

Washington Cunha, Ph.D. Student

✉ washingtoncunha@dcc.ufmg.br

in <https://www.linkedin.com/in/washington-l-m-cunha/>

📞 +5531992179467

🔗 <https://www.github.com/waashk>

Research Interest

Washington 📌 I'm a Ph.D. Student in the Machine Learning and Databases Laboratory (LBD), advised by professors Marcos Goncalves and Leonardo Rocha at Federal University of Minas Gerais. My main research goal focuses on an under-investigated data engineering technique, but whose potential is enormous in the current scenario known as Instance Selection (IS). The IS goal is to reduce the training set size by removing noisy or redundant instances while maintaining the effectiveness of the trained models and reducing the training process cost. In this context, we provided (Accepted on **CSUR'23**) a comprehensive and scientifically sound comparison of IS methods applied to Text Classification, considering several classification solutions and many datasets, answering questions that reveal an enormous unfulfilled potential for IS solutions. We also proposed (Accepted on **SIGIR'23**) a two-step framework aimed at large datasets with a particular focus on Transformer architectures. Our solution managed to reduce the training sets by almost 30% on average while maintaining the same levels of effectiveness in all datasets, with speedup improvements of 37% (up to 70%), scaling for datasets with hundreds of thousands of documents.

Education

- 2019 – Actual 📌 **Ph.D. Student**, Federal University of Minas Gerais.
Project Title: *A Comprehensive Exploitation of Instance Selection Methods for Automatic Text Classification*
- 2018 – 2019 📌 **Master Degree in Computer Science** in Federal University of Minas Gerais
Thesis title: *Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling.*
- 2014 – 2017 📌 **Bachelor Degree in Computer Science** in Federal University of São João del-Rei.
Thesis title: *A Feature-oriented Sentiment Rating for Mobile App Reviews.*

Research Publications

Journal Articles

- 1 A comparative survey of instance selection methods applied to nonneural and transformer-based text classification. (2023). *ACM Computing Surveys - Imp. Fac.*: 14.324.
- 2 On the class separability of contextual embeddings representations—or “the classifier does not matter when the (text) representation is so good!” (2023). *IP&M – Imp. Fac.*: 7.466, 60(4), 103336.
- 3 On the cost-effectiveness of neural and non-neural approaches and representations for text classification. (2021). *IP&M*. [🔗 doi:10.1016/j.ipm.2020.102481](https://doi.org/10.1016/j.ipm.2020.102481)
- 4 Stroke outcome measurements from electronic medical records: Cross-sectional study on the effectiveness of neural and nonneural classifiers. (2021). *JMIR Med Inform.* [🔗 doi:10.2196/29120](https://doi.org/10.2196/29120)
- 5 Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. (2020). *IP&M*. [🔗 doi:10.1016/j.ipm.2020.102263](https://doi.org/10.1016/j.ipm.2020.102263)

Conference Proceedings

- 1 An Effective, Efficient, and Scalable Confidence-Based Instance Selection Framework for Transformer-Based Text Classification. (2023), In **SIGIR'23** – *h5-index: 75*.
- 2 CluHTM - Semantic Hierarchical Topic Modeling based on CluWords. (2020), In **ACL'20** – *h5-index: 169*. [doi:10.18653/v1/2020.acl-main.724](https://doi.org/10.18653/v1/2020.acl-main.724)
- 3 “Keep It Simple, Lazy” – MetaLazy: A New MetaStrategy for Lazy Text Classification. (2020), In **CIKM'20** – *h5-index: 69*. [doi:10.1145/3340531.3412180](https://doi.org/10.1145/3340531.3412180)
- 4 CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. (2019), In **WSDM'19** – *h5-index: 69*. [doi:10.1145/3289600.3291032](https://doi.org/10.1145/3289600.3291032)
- 5 A Feature-oriented Sentiment Rating for Mobile App Reviews. (2018), In *The world wide web conference WWW'18* – *h5-index: 98*. [doi:10.1145/3178876.3186168](https://doi.org/10.1145/3178876.3186168)
- 6 Semantically-Enhanced Topic Modeling. (2018), In **CIKM'18** – *h5-index: 69*. [doi:10.1145/3269206.3271797](https://doi.org/10.1145/3269206.3271797)

Skills

Coding	📖 Python, C, C++, CUDA, R, AWK, Shellscript,
Research	📖 Machine Learning and Data Mining, focusing in Automatic Text Classification, Feature Engineering, Sentiment Analysis and Topic Modeling.
Libraries	📖 Python: Scikit-learn, Numpy, Pandas, Pytorch, Keras, Scipy, Mxnet, Matplotlib, Seaborn.
Dev	📖 AWS, Linux, Git, Docker, Jupyter-notebook, Google Colab.
Languages	📖 Portuguese and English.

Miscellaneous Experience

Awards and Achievements

- 2023 📖 **SIGIR Student Travel Awards** for present an accepted full research paper at SIGIR'23.
- 2021 📖 **Honorable Mention** in the Masters Theses Contest of the Brazilian Database Symposium. **CTDBD – SBBD'21**
- 2019 📖 **Among of 10 best Brazilian Scientific Initiations** Research selected by Brazilian Computer Society (As co-advisor). **CTIC'19 – SBC**.




Advanced Courses

- 2020 📖 Natural Language Processing - Deep Learning Algorithms
- 2019 📖 Deep Learning Algorithms
- 📖 Information Theory
- 2018 📖 Information Retrieval and Social Computing
- 📖 Quantitative Methods of Experimental Research in Computer Science
- 📖 Machine Learning
- 📖 Fundamentals of Statistics for Data Science












Certification

- 2020 📖 Coursera Specialization: Deep Learning - Coursera
- 📖 Coursera Specialization: Applied Data Science with Python - Coursera

Miscellaneous Experience (continued)

- 2019  Coursera Specialization: AWS Fundamentals - Coursera
-  Practical Project Management - Udemy
- 2019  Fundamentals of Accelerated Computing with CUDA C/C++ - NVIDIA

Academic Service

- 2023  International Conference on the Theory of Information Retrieval (**ICTIR**) - PC member
-  Conference on Neural Information Processing Systems (**NeurIPS**) - Reviewer
-  TheWebConf2023 - Reviewer
- 2022-Actual  Connection Science Journal - Reviewer
- 2021-Actual  Association for Computational Linguistics (**ACL**) - Reviewer
-  Conference on Empirical Methods in Natural Language Processing (**EMNLP**) - Reviewer
-  The Conference on Information and Knowledge Management (**CIKM**) - Sub-Reviewer
-  The European Conference on Information Retrieval (**ECIR**) - Sub-Reviewer
-  Association for Computational Linguistics (**ACL**) Mentorship - Participant
-  ACM and SBC Membership
- 2021  WebMedia Conference - Volunteer - Squad Leader