Infraestrutura em Nuvem para Experimentação Científica em Larga Escala: Relato de Experiência de um Projeto sobre Classificação Automática de Texto

Leonardo Rocha[1], **Washington Cunha**[2], **Vitor Mangaravite**[3], **Marcos André Gonçalves**[3]

[1]Universidade Federal de São João del Rei (UFSJ), Brasil

[2]Universidade Estadual de Campinas (Unicamp), Brasil

[3]Universidade Federal de Minas Gerais (UFMG), Brasil

lcrocha@ufsj.edu.br, wcunha@unicamp, {mangaravite,mgoncalv}@dcc.ufmg.br

Resumo. *Este artigo apresenta um relato de experiência sobre o uso de computação em nuvem no projeto Comparando a Efetividade de Abordagens Neurais e Não-Neurais em Tarefas de Classificação Automática de Texto, desenvolvido entre 2020 e 2022 com apoio do CNPq e da AWS. A infraestrutura em nuvem viabilizou uma avaliação em larga escala, com múltiplas representações textuais, algoritmos e coleções, totalizando mais de 10.000 execuções em instâncias otimizadas para GPU, CPU e memória. O relato destaca a importância da elasticidade da nuvem, da escolha adequada de instâncias e da análise conjunta entre efetividade e custo computacional.*

Abstract. *This paper presents an experience report on the use of cloud computing in the project Comparing the Effectiveness of Neural and Non-Neural Approaches in Automatic Text Classification Tasks, developed between 2020 and 2022 with support from CNPq and AWS. Cloud infrastructure enabled a large-scale evaluation involving multiple text representations, algorithms, and datasets, totaling more than 10,000 runs on GPU-, CPU-, and memory-optimized instances. The report highlights the importance of cloud elasticity, suitable instance selection, and the joint analysis of effectiveness and computational cost.*

1. Introdução

A classificação automática de texto é componente central de diversas aplicações de recuperação da informação, mineração de dados e processamento de linguagem natural [Aggarwal and Zhai 2012]. Em contextos institucionais e organizacionais, ela pode apoiar a triagem de documentos, a identificação temática de acervos e a automação de fluxos de análise textual. Ao longo da última década, a área passou por forte expansão com o surgimento de novas representações de texto e de arquiteturas neurais [Lin 2019], levando à percepção de que métodos baseados em redes neurais se tornariam, de forma quase universal, a melhor escolha para tarefas de classificação.

Entretanto, a consolidação dessa percepção nem sempre foi acompanhada do mesmo rigor experimental. Uma parte importante da literatura recente passou a comparar novas abordagens com linhas de base fracas, parâmetros insuficientemente ajustados, protocolos sem repetições e métricas pouco adequadas a cenários desbalanceados.

Como consequência, tornou-se difícil responder de modo confiável a questões fundamentais para a prática científica e aplicada: métodos neurais são de fato superiores em qualquer cenário? O custo computacional adicional é justificado por ganhos consistentes? O porte e as características das coleções alteram o balanço entre efetividade e eficiência?

Responder a essas perguntas exigia mais do que uma nova combinação de algoritmos. Exigia um desenho experimental amplo [Strubell et al. 2019], capaz de combinar múltiplas coleções, diferentes estratégias de representação textual, classificadores neurais e não neurais, ajuste adequado de parâmetros e análise sistemática dos resultados. Esse tipo de investigação, embora metodologicamente desejável, impõe demandas computacionais incompatíveis com a infraestrutura local tipicamente disponível em muitos grupos de pesquisa.

Nesse contexto, este artigo apresenta um relato de experiência sobre a execução do projeto *Comparando a Efetividade de Abordagens Neurais e Não-Neurais em Tarefas de Classificação Automática de Texto*, desenvolvido entre 2020 e 2022 com financiamento do CNPq e apoio de infraestrutura da AWS. Mais do que discutir apenas os achados científicos do estudo, o foco aqui está em descrever como a computação em nuvem foi determinante para transformar um plano metodologicamente ambicioso em uma investigação executável em larga escala.

As contribuições deste relato são três. Primeiro, documentamos como a infraestrutura em nuvem foi organizada para sustentar mais de 10.000 execuções experimentais com diferentes perfis de processamento. Segundo, descrevemos decisões práticas de alocação de recursos computacionais associadas a restrições de prazo, custo e tipo de algoritmo. Terceiro, sintetizamos os principais aprendizados dessa experiência para projetos acadêmicos que demandam experimentação intensiva em inteligência artificial e ciência de dados.

2. Contexto e motivação do projeto

O projeto nasceu da constatação de que a literatura sobre classificação automática de texto vinha acumulando um número crescente de propostas neurais, mas nem sempre fornecia evidências suficientemente robustas sobre sua superioridade prática. Em paralelo, a adoção crescente de métodos de aprendizado de máquina em cenários reais, tais como análise documental, organização de acervos e apoio a fluxos institucionais, tornava ainda mais relevante compreender o trade-off entre custo computacional e qualidade dos resultados [Aggarwal and Zhai 2012]. A pergunta que orientou o projeto foi direta: *abordagens neurais para classificação automática de texto são universalmente mais eficazes do que abordagens não neurais, independentemente das características das coleções de dados e dos custos envolvidos?* Associada a essa pergunta, havia um pressuposto importante: as características dos dados e as restrições de infraestrutura influenciam de maneira decisiva a escolha do método mais apropriado.

A partir disso, o projeto estabeleceu como objetivo geral fornecer insumos para apoiar a seleção de algoritmos de classificação automática de texto em ambientes acadêmicos e organizacionais, considerando simultaneamente eficácia e eficiência. Em vez de buscar apenas o melhor resultado absoluto em uma base específica, a proposta era construir uma comparação ampla e cientificamente sólida, capaz de orientar escolhas em cenários diversos.

A execução dessa agenda implicava desafios concretos [Lin 2019]. Era necessário trabalhar com diferentes famílias de representações textuais, combinar métodos tradicionais e arquiteturas recentes, considerar bases de diferentes tamanhos e perfis e repetir experimentos sob protocolos comparáveis. Em termos operacionais, isso significava lidar com um volume muito elevado de execuções e com necessidades computacionais heterogêneas, variando de tarefas intensivas em GPU a rotinas mais dependentes de CPU ou memória principal.

Foi precisamente nesse ponto que a computação em nuvem deixou de ser apenas uma conveniência tecnológica e passou a compor a própria viabilidade metodológica do projeto.

3. Caracterização da experiência

3.1. Escopo experimental

A experiência relatada neste artigo teve como núcleo a realização de um estudo comparativo abrangente sobre métodos de classificação automática de texto. No artigo principal resultante do projeto, foram avaliadas nove coleções amplamente utilizadas pela comunidade [Zhang et al. 2016], sendo cinco de grande escala (i.e., AG's News, Sogou News, Yelp Review 2015, IMDB Reviews e Yahoo! Answers) e quatro coleções menores, mas clássicas (i.e., 20 Newsgroups, WebKB, Reuters e ACM Digital Library).

A escolha de bases grandes e menores foi intencional. Em aplicações reais, especialmente fora de plataformas consolidadas, é comum que os conjuntos de treinamento sejam limitados em tamanho devido ao custo de rotulação manual. Assim, comparar o comportamento dos métodos em ambos os cenários era necessário para obter conclusões mais úteis praticamente.

No que se refere às representações textuais, o projeto contemplou desde abordagens tradicionais, como TF-IDF, até embeddings distribuídos e representações mais semânticas, incluindo FastText [Bojanowski et al. 2017], PTE [Tang et al. 2015], SWEM [Shen et al. 2018], TextGCN [Yao et al. 2019] e MetaFeatures [Canuto et al. 2018]. Quanto aos classificadores, foram explorados métodos não neurais e neurais representativos, como SVM [Joachims 1998], CNN [Zhang et al. 2016], LSTM [Hochreiter and Schmidhuber 1997], HAN [Yang et al. 2016], BERT [Devlin et al. 2019] e XLNet [Yang et al. 2019]. Em extensões e desdobramentos do projeto, outros modelos e combinações também foram investigados, ampliando ainda mais a carga experimental.

3.2. Magnitude da execução

A combinação de múltiplas bases de dados, diferentes representações, algoritmos de classificação e configurações experimentais resultou em mais de 10.000 execuções distintas ao longo do projeto. Esse volume excedia, de forma clara, a capacidade operacional de infraestrutura local convencional, sobretudo quando se consideram simultaneamente o tempo de processamento, a necessidade de paralelização, a diversidade de requisitos computacionais e a pressão por prazos de publicação.

Além da quantidade de execuções, outro aspecto relevante foi a heterogeneidade dos experimentos. Modelos baseados em transformers e outras arquiteturas neurais

exigiam aceleração por GPU; métodos tradicionais e pipelines de pré-processamento pesado demandavam instâncias com boa capacidade de CPU; e alguns algoritmos, como o XGBoost, beneficiavam-se especialmente da disponibilidade ampliada de memória RAM. Em outras palavras, não havia uma única configuração de máquina adequada para todo o projeto.

3.3. Objetivo do relato

Diante desse cenário, o objetivo deste relato não é reproduzir em detalhe todos os resultados técnicos do estudo original, mas sim documentar a experiência de planejamento, seleção e uso da infraestrutura em nuvem para viabilizar uma agenda experimental ampla. O interesse central está em mostrar como a nuvem foi incorporada à rotina da pesquisa, quais decisões práticas precisaram ser tomadas e que aprendizados emergiram desse processo.

4. Infraestrutura em nuvem adotada

A infraestrutura da AWS foi organizada em três grupos principais de recursos, selecionados de acordo com o perfil de processamento dos experimentos [Strubell et al. 2019].

4.1. Instâncias aceleradas por GPU

Para projetos relacionados a *deep learning*, foram utilizadas as instâncias `g4dn.2xlarge` e `p3.2xlarge`. Essas instâncias foram empregadas em experimentos com modelos como RoBERTa, GPT, DistilBERT, ALBERT, BART, BERT, LSTM, XLNet, VDCNN, CNN e HAN.

Embora ambas disponibilizassem 16 GB de memória de GPU, havia diferenças práticas importantes entre elas. As instâncias da família P3 apresentavam custo mais elevado, mas também maior poder de processamento. Na prática, isso levou a uma estratégia de uso orientada por prazo: instâncias `p3.2xlarge` eram priorizadas em fases próximas às submissões de artigos ou em momentos que exigiam respostas mais rápidas, enquanto instâncias `g4dn.2xlarge` eram adotadas em experimentos de horizonte mais longo, com melhor relação custo-desempenho.

4.2. Instâncias otimizadas para computação

Para métodos fortemente dependentes de CPU, foram utilizadas instâncias `c5a.8xlarge`. Esse grupo foi particularmente importante para experimentos com SVM, PTE, SWEM, FastText, MetaFeatures e CluWords. Em tais cenários, a disponibilidade de processadores robustos desempenhou papel decisivo para manter o fluxo experimental em escala e reduzir gargalos nas tarefas de treinamento e avaliação.

4.3. Instâncias otimizadas para memória

Algoritmos com uso intensivo de memória principal foram alocados em instâncias `r5a.4xlarge`. Um exemplo citado no projeto foi o XGBoost, cujo desempenho pode se beneficiar de maior disponibilidade de RAM, especialmente em cenários com conjuntos de dados maiores ou estruturas intermediárias mais pesadas.

4.4. Critérios de escolha

A experiência mostrou que a escolha da infraestrutura não pode ser feita apenas com base em uma noção genérica de "mais poder computacional". O critério adotado no projeto combinou, ao menos, quatro fatores: natureza do algoritmo, tamanho do conjunto de dados, urgência associada ao cronograma científico e custo do recurso. Essa lógica permitiu distribuir melhor os experimentos e evitar tanto subutilização quanto alocação excessivamente onerosa.

5. Execução do projeto e resultados alcançados

A disponibilidade da infraestrutura em nuvem permitiu a execução de uma agenda experimental coerente com a ambição metodológica originalmente proposta. Em vez de restringir o estudo a poucas bases ou a um conjunto reduzido de modelos, foi possível realizar uma comparação extensa entre abordagens neurais e não neurais, abrangendo diferentes representações textuais e distintos cenários de aplicação.

No artigo principal derivado do projeto, o estudo reuniu mais de 1,5 mil medições consolidadas, produzindo achados relevantes para a área. Um dos resultados mais expressivos foi a constatação de que SVM com TF-IDF, quando adequadamente parametrizado, manteve-se altamente competitivo, tanto em bases menores quanto em vários cenários de maior escala. Esse resultado relativizou a percepção de que métodos tradicionais seriam necessariamente superados por arquiteturas mais recentes.

Entre as abordagens neurais *end-to-end*, arquiteturas baseadas em transformers, como BERT e XLNet, apresentaram desempenho elevado, mas sem dominância universal. Em diversos casos, outras arquiteturas neurais mais simples ou menos custosas alcançaram efetividade semelhante. Além disso, abordagens em duas etapas, como MetaFeatures, mostraram-se competitivas em diferentes cenários, inclusive do ponto de vista de custo-benefício.

Esses resultados reforçaram uma conclusão importante: não existe uma única solução melhor para todos os casos. A escolha entre as abordagens deve considerar simultaneamente o perfil da coleção, as exigências de qualidade e os recursos computacionais disponíveis. Em outras palavras, a infraestrutura não é um aspecto externo ao problema; ela integra a própria decisão metodológica.

Do ponto de vista da produção científica, a experiência resultou em pelo menos dois desdobramentos diretos. O primeiro, mais diretamente vinculado ao projeto, foi o artigo *On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study*, publicado em *Information Processing & Management*, em 2021 [Cunha et al. 2021]. O segundo foi o artigo *On the Cost-Effectiveness of Stacking of Neural and Non-Neural Methods for Text Classification: Scenarios and Performance Prediction*, publicado na ACL 2021 [Gomes et al. 2021], como desdobramento complementar.

6. Lições aprendidas

A execução do projeto permitiu consolidar um conjunto de aprendizados relevantes para pesquisas que dependem de infraestrutura digital em larga escala [Strubell et al. 2019].

6.1. A nuvem como elemento metodológico

A principal lição foi perceber que a nuvem não atuou apenas como suporte operacional. Ela passou a compor a própria estratégia metodológica do projeto, permitindo ampliar o número de comparações, controlar melhor o cronograma e sustentar um desenho experimental mais robusto. Sem essa infraestrutura, o estudo provavelmente teria sido reduzido em escopo, número de métodos ou diversidade de bases.

6.2. Elasticidade e gestão de prazos acadêmicos

Outra lição importante foi o valor da elasticidade em contextos acadêmicos. Projetos de pesquisa frequentemente alternam períodos de menor atividade com momentos de alta demanda, especialmente em torno de submissões de artigos, revisões ou consolidação de resultados. A possibilidade de variar o tipo de instância conforme a urgência do momento mostrou-se particularmente útil para equilibrar custo e tempo.

6.3. Alinhamento entre algoritmo e instância

A experiência também evidenciou que a escolha da infraestrutura deve ser sensível ao perfil computacional do método. Nem todo experimento em aprendizado de máquina requer GPU, e nem todo gargalo se resolve com mais núcleos de processamento. Em projetos com grande diversidade de algoritmos, mapear previamente o perfil de consumo de cada grupo de métodos reduz desperdícios e melhora a previsibilidade da execução [Mendes et al. 2020].

6.4. Escala exige organização

O aumento do poder computacional, por si só, não resolve o problema da pesquisa em larga escala. O projeto exigiu cuidadosa organização das execuções, definição de prioridades, controle de versões de experimentos e clareza sobre o que deveria ser executado em cada etapa. A nuvem ampliou a capacidade de execução, mas também tornou ainda mais importante o planejamento experimental.

6.5. Custo-efetividade como critério científico

Por fim, um aprendizado conceitual importante foi a necessidade de incorporar o custo computacional como dimensão legítima da avaliação científica. Em muitos cenários, o ganho marginal de efetividade de um método mais caro pode não justificar sua adoção. Projetos apoiados na nuvem tornam essa comparação ainda mais concreta, pois evidenciam o impacto das escolhas algorítmicas no consumo de recursos.

7. Desafios e limitações

Apesar dos ganhos obtidos, a experiência também envolveu desafios. O primeiro deles foi a necessidade de gerenciar diferentes perfis de máquina e de monitorar continuamente o uso dos recursos, evitando tanto a ociosidade quanto o sobrecusto. Em projetos com muitas execuções, pequenas ineficiências podem se acumular rapidamente.

Outro desafio diz respeito à reprodutibilidade operacional. Embora a nuvem facilite a disponibilidade de infraestrutura, ela não elimina a necessidade de documentar cuidadosamente pipelines, dependências, parâmetros e versões de bibliotecas. Em

estudos extensos, essa documentação é fundamental para permitir reexecução, auditoria e extensão futura dos experimentos.

Também é importante reconhecer que o relato enfatiza um caso bem-sucedido em um projeto com apoio institucional e financiamento dedicado. Outros grupos podem enfrentar restrições adicionais de crédito, de capacitação técnica ou de apoio à gestão da infraestrutura. Ainda assim, os princípios extraídos da experiência, tais como a adequação entre instância e tarefa, a elasticidade orientada por cronograma e a avaliação de custo-benefício, permanecem amplamente úteis.

8. Conclusão

Este artigo apresentou um relato de experiência sobre o uso de infraestrutura em nuvem na execução de um projeto de pesquisa voltado à comparação entre abordagens neurais e não neurais para classificação automática de texto. Desenvolvido entre 2020 e 2022, o projeto exigiu grande volume de experimentação, diversidade de recursos computacionais e capacidade de adaptação a diferentes perfis de processamento.

A experiência mostrou que a AWS foi decisiva para transformar um plano experimental ambicioso em uma investigação efetivamente realizável. A possibilidade de combinar instâncias aceleradas por GPU, instâncias otimizadas para computação e instâncias orientadas a memória permitiu alinhar infraestrutura e método, suportando mais de 10.000 execuções distintas. Além disso, a elasticidade da nuvem contribuiu para lidar com picos de demanda associados ao calendário acadêmico.

Para além dos resultados científicos do projeto, o relato reforça uma mensagem prática: nas pesquisas em inteligência artificial e ciência de dados, a infraestrutura não deve ser tratada apenas como suporte técnico, mas como parte integrante da estratégia metodológica. Em especial, projetos que dependem de avaliações comparativas amplas podem se beneficiar significativamente de abordagens em nuvem quando utilizadas de forma planejada, criteriosa e sensível ao trade-off entre custo e efetividade.

Agradecimentos

Este trabalho foi apoiado por: CNPq, CAPES, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR), FAPEMIG, AWS, Google, NVIDIA, CIIASaúde e FAPESP.

Declaração sobre uso de Inteligência Artificial

Os autores utilizaram ferramenta de inteligência artificial generativa como apoio à organização inicial do manuscrito e à revisão textual. Todas as decisões relacionadas ao conteúdo, à estrutura argumentativa, à seleção das informações relatadas e à redação final do artigo foram realizadas e revisadas pelos autores, que se responsabilizam integralmente pelo conteúdo submetido.

Referências

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining Text Data*, pages 163–222.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Canuto, S., Sousa, D. X., Gonçalves, M. A., and Rosa, T. C. (2018). A thorough evaluation of distance-based meta-features for automated text classification. *IEEE Transactions on Knowledge and Data Engineering*, 30(12):2242–2256.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., Rosa, T., Rocha, L., and Gonçalves, M. A. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Gomes, C., Goncalves, M., Rocha, L., and Canuto, S. (2021). On the cost-effectiveness of stacking of neural and non-neural methods for text classification: Scenarios and performance prediction. In *ACL-IJCNLP 2021*. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML*, pages 137–142.

Lin, J. (2019). The neural hype and comparisons against weak baselines. *ACM SIGIR Forum*, 52(2):40–51.

Mendes, L. F., Gonçalves, M. A., Cunha, W., Rocha, L. C., Rosa, T. C., and Martins, W. (2020). Keep it simple, lazy: Metalazy, a new metastrategy for lazy text classification. In *Proceedings of CIKM*, pages 1125–1134.

Shen, D., Wang, G., Wang, W., Min, M. R., Su, Q., Zhang, Y., Li, C., Henao, R., and Carin, L. (2018). Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of ACL*, pages 440–450.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. In *Proceedings of ACL*, pages 3645–3650.

Tang, J., Qu, M., and Mei, Q. (2015). Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of KDD*, pages 1165–1174.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5754–5764.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489.

Yao, L., Mao, C., and Luo, Y. (2019). Graph convolutional networks for text classification. In *Proceedings of AAAI*, pages 7370–7377.

Zhang, X., Zhao, J., and LeCun, Y. (2016). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657.