CLASSIFICAÇÃO AUTOMÁTICA BI-RADS DE REGISTROS MÉDICOS DE RESSONÂNCIA MAGNÉTICA DE MAMA USANDO MODELOS BASEADOS EM ARQUITETURA TRANSFORMERS PARA O PORTUGUÊS DO BRASIL

Ricardo Gomes de Oliveira[1]; Bruno Leonardo Santos Menezes [2]; Júnia Ortiz [2]; Erick Giovani Sperandio Nascimento [2]
[1]Pós-graduando; TCC (Trabalho de Conclusão de Curso); ricardo.oliveira@aln.senaicimatec.edu.br
[2] Centro Universitário SENAI CIMATEC; Salvador-BA; bruno.menezes@fieb.org.br
[2] Centro Universitário SENAI CIMATEC; Salvador-BA; junia.matos@fieb.org.br
[2] Centro Universitário SENAI CIMATEC; Salvador-BA; erick.sperandio@fieb.org

RESUMO

Este estudo tem como objetivo apresentar um modelo de classificação para categorização de registros clínicos textuais de ressonância magnética de mama, baseado na análise lexical, sintática e semântica de relatórios clínicos de acordo com a classificação do *Breast Image Data and Reports System* (BI-RADS) [1], usando *deep learning* e Processamento de Linguagem Natural (PNL). O modelo foi desenvolvido a partir de *transfer learning* baseado no modelo BERTimbau [2] pré-treinado, modelo BERT (*Bidirectional Encoder Representations from Transformers*) [3] treinado em português do Brasil. O conjunto de dados consiste em laudos médicos em português do Brasil classificados em seis categorias: Inconclusivo; Normal ou Negativo; Achados Certamente Benignos; Achados provavelmente benignos; Achados Suspeitos; Alto risco de câncer; Lesão maligna previamente conhecida. Os seguintes modelos foram implementados e comparados: *Random Forest, SVM, Naïve Bayes*, BERTimbau com e sem ajuste fino. O modelo BERTimbau apresentou melhores resultados, com melhor desempenho após o ajuste fino.

PALAVRAS-CHAVE: BI-RADS, Deep Learning, Transformers, BERTimbau.

1. INTRODUÇÃO

Na área da saúde, há muitos dados importantes nos prontuários médicos, como laudos e resultados de exames. A correta classificação desses dados pode ajudar os profissionais da saúde a gerenciar as informações dos pacientes de maneira mais eficiente. Este trabalho apresenta um sistema de classificação para categorização do BI-RADS [1], baseado na análise lexical, sintática e semântica de documentos, derivados de registros clínicos textuais. Esse sistema foi desenvolvido com o uso de tecnologias como Deep Learning e NLP (Processamento de Linguagem Natural). Foi utilizado o modelo BERTimbau para classificar as categorias do BI-RADS e compará-lo com outros modelos de aprendizado de máquina como Random Forest, SVM, Naïve Bayes, BERTimbau com e sem finetuning. Após o treinamento do modelo com milhares de registros médicos, um novo modelo especialista foi criado para classificar os relatórios de ressonância magnética de mama.

2. METODOLOGIA

O BI-RADS é uma ferramenta usada na mamografia para avaliar o risco de uma paciente desenvolver câncer de mama. Ele padroniza relatórios clínicos e tem seis categorias de classificação: Categoria 0 é inconclusiva, Categoria 1 é normal ou negativa, Categoria 2 são achados certamente benignos, Categoria 3 são achados provavelmente benignos, Categoria 4 são achados suspeitos e Categoria 5 são de alto risco de câncer. Cada categoria indica um risco crescente de malignidade, sendo que a Categoria 5 indica um risco muito alto de câncer e a Categoria 6 é usada quando o diagnóstico de câncer já foi confirmado.

Foram coletados dados de 8.813 relatórios de ressonâncias magnéticas de mama, emitidos por um serviço de radiologia. A maioria dos relatórios (83,51%) foi de exames bilaterais e o menor número (0,52%) foi de exames de mamotomia. Esses relatórios possuem uma classificação chamada BI-RADS, que indica o risco de câncer de mama. Para facilitar a análise dos dados, uma nova variável foi criada para representar essa classificação de forma numérica. A média de palavras encontradas nos relatórios foi de 202, e as informações foram extraídas funções e métodos específicos para localização de palavras em sentenças.

BI-RADS por Categoria:

- 0: 10 instancias;
- 1: 586 instancias;
- 2: 3.387 instancias;
- 3: 1.115 instancias;
- 4: 971 instancias;
- 41: 280 instancias;
- 43: 9 instancias;
- 5: 352 instancias;

- 6: 723 instancias.

O estudo utilizou dados de exames de ressonância magnética de mama para diagnóstico do câncer. Os dados foram ajustados para que as categorias estivessem equilibradas e foram processados por diferentes algoritmos de aprendizado de máquina, incluindo BERT, Random Forest [4], SVM [5] e Naïve Bayes [6]. BERT é um algoritmo que aprende representações precisas de palavras em um corpus de texto, considerando tanto o contexto esquerdo como o direito de cada palavra. Random Forest é um algoritmo popular de aprendizado de máquina que resolve problemas complexos combinando vários classificadores. SVM é um algoritmo popular de aprendizado de máquina usado principalmente para classificação. Naïve Bayes é um algoritmo de aprendizado supervisionado que utiliza o teorema de Bayes para resolver problemas de classificação, principalmente na classificação de texto de alta dimensão. A título de verificação de desempenho dos resultados encontrados, foram utilizadas as métricas *precision* (que avalia a taxa de verdadeiros positivos em relação aos falsos positivos), acurácia (quantidade de acertos dividida pelo total de uma amostra), *recall* (que avalia a taxa de verdadeiros positivos em relação aos falsos negativos), e F1-Score (que realiza a média harmônica entre *precision* e *recall* a fim de trazer um número único que determine a qualidade geral de um modelo). O intervalo das métricas varia entre 0 a 1, e quanto mais próximo de 1, melhor o desempenho do modelo para aquela métrica.

3. RESULTADOS E DISCUSSÃO

O conjunto de dados original foi usado em diferentes algoritmos de aprendizado de máquina, como Random Forest, SVM e Naïve Bayes, além de um algoritmo de aprendizado profundo chamado BERTimbau. As métricas usadas para avaliar o desempenho dos modelos foram precisão [8], recall [8], F1-score [8] e acurácia [8]. Para o algoritmo Random Forest, técnicas como Randomized Search Cross Validation e Grid Search Cross Validation foram aplicadas para encontrar os melhores hiperparâmetros. Para o SVM, foi usada a técnica Randomized Search Cross Validation para encontrar os melhores hiperparâmetros. A Tabela 1 resume os resultados dos modelos de aprendizado de máquina, mostrando que o algoritmo Random Forest apresentou o melhor resultado.

Resultados dos Testes para Algoritmos de Machine Learning		
Modelo	F1-Score	Acurácia
Random Forest	0,787	0,876
Naive Bayes	0,556	0,753
SVM	0,519	0,674

Tabela 1 - Resumo dos resultados dos modelos de aprendizado de máquina.

O conjunto de dados foi analisado por quatro algoritmos de aprendizado de máquina: Random Forest, SVM, Naïve Bayes e BERTimbau. Para aplicar o BERTimbau, as variáveis categóricas foram transformadas em binárias usando a técnica One-Hot Encoding. Foram realizados testes específicos com e sem ajuste fino, usando o otimizador AdamW com parâmetros personalizados. Com o ajuste fino, 1.819 novos tokens foram adicionados e um novo modelo foi criado após quatro épocas de treinamento, com uma perplexidade de 2,17. A perplexidade é uma medida de quão bem um modelo prevê uma amostra. O modelo BERTimbau teve melhores resultados do que os algoritmos de aprendizado de máquina. A Tabela 2 apresenta os valores comparativos das etapas do modelo BERTimbau.

	1 - MODELO BERTimbau ORIGINAL BERTimbau Tokenizer and Model			2 - MODELO PÓS AJUSTE FINO BIRADS Tokenizer and Model			Qtd. de Amostras
	precision	recall	f1-score	precision	recall	f1-score	
birads0	0	0	0	0	0	0	10
birads1	1	0,97	0,98	1	**0,98**	**0,99**	586
birads2	0,99	0,98	0,99	**1**	**0,99**	0,99	3387
birads3	0,95	0,97	0,96	**0,98**	**0,99**	**0,98**	1115
birads4	0,86	0,95	0,9	**0,95**	**0,98**	**0,97**	971
birads4a	1	0,7	0,83	0,95	**0,96**	**0,96**	280
birads4c	0	0	0	0	0	0	9
birads5	0,51	0,79	0,62	**0,95**	**0,8**	**0,87**	352
birads6	0,93	0,55	0,7	0,93	**0,96**	**0,95**	723
micro avg	0,93	0,91	0,92	**0,98**	**0,97**	**0,98**	7433
macro avg	0,69	0,66	0,66	**0,75**	**0,74**	**0,75**	7433
weighted avg	0,94	0,91	0,92	**0,97**	**0,97**	**0,97**	7433
samples avg	0,91	0,91	0,91	**0,97**	**0,97**	**0,97**	7433

Tabela 2 - Comparativo entre o BERTimbau e o BERTimbau ajustado. O melhor valor para cada métrica está evidenciado em negrito.

Observando os valores apresentados na Tabela 2, percebe-se claramente que na grande maioria das situações em que as classes estiveram presentes, o desempenho do modelo ajustado foi superior a todos os modelos testados anteriormente.

4. CONSIDERAÇÕES FINAIS

A arquitetura *Transformer* apresenta-se como a dominante para o processamento de linguagem natural, com modelos pré-treinados facilmente adaptáveis a diferentes tarefas. É indicado para o uso de aprendizado de máquina em tarefas de classificação textos em um conjunto inicial de dados rotulados, sendo o BERT um avanço significativo no processamento de linguagem natural, especialmente para o português. No caso deste estudo, o BERTimbau ajustado conseguiu captar informações específicas para uma área generalista, aumentando seu vocabulário e tornando-se um bom modelo para classificar dados de textos médicos, estruturando dados normalmente não estruturados.

5. REFERÊNCIAS

1. CASTRO, Sergio M. TSEYTLIN, Eugene. MEDVEDEVA, Olga. MITCHELL, Kevin. VISWESWARAN, Shyam . BEKHUIS, Tanja. JACOBSON, Rebecca S. Automated annotation and classification of BI-RADS assessment from radiology reports, Journal of Biomedical Informatics, April 2017.
2. SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear). [S.l.: s.n.], 2020.
3. DEEP LEARNING BOOK, disponível em https://www.deeplearningbook.com.br/o-que-e-bert-bidirectional-encoder-representations-from-transformers/
4. HO, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International 9 Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
5. CORTES, Corina ; VAPNIK, Vladimir (1995). "Support-Vector Networks". Machine Learning. 20 (3): 273–297. doi:10.1007/BF00994018.
6. MCCALLUM, Andrew. "Graphical Models, Lecture2: Bayesian Network Representation" (PDF). Retrieved 22 October 2019.
7. DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Disponível em https://arxiv.org/abs/1810.04805.
8. PÁDUA, Mateus. Machine Learning -Métricas de avaliação: Acurácia, Precisão e Recall, F1-score. Disponível em https://medium.com/@mateuspdua/machine-learning-m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-e-recall-d44c72307959.