Avanços no tratamento de dados textuais na saúde com técnicas de Inteligência Artificial: Um algoritmo para agrupamento de dados

Alisson I. Dias[1], **Denise S. de Sousa**[1],
Josimar A. de Oliveira[1], **Larissa G. Cardoso**[1]
Sara L. de Farias[2], **Alan R. dos Santos**[1],
Elton C. S. Morais[1]

[1]Instituto Acadêmico de Ciências Tecnológicas – Universidade Estadual de Goiás (UEG)
Unidade Universitária de Ceres – 76.300-000 – Ceres, GO – Brazil

[2]Instituto Federal Goiano (IFGoiano) – Campus Ceres
GO-154, km 218 - Zona Rural, Ceres – 76.300-000, GO – Brazil

{alissonivo, denise.sousa, josimar, larissa.320}@aluno.ueg.br

fariassara012@gmail.com, alan.rib.san@gmail.com, eltoncsmorais@ueg.br

Abstract. *The advance of Information Technology (IT) in healthcare has generated a large volume of data, often without adequate processing. In view of this, Artificial Intelligence (AI) helps to harness this data, but dealing with free and heterogeneous clinical texts is still challenging. This study developed a Python algorithm for pre-processing and clustering 217,000 clinical diagnoses by structural similarities, focusing on terms related to Dengue and COVID-19. Consequently, preliminary results show that this approach effectively organizes the data, facilitating further analysis. Despite the initial success, challenges such as the configuration of terms and the heterogeneity of the texts indicate the need for improvements to improve the accuracy of the process.*

Resumo. *O avanço da Tecnologia da Informação (TI) na saúde gerou grande volume de dados, muitas vezes sem processamento adequado. À vista disso, a Inteligência Artificial (IA) ajuda no aproveitamento desses dados, mas lidar com textos clínicos livres e heterogêneos ainda é desafiador. Este presente estudo desenvolveu um algoritmo em Python para o pré-processamento e agrupamento de 217 mil diagnósticos clínicos por similaridades estruturais, com foco em termos relacionados à Dengue e COVID-19. Consequentemente, resultados preliminares mostram que essa abordagem organiza de forma eficaz os dados, facilitando análises posteriores. Apesar do sucesso inicial, desafios como a configuração de termos e a heterogeneidade dos textos indicam a necessidade de aprimoramentos para melhorar a precisão do processo.*

1. Introdução

A era da informação trouxe um desenvolvimento rápido na utilização de Tecnologia da Informação (TI) na saúde. Diversas ferramentas tecnológicas são utilizadas e consequentemente, produzem uma quantidade massiva de dados [Dou and Meng 2023]. Esses dados geralmente são vídeos, imagens, áudio e texto [Haraty et al. 2015], promovendo uma diversidade de dados distintos, que muitas vezes são apenas armazenados.

Com o avanço da Inteligência Artificial (IA), esses dados começaram a serem mais bem aproveitados através de Aprendizados de Máquinas (AM) baseados em imagens [Godinho et al. 2019, Napravnik et al. 2024], textos [Dobrakowski et al. 2021], dentre outros. Entretanto, o volume de dados é extremamente grande e cresce a uma velocidade impressionante [Concepcion et al. 2024], o que torna difícil tratá-los de forma adequada, uma vez que estão desorganizados e mistos [Singh et al. 2019].

Olhando especificamente para os dados textuais, torna-se ainda mais difícil esse tratamento. Usualmente esses registros textuais são livres, ou seja, são carregados de termos específicos da área da saúde, assim como também das particularidades individuais de escrita daqueles que o fazem. Devido a essa heterogeneidade, volume e complexidade [Haraty et al. 2015], torna-se extremamente difícil o processamento desses dados [Dobrakowski et al. 2021, Siouda et al. 2024], tornando esse processo oneroso [Ghaddar and Naoum-Sawaya 2018] e muitas vezes ineficiente. Apesar de existirem diversas técnicas para lidarem com tais situações, como algoritmo híbrido, lógica fuzzy, rede neural e técnicas de agrupamentos [Thangarasu and Dominic 2015], ainda assim o volume de dados é extremamente grande, considerando que esses não tiveram nenhum tipo de tratamento prévio,ou seja, não passaram por etapas iniciais de padronização, limpeza ou organização que poderiam torná-los mais estruturados e homogêneos.

Dados poluídos e errôneos [Tripathi et al. 2023] podem ser caros e difíceis de serem processados. Logo, para o tratamento inicial desses dados, foi desenvolvido um algoritmo capaz de agrupar diagnósticos clínicos por similaridades. Trata-se da identificação de padrões de estruturas em um conjunto dados [Waqas et al. 2022], permitindo assim agrupar diagnósticos com os mesmos padrões estruturais. O desenvolvimento desse algoritmo capaz de realizar esse pré-processamento de dados, promove uma organização inicial e essencial, estruturando as informações de forma mais consistente antes de uma análise mais aprofundada. Para valida sua aplicação, foram utilizados termos relacionados as duas doenças de alta relevância no Brasil: Dengue e Covid-19 [Paula et al. 2023], dada sua alta incidência e impacto significativo na saúde pública. Essa abordagem visa facilitar o manejo e a extração de informações úteis a partir de dados inicialmente desorganizados e heterogêneos. Para o seu funcionamento, o algoritmo recebe termos clínicos previamente formatados, que são utilizados para classificar e agrupar os diagnósticos conforme suas características em comum.

2. Métodos

A análise de dados, principalmente os não tratados, podem ocasionar perdas no processamento ou até mesmo enviesar resultados. Nesse sentido, este estudo desenvolveu um algoritmo utilizando Python, e suas principais bibliotecas para análise de tratamentos de dados (Figura 1), como pandas, para o pré-processamento e agrupamento de dados textuais de diagnósticos clínicos, visando uma organização inicial dos dados antes de uma análise detalhada.

O algoritmo, integrado a uma aplicação desenvolvida com o framework Django, recebe termos clínicos previamente categorizados, faz seu processamento e agrupa todos os diagnósticos clínicos que contenham tais termos. A aplicação utiliza um modelo que armazena cada registro clínico com campos para o termo clínico e o texto do diagnóstico. Para os testes preliminares, foi utilizada uma base de dados contendo 217 mil registros,

```python
import django_filters
from django.db.models import Q
form .models import Patient

class PatientFilter(django_filters.FilterSert):
    diagnostic = django_filters.CharFilter(method='filter_by_keywords')

    class Meta:
        model = Patient
        fields = []

    def filter_by_keywords(self, queryset, name, value):
        keywords = value.split(",")
        query = Q()
        for keyword in keywords:
            query |= Q(diagnostic__icontains=keyword)
        return queryset.filter(query)
```

Figura 1. Parte do código algoritmo

todos em seu estado bruto, sem nenhum tipo de pré-tratamento.

Essa base de dados é previamente carregada em um modelo relacional, permitindo que aplicação, por meio do algoritmo de agrupamento, acesse e analise os dados com base nos termos forncedidos. O resultado do processo é consolidado e transportado para um arquivo CSV (*Comma Separated Values*).

3. Resultados e Discussões

Os resultados iniciais são promissores. O uso do algoritmo de agrupamento para pré-processamento de dados textuais clínico se mostrou eficiente na organização preliminar, o que possibilita melhor análise subsequente desses dados. Nos testes iniciais, o algoritmo conseguiu identificar padrões textuais com base em estruturas ou termos relacionados à diagnósticos clínicos de Dengue e Covid-19. Esse agrupamento permite redução significativa do volume de dados, possibilitando assim um melhor tratamento desses dados para um processamento posterior, como o treinamento de inteligência artificial.

Contudo, apesar dos resultados iniciais serem satisfatórios, o estudo identificou limitações importantes que sugerem aprimoramentos futuros. Dentre as principais limitações identificada foi a configuração dos termos iniciais de busca, considerando que precisa ser abrangente o suficiente para alcançar o máximo de resultados, no entanto, precisa ser coerente e conciso com o resultado esperado. Aliado a isso, por ser textos clínicos livres, por consequência heterogêneos, deve-se levar em consideração, ambuiguidade nos textos, erros ortográficos e abreviações inconsistentes.

Embora o método tenha logrado êxito em sua proposta preliminar, melhorias no processo são necessárias para aumentar a precisão e escalabilidade no processo. A utilização de termos técnicos, jargões, abreviações deverão ser catalogadas previamente através de profissionais da área da saúde, que estão familiarizados com tais possibilidades.

4. Conclusão

A construção de um algoritmo para agrupar dados textuais clínicos representa um avanço significativo na organização inicial desses dados, especialmente diante do aumento dos registros na área da saúde. A aplicação do código permitiu agrupar diagnósticos clínicos com base em termos comuns, resultando na redução de uma quantidade considerável de dados desorganizados. No entanto, a complexidade dos textos clínicos, com suas variações terminológicas e erros, destaca a necessidade de ajustes contínuos.

Para trabalhos futuros, é essencial estabelecer parcerias com profissionais da área da saúde, visando uma maior padronização dos dados. Além disso, este estudo contri-

bui para a otimização do uso de dados textuais na saúde, enfatizando a importância de estratégias eficazes de pré-processamento em grandes volumes de dados.

Referências

Concepcion, M. B. S., Gerardo, B. D., Elijorde, F. I., Castro, J. T. D., and Cruz, N. B. D. (2024). Development of big data classifier for biomedicine early diagnosis: An experimental approach using machine learning methods. *Journal of Computer Science*, 20:379–388.

Dobrakowski, A. G., Mykowiecka, A., Marciniak, M., Jaworski, W., and Biecek, P. (2021). Interpretable segmentation of medical free-text records based on word embeddings. *Journal of Intelligent Information Systems*, 57:447–465.

Dou, Y. and Meng, W. (2023). Comparative analysis of weka-based classification algorithms on medical diagnosis datasets. *Technology and health care : official journal of the European Society for Engineering and Medicine*, 31:397–408.

Ghaddar, B. and Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, 265:993–1004.

Godinho, T. M., Lebre, R., Almeida, J. R., and Costa, C. (2019). Etl framework for real-time business intelligence over medical imaging repositories. *Journal of Digital Imaging*, 32:870–879.

Haraty, R. A., Dimishkieh, M., and Masud, M. (2015). An enhanced k-means clustering algorithm for pattern discovery in healthcare data. *International Journal of Distributed Sensor Networks*, 2015.

Napravnik, M., Hržić, F., Tschauner, S., and Štajduhar, I. (2024). Building radiologynet: an unsupervised approach to annotating a large-scale multimodal medical database. *BioData Mining*, 17.

Paula, F. D. A. P., Ferreira, J. Z., Júnior, E. L. D. S., Alves, I. G., Narvaes, J. V. R., Paula, C. D. A. P., Baretta, I. P., and Pacheco, R. B. (2023). Incidência da dengue durante a covid-19.

Singh, P., Singh, S. P., and Singh, D. S. (2019). An introduction and review on machine learning applications in medicine and healthcare.

Siouda, R., Nemissi, M., and Seridi, H. (2024). Diverse activation functions based-hybrid rbf-elm neural network for medical classification. *Evolutionary Intelligence*, 17:829–845.

Thangarasu, G. and Dominic, P. D. D. (2015). Diabetic deduction through non-parametric analysis. *International Journal of Business Information Systems*, 20:325–347.

Tripathi, M. A., Tripathi, R., Effendy, F., Manoharan, G., Paul, M. J., and Aarif, M. (2023). An in-depth analysis of the role that ml and big data play in driving digital marketing's paradigm shift.

Waqas, S. M., Hussain, K., Mostafa, S. A., Nawi, N. M., and Khan, S. (2022). Fuzzy density-based clustering for medical diagnosis. volume 457 LNNS, pages 264–271.