

## **TREINAMENTO DO MODELO ROBERTA PARA ASSOCIAÇÃO DE ESQUEMAS DE DADOS GEOESPACIAIS: ESTUDOS PRELIMINARES DO USO DE *LARGE LANGUAGE MODEL* EM GEOSEMÂNTICA**

FABÍOLA ANDRADE SOUZA<sup>1,2</sup>

HIDEO ARAKI<sup>2</sup>

SILVANA PHILIPPI CAMBOIM<sup>2</sup>

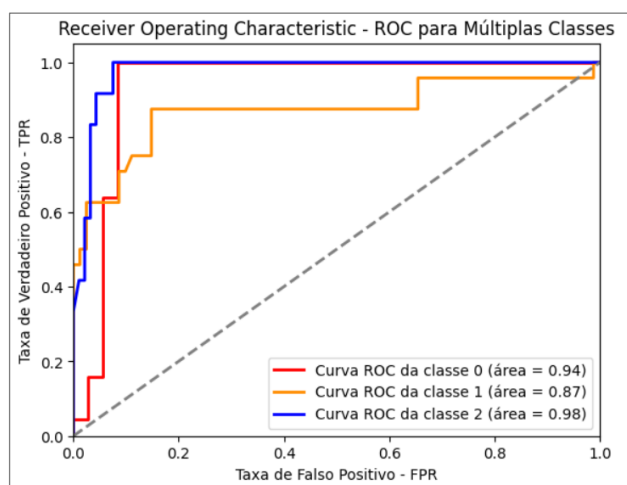
<sup>1</sup>Universidade Federal da Bahia (UFBA) – fabiola.andrade@ufba.br

<sup>2</sup>Universidade Federal do Paraná (UFPR) – {haraki@ufpr.br; silvanacamboim@ufpr.br}

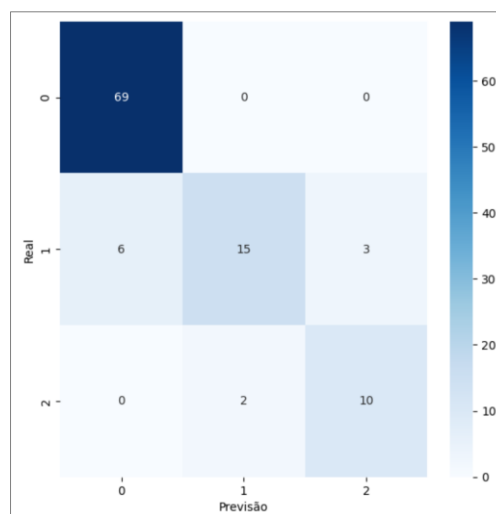
A aplicação de tecnologia baseada em Inteligência Artificial (IA), recentemente, trouxe uma revolução para o comportamento social e a percepção e determinação de novas frentes de atuação no mercado de trabalho, ou mesmo a redefinição de processos para quem já atuava na área [1]. No tocante à IA geoespacial (*Geospatial Artificial Intelligence* - GeoAI), pouco foi explorado até o momento e existe uma ampla gama de questões a serem discutidas, tanto na definição de novas metodologias de atuação quanto para abordagens que considerem as múltiplas modalidades de dados geoespaciais e suas aplicações [2-3]. Entre os diversos usos da IA, destaca-se a interpretação da linguagem humana no contexto do aprendizado de máquina (*machine learning*), onde o Processamento de Linguagem Natural (PLN) tem expandido suas possibilidades de emprego. Em PLN, os algoritmos utilizados identificam padrões e elaboram respostas a partir do uso de modelagem linguística. A modelagem clássica aprende a distribuição de probabilidade sobre sequências de texto, interpretando-o sintática e semanticamente, ao aplicar modelos matemáticos fundamentais sobre as bases de dados de treinamento pré-existentes, para posteriormente desenvolver novas classificações [3-4]. Os mais recentes modelos matemáticos aplicados são baseados em redes neurais, sendo que a arquitetura *Transformers*, apresentada por [5], revolucionou a forma de analisar um texto, reduzindo a dificuldade dos modelos anteriores em lidar com conexão entre termos distantes, além da redução do tempo de processamento devido a seu paralelismo. Esta arquitetura tem sido aplicada na geração de grandes modelos de linguagem (*Large Language Model* - LLM), permitindo processar volumes expressivos de dados, a exemplo dos modelos BERT (*Bidirectional Encoder Representations from Transformers*) e GPT (*Generative Pre-Trained Transformer*), com aplicações práticas variadas [6-7]. Entretanto, até o momento, não foram identificados trabalhos específicos que tratem da realização de alinhamento semântico de modelagens conceituais de dados geoespaciais utilizando LLM. O alinhamento semântico é importante para garantir uma utilização integrada de bases de dados geoespaciais que podem ter sido geradas por produtores distintos, utilizando diferentes hierarquias conceituais para os dados [8-9]. Contudo, estudos para automatizar este processo são recentes [10-12] e não consideram GeoAI. Neste contexto, este resumo apresenta estudos preliminares para o treinamento de um LLM a partir de alinhamento semântico efetivado por humanos entre dois esquemas conceituais de dados geoespaciais para, posteriormente, avaliar a capacidade do modelo em distinguir a adequação de novas associações semânticas. O LLM escolhido como método de pré-processamento para o treinamento foi o RoBERTa (*Robustly Optimized BERT Pre-Training Approach*), que apresenta vantagens de ganho no processamento em relação a outros modelos fundamentais [13]. Seu uso ocorreu através da linguagem de programação *Python*, em ambiente *Google Colab*, com aplicação da função *RobertaTokenizer* da biblioteca *Transformers*. Após o treinamento do modelo, a avaliação de desempenho do mesmo foi efetuada com o cálculo de alguns parâmetros, como a curva ROC (*Receiver Operating Characteristic*) e a matriz de confusão [14-15]. Os esquemas conceituais utilizados foram a Especificação Técnica para Estruturação de Dados Geoespaciais Vetoriais - ET-EDGV [16] e o *OpenStreetMap* – OSM [17], uma vez que já havia alinhamento semântico prévio destes modelos realizado por [12], o que pode ser utilizado como referência no treinamento. Após o treinamento e a validação de desempenho, a classificação de novas associações incluiu a avaliação da confiabilidade e suas probabilidades com aplicação das funções de *logits* e *softmax* [18]. O treinamento utilizou um recorte da ET-EDGV para as classes: Edificação Agropecuária de Extrativismo Vegetal e/ou Pesca, de Comércio ou Serviços, Construção Turística, Religiosa, Construção Aeroportuária, Construção Portuária, Construção de Lazer, Metroferroviária, de Polícia, Pública Civil e Rodoviária. Em cada classe, apenas o domínio do atributo que classifica o tipo de edificação foi aplicado, sendo associado em uma planilha ao *value* da *tag* (par *key+value*) do OSM, conforme correspondências propostas por [12]. Cada associação sugerida pelas autoras foi rotulada como ‘1- Adequada’, a exemplo de ‘Banco – *Atm*’ ou ‘Igreja -

Church'. Depois, todas as combinações entre as opções disponíveis que não foram explicitadas pelas autoras receberam rótulo de '0 - Inadequada' (casos não aplicáveis como 'Curral - Bank' ou 'Mesquita - cinema') ou de '2- Parcialmente adequada' (situações que tem similaridade pelo tema, mas em outro nível hierárquico de conhecimento, como 'Sinagoga - Cathedral' ou 'Universidade - Kindergarten'). Houve 351 associações para o treinamento, sendo que 74,6% destas com rótulo '0 - Inadequada', motivo pelo qual a base de treinamento foi separada na proporção 70-30% em relação à validação. Como resultado, após o treinamento desta base com o modelo RoBERTa, os parâmetros de avaliação de desempenho mostraram uma acurácia de 0.895 e precisão de 0.894, para seis épocas de execução. Em relação à curva ROC (Figura 1a), a rotulação '2 - Parcialmente adequada' teve o melhor desempenho ( $AUC = 0.98$ ), seguida da '0 - Inadequada' ( $AUC = 0.94$ ) e deixando os dados com semântica adequadamente associada com o menor desempenho, embora com uma  $AUC = 0.87$  que define um resultado de classificação melhor que o acaso. Já a matriz de confusão (Figura 1b), em sua linha diagonal, mostra que a maioria dos dados foram classificados corretamente. Os equívocos foram seis associações adequadas que apareceram como inadequadas; três associações adequadas apareceram como parcialmente adequadas; e duas associações parcialmente adequadas classificadas como adequadas. Nenhuma associação inadequada foi equivocadamente correlacionada com as demais rotulações. Neste escopo, considerou-se os resultados do treinamento como coerentes e efetuou-se testes para realização de novas associações semânticas sobre o modelo treinado, aplicando o mesmo tokenizador sobre elas. Dentre as novas associações semânticas avaliadas, embora seis das quatorze sugeridas (43%) tenham apresentado respostas diferentes do esperado, os resultados foram considerados satisfatórios para testes preliminares, diante das limitações do treinamento e da avaliação de desempenho efetuada. Observou-se que as associações incoerentes fizeram confusão com rótulos mais próximos. Por exemplo, 'polícia - courthouse' e 'escola - university' foram associações consideradas adequadas, onde esperava-se parcialmente adequadas. Enquanto 'hotel - love\_hotel', 'banco - atm', 'banco - bank' e 'filmes - cinema' foram consideradas inadequadas, embora sejam adequadas. Nenhuma associação inadequada foi classificada como adequada ou parcial, alinhando-se às expectativas dos resultados das análises pela curva ROC e matriz de confusão, sobre os limiares de erro. Portanto, mesmo com uma taxa de erro alta, este estudo inicial para o treinamento do modelo RoBERTa para associação semântica entre dois esquemas conceituais de dados geoespaciais alcançou o objetivo principal de avaliar a capacidade do modelo aprender a realizar associações. Alguns fatores que podem ter contribuído no desempenho do modelo são: (i) pequena quantidade de dados de treinamento e maior rotulação dos dados considerados inadequados; (ii) estruturação dos dados de entrada, associando domínio da ET-EDGV e *value* do OSM, que pode não ter sido uma estratégia interessante para entendimento do contexto de uso dos termos; (iii) uso de hiper parâmetros mínimos no treinamento, como quantidade de épocas e taxa de aprendizado, por conta das limitações do ambiente computacional. Ainda há lacunas que precisam evoluir em estudos futuros, principalmente pela tecnologia envolvida ser recente e inovadora. No âmbito de uso de PLN e LLM para alinhamento semântico entre dados geoespaciais, os resultados são promissores e cabe ampliar as análises tanto em relação à melhoria dos parâmetros e dados de entrada (a exemplo de ontologias), quanto na análise de outros LLM, permitindo comparação de seu comportamento e determinação das vantagens e desvantagens de cada um.

Figura 1: Avaliação do Desempenho do Treinamento do modelo RoBERTa.



(a) Curva ROC



(b) Matriz de Confusão

Fonte: Autores (2024).

**Palavras-chaves:** Processamento de linguagem natural; *Large Language Model*; *Transformers*; mapeamento topográfico; *OpenStreetMap*.

## Referências

- [1] MEIRA, S. **Estamos na era da pedra lascada da IA, mas o futuro chega em 800 dias.** Entrevista. Brazil Journal. 16/03/2024. Disponível em <https://braziljournal.com/silvio-meira-estamos-na-era-da-pedra-lascada-da-ia-mas-o-futuro-chega-em-800-dias/>
- [2] HU, Y., GOODCHILD, M., ZHU, A.-X., YUAN, M., AYDIN, O., BHADURI, B., GAO, S., LI, W., LUNGA, D., & NEWSAM, S. (2024). **A five-year milestone: reflections on advances and limitations in GeoAI research.** Annals of GIS, 30(1), 1-14. DOI: 10.1080/19475683.2024.2309866
- [3] MAI, G., HUANG, W., SUN, J., SONG, S., MISHRA, D., LIU, N., GAO, S., LIU, T., CONG, G., HU, Y., CUNDY, C., LI, Z., ZHU, R., & LAO, N. (2024). On the Opportunities and Challenges of Foundation Models for GeoAI. **ACM Transactions on Spatial Algorithms and Systems**, 10(2), 46 pages. DOI: 10.1145/3653070
- [4] JOZEFOWICZ, R.; VINYALS, O.; SCHUSTER, M.; SHAZEER, N.; WU, Y. **Exploring the Limits of Language Modeling.** 11. fev. 2016. arXiv. Disponível em: <<http://arxiv.org/abs/1602.02410>>. Acesso em: 21/04/2023.
- [5] VASWANI, A.; SHAZEER, N.; PARMAR, N.; et al. **Attention is All you Need.** 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 2017.
- [6] ZHANG, Q.; ZHANG, T.; ZHAI, J.; et al. **A Critical Review of Large Language Model on Software Engineering: An Example from ChatGPT and Automated Program Repair.** 17. abr. 2024. arXiv. Disponível em: <<http://arxiv.org/abs/2310.08879>>. Acesso em: 28/04/2024.
- [7] STRAFFORELLO, F. **How Transformers and Large Language Models (LLMs) Work — A Comprehensive Guide Using BERT, GPT, and T5.** Medium Magazine. 11/06/2023. Disponível em <https://blog.gopenai.com/how-transformers-and-large-language-models-llms-work-3f20bb41c1ff>
- [8] KUHN, W. **Semantic Reference Systems.** In: Gould, M. F.; Laurini, R.; Coulonde, S. (Eds.). AGILE 2003: 6th AGILE Conference on Geographic Information Science. Heidelberg: Springer, 2003. p. 63-72.
- [9] YU, L.; QIU, P.; LIU, X.; LU, F.; WAN, B. **A holistic approach to aligning geospatial data with multidimensional similarity measuring.** International Journal of Digital Earth, v. 11, n. 8, p. 845–862, 2018.
- [10] ANAND, S.; MORLEY, J.; JIANG, W.; DU, H.; HART, G. **When worlds collide: combining Ordnance Survey and Open Street Map data.** In: AGI Geocommunity '10, London, UK. 2010.
- [11] DU, H.; ALECHINA, N.; JACKSON, M.; HART, G. **Matching Formal and Informal Geospatial Ontologies.** In: D. Vandenbroucke; B. Bucher; J. Crompvoets (Orgs.); Geographic Information Science at the Heart of Europe, Lecture Notes in Geoinformation and Cartography. p.155–171, 2013. Cham: Springer International Publishing. Disponível em: <[https://link.springer.com/10.1007/978-3-319-00615-4\\_9](https://link.springer.com/10.1007/978-3-319-00615-4_9)>. Acesso em: 03/06/2023.
- [12] MACHADO, A. A.; CAMBOIM, S. P. **Semantic Alignment of Official and Collaborative Geospatial Data: A Case Study in Brazil.** Revista Brasileira de Cartografia, v. 76, n. 1, 2024.
- [13] LIU, Y. et al. **RoBERTa: A robustly optimized BERT pre-training approach.** 2019. Disponível em <https://arxiv.org/abs/1907.11692>
- [14] FAWCETT, T. **An introduction to ROC analysis.** Pattern Recognition Letters, v. 27, n. 8, p. 861-874, 2006.
- [15] DÜNTSCH, I.; GEDIGA, G. **Confusion Matrices and Rough Set Data Analysis.** [cs.AI], 2019. Disponível em: <https://arxiv.org/abs/1902.01487>. Acesso em: 18 jul. 2024.
- [16] CONCAR. Comissão Nacional de Cartografia. **Especificações Técnicas para Estruturação de Dados Geoespaciais Vetoriais (ET-EDGV 3.0).** NCB-CC/E 0001B08. Versão 3.0 - 2017.
- [17] OSM. OpenStreetMap. **Map Features.** Disponível em [https://wiki.openstreetmap.org/wiki/Map\\_features](https://wiki.openstreetmap.org/wiki/Map_features) Acesso em 29/05/2024.
- [18] SILVA, J. C. da; VIEIRA, R. O. **Introdução às Redes Neurais Profundas com Python.** In: TELES, A. S.; SILVA, D. B. da; ESMERALDO, G. A. R. M. (Eds.). Minicursos da ERCEMAPI. Porto Alegre: Sociedade Brasileira de Computação (SBC), 2022. Disponível em: <https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/110/498/771-1>. Acesso em: 18 jul. 2024.