130–138, 2018.