A Comprehensive Exploitation of Instance Selection Methods for Automatic Text Classification

"Doing More with Less"

Washington Cunha[1], **Leonardo Rocha**[2] **(Co-advisor), Marcos A. Gonçalves**[1] **(Advisor)**

[1]Department of Computer Science – Federal University of Minas Gerais – Brazil

[2]Department of Computer Science – Federal University of São João del Rei – Brazil

{washingtoncunha,mgoncalv}@dcc.ufmg.br, lcrocha@ufsj.edu.br

Abstract. *Recent progress in NLP has followed a "more is better" trend (more data, computing power, and model complexity) best exemplified by the Large Language Models (LLMs). However, training such models remains resource-intensive. This Ph.D. dissertation explores Instance Selection (IS), a promising yet underexplored data engineering technique that reduces training set size by removing noisy or redundant instances, lowering computational cost without sacrificing performance. We evaluate comprehensively the IS methods for Automatic Text Classification (ATC) across several classifiers and 22 datasets, uncovering significant untapped potential. Additionally, we propose two novel IS methods tailored for large datasets and LLMs. Our best solution cut training set sizes by 41% on average while preserving effectiveness, and achieved speedups of up to 2.46x, proving its scalability.*

Resumo. *Progresso recente em PLN seguiu a tendência de "quanto mais, melhor" (mais dados, poder computacional e complexidade de modelos), exemplificada pelos Grandes Modelos de Linguagem. Contudo, o treinamento desses modelos continua sendo um processo intensivo em recursos. Esta tese de doutorado explora a Seleção de Instâncias (SI), uma técnica de engenharia de dados promissora, porém pouco explorada, que reduz o tamanho do conjunto de treinamento removendo instâncias ruidosas ou redundantes, reduzindo o custo computacional sem sacrificar a qualidade. Avaliamos de forma abrangente os métodos de SI para classificação automática de texto em diversos modelos e 22 conjuntos de dados, revelando um potencial significativo inexplorado. Além disso, propomos dois novos métodos de SI com foco em grandes conjuntos de dados e LLMs. Nossa melhor solução reduziu o tamanho dos conjuntos de treinamento em 41% em média, preservando a qualidade, e alcançou speed-ups de até 2,46x, comprovando sua escalabilidade.*

1) Thesis Defense Data and Highlights : **Category:** Ph.D.; **Defended** on August 26th, 2024

Advisor: Marcos A. Gonçalves (UFMG); **Co-advisor:** Leonardo Rocha (UFSJ)

Board Members: Franco Maria Nardini (CNR - Italy); Thierson Couto Rosa (UFG); Rodrygo Luis Teodoro Santos (UFMG); Anisio Mendes Lacerda (UFMG)

Highlights:

- We thoroughly compared Instance Selection (IS) methods applied to Text Classification
- We survey and compare 10 IS baselines along with 9 LLMs applied to 22 datasets
- Our novel solutions reduce the training set size by up to 60% while maintaining effectiveness
- We do so while achieving speedups up to 2.46x and 55% reduction in carbon emission
- Our methods set the state-of-the-art in NLP Instance Selection based on 5000+ measurements

Dissertation available at: `http://hdl.handle.net/1843/76441`

Publications available at: `http://bit.ly/3WvX1T5`

Code and Data available at: `github.com/waashk/instanceselection`

2) Context and problem: The rapid data growth on the Web, social network platforms, companies, and governmental institutions has made organizing and retrieving content extremely challenging. Automatic Text Classification (ATC) offers a solution to this problem by mapping textual documents into predefined semantic categories. Accurate ATC models have become crucial for many emerging applications [Cunha et al. 2023b], such as spam, fake news and hate speech detection, relevance feedback, sentiment and product review analysis, to cite just a few. As a supervised task, ATC benefits from applications generating large volumes of *labeled data*, such as social networks (e.g., X, Facebook, WhatsApp). Crowdsourcing and soft labeling [Roy and Cambria 2022] further reduce the costs of acquiring labeled data. Thus, labeling has become less of an issue, while the abundance of labeled data is.

According to Andrew Ng [Ng 2016], the success of Transformer-based architectures, the state-of-the-art (SOTA) in ATC, best exemplified by Small and Large Language Models (SLMs and LLMs) such as RoBERTa and Llama 4, is due to extensive pre-training on massive datasets (e.g., 45PB for GPT-4) and the adaptability of pre-trained models via fine-tuning. This approach enables faster task-specific training compared to starting from scratch [Uppaal et al. 2023]. However, fine-tuning remains resource-intensive. Despite being faster than full training, it still requires significant computational power. For instance, fine-tuning the SLM XLNet in the MEDLINE dataset took 80 GPU hours in our experiments. Resource limitations in companies and research groups also restrict experimentation with such models. For instance, in our PhD dissertation alone, we ran over 5,000 experiments that took approximately 5,600 hours in a specialized (GPU-based) architecture. Reducing financial, computational, and environmental costs is crucial, given the significant energy consumption and carbon emissions associated with generating and using (large) language models.

3) Objective: Given increasing data volumes, re-training demands, and environmental concerns, proposing scalable and cost-effective NLP and ATC strategies has become essential. These include creating efficient deep learning algorithms, using advanced hardware, or improving data preprocessing techniques. The recent success and real-world impact, including financial, of DeepSeek [DeepSeek et al. 2025], which matched or surpassed the effectiveness of SOTA LLMs while reducing computational demands, highlights the importance of the trade-off effectiveness vs. cost to the research and practitioner communities. This PhD dissertation focuses exactly on this trade-off, one of the SBC 2025-2035 Grand Challenges on Computer Science, from a *data engineering perspective*, aiming to enhance model performance while reducing costs. In particular, we focus on Instance Selection (IS), an understudied (in NLP and ATC at least), yet promising, set of techniques and growing research area [Cunha et al. 2020, Cunha et al. 2021], focused on selecting the most representative instances (documents) for a training set [Garcia et al. 2012]. The intuition behind IS is to remove potentially noisy or redundant instances from the original training set and improve performance in terms of total training time while keeping or even improving effectiveness. IS methods have three main concomitant goals: *(i) to reduce instances by selecting the most representative ones; (ii) to maintain effectiveness by removing noise and redundancy;* and *(iii) to reduce the total time for applying an end-to-end model (from preprocessing to model training to deployment (test).*

4) Summary of the proposed solution: The main contributions of our PhD dissertation are fourfold: (i) a comprehensive survey of the IS methods applied to ATC, including (ii) an extended IS taxonomy; and (iii-iv) two novel state-of-the-art (SOTA) IS approaches applied to NLP/ATC. Due to space limitations, we focus on the latter two, noticing that our article on the ACM Computing Surveys [Cunha et al. 2023b], derived from the dissertation, has been highly cited (52 citations as of April/2025). First, we proposed **E2SC** [Cunha et al. 2023a], a two-step IS framework aimed at large datasets with a special focus on transformer-based architectures. E2SC's first step assigns a probability to each instance being removed from the training set. We adopted an exact KNN model solution to estimate the probability of removing (training) instances, as KNN is considered a calibrated [Rajaraman et al. 2022] and computationally cheap (fast) classifier. Our first hypothesis (H1) was that *high confidence (if the model is calibrated to the correct class known in training) positively correlates with redundancy for the sake of building a stronger classification model.* In the second E2SC step, we estimate a near-optimal reduction rate that does not degrade the Transformer's effectiveness by employing a validation set and a weak but fast classifier. Our second hypothesis (H2) was that *we can estimate the effectiveness of a robust model through the analysis and variation of selection rates in a weaker model*. Again, we explored KNN to gather evidence for this hypothesis by introducing an iterative method that statistically compared, using a validation set, the KNN model's effectiveness without any data reduction against the model with iterative data reduction rates. In this way, we could estimate a reduction rate that did not affect the KNN model's effectiveness.

E2SC focused only on *redundancy*. Despite excellent results regarding the trade-off effectiveness-efficiency-reduction, other aspects such as **noise** – defined as instances incorrectly labeled by humans [Martins et al. 2021] as well as outliers that do not contribute to model learning – were not explored in our first solution. To fill this gap, we proposed **biO-IS** [Cunha et al. 2025], built on top of E2SC, aimed at simultaneously removing redundant and noisy instances. **biO-IS** has three main components: (i) a weak classifier; (ii) a redundancy-based approach; and (iii) an entropy-based approach. We departed from E2SC, considering the Logistic Regression (LR) as the calibrated weak classifier instead of KNN – in further experiments described in the dissertation, LR proved to be the best classifier for effectiveness and calibration. To address the second objective of noise removal, we proposed a new step to be combined with our previous IS solution based on entropy, as well as a novel iterative process to estimate near-optimum reduction rates. Considering wrongly predicted instances by the weak classifier, the main objective is to assign a probability to each of them being removed from the training set based on the probability of the instance being noisy. For this, we proposed using entropy as a proxy to determine the reduction behavior for the sake of training a stronger model.

5) Evaluation: We compared our proposals with 13 of the most robust SOTA IS baseline methods in the ATC domain based on our systematic literature review, considering 22 datasets and 7 SOTA small and large language models (including BERT, RoBERTa, Llama, among others). Our experimental evaluation showed that **E2SC** was able to reduce the size of the training sets by 29% on average while maintaining the same levels of effectiveness in almost all datasets, with speedups of 1.37x on average. The framework scaled to large datasets, reducing them by up to 40% while statistically maintaining the same effectiveness with speedups of 1.70x. E2SC focused only on redundancy, however. **biO-IS**, in turn, extended E2SC, being capable of removing, besides redundancy, also noisy instances in up to 66.6%. biO-IS managed to significantly reduce the training sets (by 40.1% on average; up to 60%) while maintaining the same effectiveness levels in **all** of the considered datasets. biO-IS was also capable of consistently producing speedups of 1.67x on average (maximum of 2.46x). No baseline, not even E2SC, was able to achieve results with this level of quality, considering all tripod criteria. biO-IS improved over E2SC by 41% regarding reduction rate and from 1.42x to 1.67x (on average) regarding speedup, being the current state-of-the-art (SOTA) in Instance Selection applied to NLP.

6) Contributions and State-of-the-art advancement: In the Ph.D. dissertation, we conducted a rigorous comparative study of classical and SOTA IS methods applied to ATC. The study evaluated tradeoffs among reduction, effectiveness, and cost, motivated by the rising costs of new ATC solutions due to contextual embeddings, Transformer architectures, and increasing data volumes. Our findings show, contrary to common beliefs, Transformers often require representative - not large - data to perform well in ATC. Overall, IS techniques effectively reduced training set sizes without compromising effectiveness. The previous SOTA in IS fell short of meeting all tripod criteria simultaneously, underscoring the need for more efficient, scalable IS solutions, especially in big data scenarios. To address these challenges and fill the gaps found in the literature, we proposed two novel IS methods focused on redundancy (only) and noise (in conjunction with redundancy). Extensive experimental evaluation confirmed our hypotheses: *small and large language models can be trained with less data without sacrificing effectiveness*. This not only enables cost savings but also contributes to reducing carbon emissions. Such experimental evaluation established our solutions as the current SOTA IS applied to NLP. Such promising results instill hope for a more sustainable (green) and efficient NLP future, where advancements in IS techniques can produce environmental and economic benefits.

7) Scientific Production: Our work on IS has been published in the main **IR** and **NLP** venues:

1. **Cunha, Washington**, et al. "A Noise-Oriented and Redundancy-Aware Instance Selection Framework." **ACM Transactions on Information Systems** (ACM TOIS) 43.2 (2025): 1-33 – h5-index: 48.0
2. **Cunha, Washington**, et al. "A quantum annealing instance selection approach for efficient and effective transformer fine-tuning." International Conference on Theory of Information Retrieval **ICTIR'24**. p. 205-214 – h5-index: 24.0
3. **Cunha, Washington**, et al. "An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification." **ACM SIGIR** 2023. p. 665-674 – h5-index: 103.0.
4. **Cunha, Washington**, et al. "A Comparative Survey of Instance Selection Methods applied to Non-Neural and Transformer-Based Text Classification." **ACM Computing Surveys** 55.13s (2023): 1-52 – h5-index: 157.0
5. **Cunha, Washington**, et al. "On the cost-effectiveness of neural and non-neural approaches and representations for text classification:A comprehensive comparative study."**IP&M** 58.3 (2021): 102481 - h5-index: 114.0
6. **Cunha, Washington**, et al. "Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling." **IP&M** 57.4 (2020): 102263 – h5-index: 114.0

In addition, the work on my Ph.D. dissertation gave me the opportunity and the expertise to contribute as a co-author to several other published journal articles (8 in total), listed below:

1. Bittencourt, G., **Cunha, W** et al. (2025). On representation learning-based methods for effective, Review-Aware Recommender Systems (RARSs): Recent Advances, Experimental Comparative Analysis, Discussions, and New Directions. **ACM Computing Surveys**. - h5-index: 157; Imp. Fac.: 23.8
2. França, C., **Cunha, W** et al. (2024). On representation learning-based methods for effective, efficient, and scalable code retrieval. **Neurocomputing**. - h5-index: 136; Imp. Fac.: 5.5
3. Andrade, C., **Cunha, W** (2023) On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". **IP&M**. h5-index: 96; IF: 7.4
4. Zanotto, B. S., **Cunha, W** et al. (2021). Pcv50 automatic classification of electronic health records for a value-based program through machine learning. **Value in Health**. – h5-index: 57.0; IF: 4.9
5. Zanotto, B. S., **Cunha, W** et al. (2021). Stroke outcome measurements from electronic medical records: Cross-sectional study on the effectiveness of classifiers. **JMIR Med** – h5-idx: 52.0; IF:4.9
6. Viegas, F., **Cunha, W** et al. (2024). Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics.**Comp. Linguistics**- h5-index:38; IF:3.7
7. Felix, L. et al., **Cunha, W** (2024). Why are you traveling? Inferring trip profiles from online reviews and domain-knowledge. **Online Social Networks and Media**. – h5-index: 28.0; IF: 4.4
8. Viegas, F., **Cunha, W** et al. (2024). Pipelining semantic expansion and noise filtering for sentiment analysis of short documents – clusent method. **Journal on Interactive Systems**. – h5-index: 9.0

Ideas, insights, and methods of our dissertation also contributed to other 20 conference papers[1]:

1. Fonseca, G., **Cunha, W** et al. (2025). Instance-Selection-Inspired Undersampling Strategies for Bias Reduction in Small and Large Language Models for Binary Text Classification. **ACL'25**. – h5-index: 215.0
2. Viegas, F., **Cunha, W** et al. (2020). Cluhtm - semantic hierarchical topic modeling based on cluwords. Meeting of the Association for Computational Linguistics (ACL) **ACL'20**. – h5-index: 215.0
3. Mendes, L., **Cunha, W** et al. (2020). "Keep it Simple, Lazy"— Metalazy: A new metastrategy for lazy text classification. **CIKM'20**. – h5-index: 91.0
4. Viegas, F., **Cunha, W** et al. (2019). Cluwords: Exploiting semantic word clustering for enhanced topic modeling. **WSDM'19**. – h5-index: 77.0
5. Andrade et al., **Cunha, W** (2024) Explaining the hardest errors of contextual embedding-based classifiers. Conference on Computational Natural Language Learning (**CoNLL'24**). – h5-index: 39.0

Technical Production: We make the documented code of our proposed methods as well as all compared methods (IS and classifiers), datasets (and fold splits), available to the community for replication and comparisons on GitHub [2] for reproducibility and comparison of future IS and ATC methods.

8) Awards and Achievements: During the Ph.D., the student has received important awards:

1. 2^{nd} place in the Thesis and Dissertation Contest of the Brazilian Computer Society (**CTD-SBC'25**)
2. Best reviewer award on **SIGIR**'24 and **ARR ACL**'24 conferences;
3. **CTIC'24** and **CTIC'19**: co-advisor of undergraduate work ranked among the top ten Brazilian Scientific Initiation research selected by the Brazilian Computer Society
4. **SIGIR Student Travel Awards** to present a full research paper at SIGIR'23 in Taipei-Taiwan;
5. **Honorable Mention** in WFA – Tools and Applications Workshop – WebMedia'23;
6. **Honorable Mention** (2nd place) in the Master's Theses Contest (CTDBD) of the **SBBD'21**.

[1] We present only the top-5 most impactful papers based on h5-index due to space constraints. For the complete list, we refer the reader to `https://scholar.google.com.br/citations?hl=pt-BR&user=TiRmr48AAAAJ`

[2] All artifacts can be accessed on GitHub –`https://github.com/waashk/instanceselection`

Acknowledgements

This work was supported by CNPq, Capes, Fapemig, Fapesp, AWS, CIIA-Saúde, Google, and Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR; 408490/2024-1)

References

Cunha, W., Canuto, S., Viegas, F., Salles, T., Gomes, C., Mangaravite, V., Resende, E., Rosa, T., Gonçalves, M. A., and Rocha, L. (2020). Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management*, 57(4):102263.

Cunha, W., França, C., Fonseca, G., Rocha, L., and Gonçalves, M. A. (2023a). An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 665–674.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., et al. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*, 58(3):102481.

Cunha, W., Moreo Fernández, A., Esuli, A., Sebastiani, F., Rocha, L., and Gonçalves, M. A. (2025). A noise-oriented and redundancy-aware instance selection framework. *ACM Transactions on Information Systems*, 43(2):1–33.

Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., and Gonçalves, M. A. (2023b). A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM Computing Surveys*, 55(13s):1–52.

DeepSeek et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Garcia, S., Derrac, J., Cano, J., and Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Martins, K., Vaz de Melo, P., and Santos, R. (2021). Why do document-level polarity classifiers fail? In *Proceedings of the 2021 Conference of the NAACL: Human Language Technologies*.

Ng, A. (2016). Nuts and bolts of building ai applications using deep learning. *NIPS Keynote Talk*, 64.

Rajaraman, S., Ganesan, P., and Antani, S. (2022). Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PloS one*.

Roy, A. and Cambria, E. (2022). Soft labeling constraint for generalizing from sentiments in single domain. *Knowledge-Based Systems*, 245:108346.

Uppaal, R., Hu, J., and Li, Y. (2023). Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. *arXiv preprint arXiv:2305.13282*.