A Comprehensive Exploitation of Instance Selection Methods for Automatic Text Classification: "Doing More with Less"

Washington Cunha[1], **Leonardo Rocha**[2], **Marcos A. Gonçalves**[3]

[1] UNICAMP [2] UFSJ [3] UFMG

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

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.*

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.*

Institution where the degree was obtained: Universidade Federal de Minas Gerais

Student Level: Ph.D., concluded on August 26th, 2024

Board Members:

1. Franco Maria Nardini, ISTI-CNR, Itália.
2. Thierson Couto Rosa, UFG, Goiás, Brasil.
3. Rodrygo Luis Teodoro Santos, UFMG, Minas Gerais, Brasil.
4. Anisio Mendes Lacerda, UFMG, Minas Gerais, Brasil.

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

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

Google Scholar: `https://bit.ly/4lzu9Vj`

1. Introduction

The rapid growth of (textual) data 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 (e.g., web pages, emails, reviews, tweets, social media messages) into predefined semantic categories. Accurate ATC models have become crucial for many emerging applications [Cunha et al. 2023a] such as fake news detection, hate speech identification, relevance feedback, sentiment analysis, product review analysis, election vote inference, assessing public agency satisfaction, 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 (former Twitter), 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.

Transformer-based architectures, including Small and Large Language Models (SLMs and LLMs) such as RoBERTa and Llama 3 represent the state-of-the-art (SOTA) in ATC, achieving outstanding results through pre-training on large datasets and fine-tuning on domain-specific tasks. While zero-shot capabilities are feasible, fine-tuning remains essential for obtaining high performance (accuracy) [de Andrade et al. 2023].

According to Andrew Ng, the success of these (language) models is due to extensive pre-training on massive datasets (e.g., 45TB for GPT-3) and the adaptability of pre-trained (aka foundation) 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 fully training foundation models, it still requires significant computational power and time. For instance, fine-tuning XLNet (one of the SLMs considered in our work) in the MEDLINE dataset took 80 GPU hours.

Resource limitations in companies and research groups also restrict experimentation with such models. For instance, in this dissertation alone, we ran 4,000 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. Indeed, given increasing data volumes, re-training demands, and environmental concerns, proposing scalable and cost-effective 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 practitioners communities.

This dissertation focuses on this trade-off, one of the SBC 2025-2035 Grand Challenges on Computer Science[1], from a *data engineering perspective*, aiming to enhance model performance while reducing costs. In particular, we focus on **Instance Selection** (IS), a promising set of techniques and growing research area [Cunha et al. 2020, Cunha et al. 2021]. In contrast to traditional Feature Selection approaches, in which the main objective is to select the most informative terms (words), **Instance Selection** methods are focused on selecting the most representative instances (documents) for a

[1] https://www2.sbc.org.br/grandesdesafios/programacao/

training set [Garcia et al. 2012]. The intuition behind this type of algorithm 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 the number of instances by selecting the most representative ones; (ii) to maintain (or even improve) effectiveness by removing noise and redundancy; and (iii) to reduce the total time for applying an end-to-end model (from traditional preprocessing to model training). Thus, IS methods must respect three fundamental constraints consisting of *reducing the amount of training without loss of effectiveness and with efficiency gains* simultaneously.

IS has been understudied in NLP and ATC despite its enormous potential [Garcia et al. 2012]. Traditional IS methods developed for other domains include **CNN** and **Drop3**. Recent methods, such as **LSSm**, **LSBo**, and **PSDSP**, have mainly been tested on small tabular datasets with weak classifiers such as KNN. In contrast, text classification datasets are unstructured, larger, and more complex, featuring high dimensionality and skewness. The high computational costs of deep learning models with large training datasets present an ideal scenario for applying IS techniques.

2. Hypothesis and Research Questions

The main hypothesis (H1) of this Ph.D. dissertation is: **It is possible to reduce training data while simultaneously maintaining (language) model quality (effectiveness) and reducing time for fine-tuning ATC models through IS** To provide evidence for this hypothesis, we propose to answer **three** research questions in our Ph.D. dissertation:

RQ1. What is the impact of applying traditional IS methods in the ATC context regarding the posed constraints? RQ1 aims to evaluate traditional IS approaches applied to the ATC task, focusing on the **(tripod)** constraints of the posed IS methods: reduction, effectiveness, and efficiency.
RQ2. Can a novel instance selection method focused on redundancy removal overcome the limitations of existing IS methods to achieve the tripod constraints in ATC? Based on limitations found on the application of SOTA IS methods to ATC, this RQ aims to demonstrate the feasibility of proposing a novel IS framework focused on redundancy. We want to demonstrate how different requirements posed by distinct scenarios, mainly those associated with big data, can be accommodated.
RQ3. Is it possible to extend the previous proposal to not only remove redundancy but also remove noise and still meet all three tripod criteria? RQ3 focuses on extending the solution proposed to answer RQ2, by designing and testing a novel IS solution capable of simultaneously removing redundant and noisy instances from training.

3. Publications, Sub-Products, and Collaborations

Our work on Instance Selection[2] has been validated and published[3] in the main Intelligent Systems journals in the last four years, including **two** published papers in the *Information Processing and Management* **(IP&M)** [Cunha et al. 2021, Cunha et al. 2020]; a survey

[2]For reproducibility sake and further comparisons, we make the documented code of all methods, ours and the implemented baselines, as well as the preprocessed and raw datasets, including the division in folds, available to the community. **Code** at: `https://github.com/waashk/`

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

paper published in the *ACM Computing Surveys* **(CSUR)** [Cunha et al. 2023c], and a paper on the *ACM Transactions on Information Systems* **(TOIS)** [Cunha et al. 2024a], the most important IR journal. The dissertation also directly resulted in papers in important conferences, including: the *ACM Int. Conference on Research and Development in Information Retrieval* **(SIGIR)** [Cunha et al. 2023b] - the main worldwide IR conference - and the *ACM Int. Conf. on Theory of IR* **(ICTIR)** [Cunha et al. 2024b]. Table 1 shows a summary of the publications and products resulting directly from the Ph.D. dissertation.

Tabela 1. Summary of publication characteristics

Journal	FI[b]	h5-idx	Year	Qualis[c]	Citations[d]
Information Processing & Management (IP&M)	6.9	105	2020	A1	73
Information Processing & Management (IP&M)	6.9	105	2021	A1	122
ACM Computing Surveys (CSUR)	28.0	184	2023	A1	93
ACM Transactions on Information Systems (TOIS)	9.1	63	2024	A1	25
Conference	TA[a]	h5-idx	Year	Qualis[c]	Citations[d]
ACM SIGIR	20%	123	2023	A1	50
ACM ICTIR	50%	24	2024	A3	24

[a] Acceptance Rate, [b] Impact Factor, [c] Qualis 2021-2024, [d] Google Scholar accessed on July 30, 2026

Our dissertation has also indirectly (by inspiration or use of results) contributed to other **10** journal papers during the doctorate period, including: Scientific Reports [Ferreira et al. 2025], Computing Surveys [Bittencourt et al. 2025], Neurocomputing [França et al. 2024], IP&M [de Andrade et al. 2023], Value in Health [Zanotto et al. 2021a], JMIR Med Infor. [Zanotto et al. 2021b], Computational Linguistics [Viegas et al. 2024b], OSNEM [Félix et al. 2025a], Journal on Interactive Systems [Viegas et al. 2024a], and the Journal of the Brazilian Computer Society [de Andrade et al. 2026]. Table 2 shows a summary of these publications. Similar indirect influence of our dissertation can be found in the following **14** conference papers. Table 3 shows a summary of these publications.

Tabela 2. Summary of publication characteristics

Journal	FI[b]	h5-idx	Year	Qualis[c]	Citations[d]
Scientific Reports	3.9	234	2025	A1	4
ACM Computing Surveys (CSUR)	28.0	184	2025	A1	4
Neurocomputing	6.5	148	2024	A1	6
Information Processing & Management (IP&M)	6.9	105	2023	A1	35
Value in Health	6.0	59	2021	A1	3
JMIR Medical Informatics	6.9	59	2021	A2	17
Computational Linguistics	5.3	41	2024	A1	6
Online Social Networks and Media (OSNEM)	2.9	33	2024	A1	2
Journal on Interactive Systems (JIS)	-	13	2024	A4	2
Journal of the Brazilian Computer Society (JBCS)	-	17	2026	A2	0

[a] Acceptance Rate, [b] Impact Factor, [c] Qualis 2021-2024, [d] Google Scholar accessed on July 30, 2026

The combined h5-index of all above publications is **2805**. The respective papers have received so far more than **900** (according to Google Scholar[4]). The h-index of the PhD is 15, which is high for someone who just obtained his PhD title.

[4] `https://scholar.google.com.br/citations?user=TiRmr48AAAAJ&hl=pt-BR`

Tabela 3. Summary of publication characteristics

Conference	TA[a]	h5-idx	Year	Qualis[c]	Citations[d]
ACL [Viegas et al. 2020]	22.7%	236	2020	A1	59
ACL [Fonseca et al. 2025]	20.3%	236	2025	A1	4
ACL [Fonseca et al. 2026]	20.3%	236	2026	A1	0
SIGIR [França et al. 2025b]	20%	123	2025	A1	1
CIKM [Mendes et al. 2020]	20%	96	2020	A1	9
WSDM [Viegas et al. 2019]	16.4%	86	2019	A1	91
ECIR [Pasin et al. 2026a]	23.01%	47	2025	A2	7
CLEF [Pasin et al. 2026b]	-	44	2026	A4	0
CLEF [Pasin et al. 2025a]	-	44	2025	A4	8
CoNLL [Andrade et al. 2024]	41%	30	2024	A1	7
WebSci [Costa et al. 2025]	26%	30	2025	A2	10
CEUR Workshop [Pasin et al. 2025b]	-	27	2025	A4	5
ENIAC [Carvalho et al. 2024]	45.85%	9	2024	A4	0
SBRC [Félix et al. 2025b]	36.5%	8	2025	A4	0
WebMedia [Júnior et al. 2022]	31%	6	2022	A3	18
WebMedia [Viegas et al. 2023]	31%	6	2023	A3	2
WebMedia [Cecilio et al. 2023]	31%	6	2023	A3	5
WebMedia [Fonseca et al. 2024b]	40.8%	6	2024	A3	4
WebMedia [Silva et al. 2024]	40.8%	6	2024	A3	1
Webmedia [Carvalho et al. 2025]	31.3%	6	2025	A3	0
IIR Workshop [Pasin et al. 2022]	82%	8	2024	-	4
SBBD [Santos et al. 2023]	49%	4	2023	A4	2
SBBD [Cunha et al. 2025]	40%	4	2025	A4	0
SBBD [Prenassi et al. 2025]	40%	4	2025	A4	0
SBBD [França et al. 2025a]	40%	4	2025	A4	0
GDComp 2025-2035 [Gonçalves et al. 2025]	-	-	2024	-	1

[a] Acceptance Rate, [b] Impact Factor, [c] Qualis 2021-2024, [d] Google Scholar accessed on July 30, 2026

This dissertation led to several international collaborations with researchers, including (1) **Fabrizio Sebastiani**, **Andrea Esuli**, and **Alejandro Moreo** (ISTI-CNR Italy) in the proposal of our extended IS framework for ATC. [Cunha et al. 2024a] (**TOIS**); (2) **Nicola Ferro** (UniPd Italy), in an Quantum Annealing implementation **ICTIR** [Cunha et al. 2024b] of the IS proposal described in [Cunha et al. 2023b]; (3) **Davide Bacciu** (UniPi Italy) studying Continuous Learning with IS support for temporal ATC; (4) **Maurizio Ferrari Dacrema** and **Paolo Cremonesi** (PoliMi-Italy), helping with the organization the 2nd edition of the QuantumCLEF [Pasin et al. 2026b] to explore Quantum Annealing in IS applied to IR tasks, with participants accessing real quantum computers from CINECA, a non-profit consortium, made up of 69 Italian universities and 27 international research centers.

4. A Comparative Survey of Instance Selection Methods Applied to NonNeural and Transformer-Based Text Classification

We present here a summary of a critical analysis (*a.k.a., rapid (systematic-based) literature review* of the most traditional and/or recent (SOTA) Instance Selection proposals. This review (Chapter 2 of the dissertation, published at ACM CSUR [Cunha et al. 2023c]) aimed at comprehensively assessing the most relevant studies on IS strategies applied to different scenarios. In particular, we focused on experimentally oriented studies, that is, studies that had strong experimental and empirical components to support their findings.

At the end of this literature review process, we ended up with a mix of the traditional and SOTA methods, comprising 13 Instance Selection strategies to be evaluated in the ATC scenario. The results of our searches and analyses reinforced our perception that IS methods were almost exclusively applied to tabular structured data. Their application to NLP tasks was very rare, which is odd since this is one of the areas that could benefit most from this type of method. In sum, of the 100 selected IS papers, 92 of them considered just tabular data. We proposed investigating the use of IS methods along with ATC models, which are paramount in many applications, as discussed.

As a contribution, we proposed a new taxonomy of IS strategies by extending a 13-year-old taxonomy [Garcia et al. 2012]. This previous IS taxonomy was proposed in a different context, considering different dataset types (mainly small tabular ones) and learning methods. Deep learning neural network methods were not even considered by the time that taxonomy was proposed. The field has significantly evolved since then. These advances were carefully documented in our new extended taxonomy, with three new categories and eight new, recently proposed methods. The new categories cover approaches based on density, spatial hyperplanes, and clustering-based strategies proposed from 2015 on.

In our comparative study, we assessed the trade-off among: (i) reduction, (ii) efficiency, and (iii) effectiveness of 13 representative IS methods applied to ATC, using large and varied datasets. The selected IS methods have been previously tested only with small structured tabular datasets. For this comparison, we considered SOTA ATC methods by the time the comparison was done (around 2022) including BERT, XLNet, RoBERTa, and other Transformer-Based classifiers. These methods have a high computational cost, mainly when dealing with large labeled training data. As such, they constituted an ideal scenario for IS application. Lastly, our work was the first to apply IS as a preprocessing step before using transformer-based architectures in the ATC context. This contribution was performed using an experimental setup whose rigor and magnitude (7 transformer methods, 13 IS approaches, 22 datasets) had not been previously reported in IS literature. The comparative scenario and experimental setup used to assess the trade-off among reduction, efficiency, and effectiveness of IS applied to the ATC is described below.

Datasets, Data Representation, and Preprocessing We used 22 real-world datasets collected from various sources in two broad ATC tasks [Li et al. 2022]: i) *topic classification*; and ii) *sentiment analysis*. The datasets covered a range of domains, diversity in size, dimensionality, classes, document density, and class distributions. The TFIDF representation was input to all IS methods. We tested other text representations, such as *static embeddings* (e.g., FastText and *contextual embeddings* built by transformer architectures (whether by forwarding documents through fine-tuned model or in a zero-shot approach). Static embeddings slowed down classification methods significantly, which is supported by [Andrade et al. 2024] while using contextualized embeddings directly as IS input was inefficient and ineffective, probably due to high dimensionality.

We removed stopwords and kept features appearing in at least two documents. We normalized the TF-IDF product result using the L2-norm. As experimental setup, we split the dataset employing a stratified k-fold cross-validation – k=10-fold partition for the smaller datasets, while for the larger ones we adopted 5 folds due to the cost. Then we constructed a TFIDF document representation matrix for the IS stage. Finally, we used the corresponding raw documents as input for the Transformers classifiers.

Text Classification Methods We compared the effectiveness among the Transformers – *RoBERTa*, *BERT*, *DistilBERT*, *BART*, *AlBERT*, and *XLNet*. We applied the methodology from [Cunha et al. 2020, Cunha et al. 2021] to determine optimum hyperparameters.

Instance Selection Methods We considered a set of 13 IS methods: CNN, ENN, ICF, IB3, Drop3, LSSm, LSBo, LDIS, CDIS, XLDIS, PSDSP, EGDIS, and CIS. Parameters were defined with grid-search using cross-validation during initial empirical experiments.

Metrics and Experimental Protocol All experiments were executed on an Intel Core i7-5820K, 64Gb RAM, and a TITAN X (12GB) and Ubuntu 19.04. Due to dataset skewness, we evaluated classification effectiveness using MacroF1. We employed the paired t-test with a 95% confidence level with Bonferroni correction to account for multiple tests. We consider reduction mean by defined as $\overline{R} = \frac{\sum_{i=0}^{k} \frac{|T_i|-|S_i|}{|T_i|}}{k}$, where T is the original training set, and S is the solution set containing the selected instances by the IS method being evaluated. Last, in order to analyze the cost-effectiveness tradeoff, we also evaluated each method's cost in terms of the total time required to build the model. Speedup is calculated as $S = \frac{T_{wo}}{T_w}$, where T_w is the total time spent on model construction using the IS approach, and T_{wo} is the total time spent on execution without the IS phase.

Experimental Results: Our findings revealed a significant potential of IS applied to ATC. Some IS methods could reduce the training set by up to 90% while maintaining effectiveness in some datasets. The IS approaches we studied achieved reductions between 15% (LSSm) to 91% (XLDIS). However, despite the motivation for noise removal, the tested IS methods were not able to improve the ATC effectiveness in any of the tested datasets. This underscored the need for further studies to investigate this issue.

We demonstrated that three traditional IS methods **(LSSm, CNN, LSBo)** were able to reduce the total classification models' construction time while keeping the effectiveness in 12 (out of 19) considered datasets – with speedups between 1.04x and 5.69x. In the other datasets, we observed that the introduction of IS caused an overhead in terms of total time for model generation (running the IS methods + model construction), making the whole process more costly. Overall, considering the three tripod constraints altogether and all datasets, the best IS method was **CNN**. Our evaluation regarding the three constraints showed that, in some datasets, specific selection methods can reduce the training set without loss of effectiveness and with efficiency improvements. **However, our experiments revealed that no IS baseline could respect all tripod constraints in all cases.** Our results motivated further investigations on exploiting IS for ATC, especially regarding Transformers, which led to the proposal of the two new methods developed and tested in our dissertation, as described next.

Final Remark: Results summarized in this Section played a crucial role in producing an impactful publication in the *ACM Computing Surveys* **(CSUR)** [Cunha et al. 2023c].

5. An Effective, Efficient, and Scalable Confidence-Based Instance Selection Framework for Transformer-Based Text Classification

In the previous section, we found that no method in the literature was able to meet all IS constraints in all cases. Additionally, in some scenarios, using these methods increased the time it takes to construct the model, which goes against IS main goals. In this context, one of the main contributions of the PhD dissertation was the proposal of **E2SC**, a novel

two-step IS framework aimed at large datasets with a special focus on transformer-based architectures. E2SC is a technique that satisfies the tripod´s constraints and is very suitable for real-world scenarios, including datasets with thousands of instances.

E2SC is comprised of two main steps. E2SC's **first** step aims to assign a probability to each instance being removed from the training set (α parameters). We adopt an exact KNN model solution[5] to estimate the probability of removing (trauning) instances, as KNN is considered a **calibrated**[6] [Rajaraman et al. 2022] and computationally cheap (fast) classifier. Our **first hypothesis (H1)** is 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 other words, if a calibrated (weak) classifier C classifies a training instance x with high confidence, there is a high probability that the information contained in x is already present in other training documents used to build C, and x can be safely removed from the training set for the sake of building a stronger (Transformer) model. Accordingly, we keep in the training the hard-to-classify instances - probably located in the decision border regions -, weighted by confidence for the next step, in which we partially remove only the easiest ones.

As the **second** E2SC step, we proposed to estimate a near-optimal reduction rate (β parameter) that does not degrade the Transformer's effectiveness by employing a validation set and a weak but fast classifier. Our **second hypothesis (H2)** is that *we can estimate the effectiveness of a robust model through the analysis and variation of selection rates in a weaker model*. Again, we explore KNN for this. More specifically, we introduce an iterative method that statistically compares, 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 can estimate a reduction rate that does not affect the KNN model's effectiveness. Last, considering the output of these two steps together, $\beta\%$ instances are randomly sampled, weighted by the α distribution, to be removed from the training set.

Experimental Results: Our experimental evaluation showed that **E2SC** managed to significantly reduce the training sets (by up to 60% – **27%** on average) while maintaining the same effectiveness levels in 18 (out of 19) considered datasets. We also found that **E2SC** was able to reduce ATC models' total construction time with speedups of **1.25x** on average, varying between 1.02x and 2.04x. Overall, considering the three tripod constraints altogether and all datasets, the best IS method was **E2SC**.

Finally, to demonstrate the flexibility of our framework to cope with large datasets, we proposed two modifications: (i) a two-rule heuristic-based β parameter for fast reduction estimation and (ii) the adoption of an approximated KNN for fast nearest neighbor search. Our enhanced solution managed to increase the reduction rate of the training sets (to **29%** on average) while maintaining the same levels of effectiveness in **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**.

Final Remark: A detailed description of **E2SC** (Chapter 4), along with the experimental results, was published in a conference paper in the *ACM SIGIR* [Cunha et al. 2023b].

[5] In our dissertation, we test and confirm that exact KNN is a reasonable proxy for redundancy.

[6] A calibrated classifier is one whose probability class predictions correlate well with its accuracy, e.g., for those instances predicted with 80% of confidence, the classifier is correct in roughly 80% of the cases.

6. An Extended Noise-Oriented and Redundancy-Aware Instance Selection Framework for Transformer-Based Automatic Text Classification

Although being able to achieve significant results regarding the trade-off effectiveness-efficiency-reduction, E2SC focused only on **redundancy**, leaving other aspects that may help to further reduce training untouched. One such aspect is **noise**, defined as instances incorrectly labeled by humans [Martins et al. 2021] as well as outliers that do not contribute (or even get in the way) to model learning. Indeed, in a simulated scenario designed to evaluate the capability of the IS baseline methods and **E2SC** to remove noise, none of the IS solutions satisfactorily performed noise removal (Chapter 5.1 of the dissertation).

Accordingly, another dissertation´s main contribution was the proposal of **biO-IS**, built on top of **E2SC**, aimed at simultaneously removing redundant and noisy instances. Our extended IS framework encompasses 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 as the calibrated weak classifier instead of KNN. An in-depth comparative analysis of several possibilities (Chapter 5.4 of the dissertation), demonstrated LR as the best classifier in terms of effectiveness-calibration-cost.

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 of this second step 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 the entropy function as a proxy to determine the reduction behavior caused by these instances for the sake of training a strong ATC model.

The hypothesis **(H3)** behind this new step is that *when the prediction provided by the calibrated weak classifier is incorrect, the entropy of the posterior probabilities negatively correlates with the classifier's confidence.* Low entropy occurs when the classifier assigns an instance with absolute certainty to a wrong class, while high entropy occurs when the classifier is uncertain among several classes. Therefore, we consider the chance of a noisy instance being removed as the inverse of the entropy of the prediction. *When an incorrect prediction is accompanied by low entropy, it is more likely to be removed; otherwise, it is more likely to be kept in the training set.*

Results: We compared **biO-IS** with seven of the most robust SOTA IS baseline methods, including our own E2SC, in the ATC domain considering **22** datasets. Our experimental evaluation revealed that, **biO-IS** was capable of satisfactorily removing noise in up to 66.6%. **biO-IS** managed to significantly reduce the training sets (by **40.1%** on average; varying between 29% and 60% of reduction) while maintaining the same effectiveness levels in **all** of the considered datasets. **biO-IS** was also capable of consistently producing speed-ups 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. Indeed, the only other method capable of maintaining effectiveness on all datasets was E2SC. **biO-IS** improved over E2SC by 41% regarding reduction rate and from 1.42x to 1.67x (on average) regarding speedup, being the current SOTA in Instance Selection applied to NLP.

Final Remark: Expanded results described in this section were published in a long journal paper in the *ACM Transactions on Information Systems* [Cunha et al. 2024a].

7. Conclusion and Future Work

In the Ph.D. dissertation, we conducted a rigorous comparative study of classical and state-of-the-art 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. Main findings based on over 5,000 experiments include: (i) existing IS methods rarely improved ATC model effectiveness; (ii) among 13 tested IS methods, LSSm, CNN, and LSBo performed best, achieving significant reductions (46.6% average) while maintaining effectiveness in 12 out of 22 datasets. 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. However, traditional IS approaches often 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 **E2SC**, a redundancy-oriented two-step IS framework designed for large datasets and transformer-based architectures, introducing: (i) calibrated weak classifiers to estimate data usefulness during transformer training; and (ii) iterative processes and heuristics to determine optimal reduction rates. E2SC reduced training sets by 29% on average while maintaining effectiveness across 22 datasets, achieving speedups of up to 70%. To overcome E2SC limitations regarding noise handling, we developed **biO-IS**, an extended framework that simultaneously removes both, *redundant* and *noisy* instances. Experimental results showed biO-IS reduced training sets by 41% on average (up to 60%), removed 66.6% of manually inserted noise, and maintained effectiveness across all datasets. Furthermore, it provided consistent speedups (1.67x average, up to 2.46x), outperforming E2SC in reduction and efficiency.

These findings confirm our main hypothesis: **small and large language models can be trained with less data without sacrificing effectiveness**. This not only enables cost and energy savings but also contributes to reducing carbon emissions (up to 65%). Such promising results instill hope for a more sustainable and efficient NLP future, where advancements in IS techniques can produce environmental and economic benefits.

Our PhD dissertation opens up several opportunities for future investigations. We have already started developing novel Quantum Annealing (QA) IS-supported approaches [Cunha et al. 2024b], where there is still a lot of room for improvements. Evaluating our proposed IS frameworks on other SLM and LLM models, such as DeepSeek, is one of these opportunities. Extending the application of our IS solutions to other IR and NLP tasks, such as ranking, recommendation, question answering, summarization, and topic modeling, is another. So far, we have measured the importance of a whole document for the training (tuning) process, but in the future, we could do this at a more fine-grained level, such as for reviews' aspects and documents' passages. We also plan to investigate using our IS solutions to build large (foundation) language model in Portuguese from scratch more efficiently. We have already started investigating IS as a way to mitigate issues of imbalance (skewness) and fairness in NLP, but there still is a lot to be done in this investigation line. Last, considering the vast number of dimensions in textual datasets, assessing how *feature selection* methods interact with IS can be very interesting.

Acknowledgments This work was supported by CNPq, CAPES, INCT-TILDIAR, AWS, FAPEMIG, Google, Finep, FAPESP and FAEPEX.

Referências

[Andrade et al. 2024] Andrade, C., Cunha, W., Fonseca, G., Pagano, A., Santos, L., Pagano, A., Rocha, L., and Gonçalves, M. (2024). Explaining the hardest errors of contextual embedding based classifiers. In *Proceedings of the 28th Conference on Computational Natural Language Learning*.

[Bittencourt et al. 2025] Bittencourt, G., Vasconcelos, N., Andrade, Y., Silva, N., Cunha, W., Colombo Dias, D. R., Gonçalves, M. A., and Rocha, L. (2025). Review-aware recommender systems (rarss): Recent advances, experimental comparative analysis, discussions, and new directions. *ACM Comput. Surv.*

[Carvalho et al. 2025] Carvalho, D., Pereira, A., Cunha, W., Tuler, E., Dias, D., and Rocha, L. (2025). Uma ferramenta para organização semântica de informações de helpdesk por meio de modelagem de tópicos. In *Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, pages 155–158. SBC.

[Carvalho et al. 2024] Carvalho, D., Pereira, A., Tuler, E., Dias, D., Cunha, W., and Rocha, L. (2024). Adapting large language models for topic modeling tasks. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 954–965. SBC.

[Cecilio et al. 2023] Cecilio, P., Pereira, A., Viegas, F., Rosa, J., Cunha, W., Testa, F., Tuler, E., and Rocha, L. (2023). Um framework para extração automática de informações em patentes farmacêuticas. In *Anais Estendidos do XXIX Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 97–100. SBC.

[Costa et al. 2025] Costa, J., Oliveira, G., Fonseca, G., Reis, D., Oliveira Teixeira, G., Cunha, W., Rocha, L., and Ferreira, C. H. (2025). Characterizing youtube's role in online gambling promotion: A case study of fortune tiger in brazil. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 42–51.

[Cunha et al. 2020] 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. *IP&M*, 57(4):102263.

[Cunha et al. 2023a] Cunha, W., França, C., Rocha, L., and Gonçalves, M. A. (2023a). Tpdr: A novel two-step transformer-based product and class description match and retrieval method. *arXiv preprint arXiv:2310.03491*.

[Cunha et al. 2023b] Cunha, W., França, C., Fonseca, G., Rocha, L., and Gonçalves, M. A. (2023b). An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23.

[Cunha et al. 2021] 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*.

[Cunha et al. 2024a] Cunha, W., Moreo, A., Esuli, A., Sebastiani, F., Rocha, L., and Gonçalves, M. (2024a). A noise-oriented and redundancy-aware instance selection framework. *ACM Transactions on Information Systems*.

[Cunha et al. 2024b] Cunha, W., Pasin, A., Goncalves, M., and Ferro, N. (2024b). A quantum annealing instance selection approach for efficient and effective transformer fine-tuning. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*.

[Cunha et al. 2025] Cunha, W., Rocha, L., and Gonçalves, M. A. (2025). Inteligência artificial sustentável baseado em engenharia de dados, aprendizado de máquina e transferência de conhecimento para processamento de linguagem natural. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 208–212. SBC.

[Cunha et al. 2023c] Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., and Gonçalves, M. A. (2023c). A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM Comput. Surv.*

[de Andrade et al. 2023] de Andrade, C. M., Belém, F. M., Cunha, W., França, C., Viegas, F., Rocha, L., and Gonçalves, M. A. (2023). On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". *IP&M*, 60(4):103336.

[de Andrade et al. 2026] de Andrade, C. M. V., Cunha, W., Reis, D., França, C., Apolinário, W. Á., de Castro Santos, L., Pagano, A. S., da Rocha, L. C. D., and Gonçalves, M. A. (2026). "call my big sibling (cmbs)--a confidence-based strategy leveraging instance selection to combine small and large language models for cost-effective text classification. *Journal of the Brazilian Computer Society*.

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

[Félix et al. 2025a] Félix, L. G., Cunha, W., de Andrade, C. M., Gonçalves, M. A., and Almeida, J. M. (2025a). Why are you traveling? inferring trip profiles from online reviews and domain-knowledge. *Online Social Networks and Media*, 45:100296.

[Félix et al. 2025b] Félix, L. G., Cunha, W., Xavier, C. R., de Melo, P. O. V., Gonçalves, M. A., and Almeida, J. M. (2025b). Geração de roteiros turísticos personalizados em tempo real através de heurística grasp adaptada. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, pages 196–209. SBC.

[Ferreira et al. 2025] Ferreira, A. A., Rocha, L., Cunha, W., Machado, A. C., Campos, J. M., Jallais, G., Viana, A. C., Tuler, E., Araújo, I., Macul, V., et al. (2025). A comprehensive qualitative analysis of patient dialogue summarization using large language models applied to noisy, informal, non-english real-world data. *Scientific Reports*.

[Fonseca et al. 2026] Fonseca, G., Cunha, W., Prenassi, G., Da Rocha, L. C. D., and Gonçalves, M. A. (2026). When high accuracy hides poor calibration: Rethinking confidence evaluation in transformer-based text classification with balanced brier score. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[Fonseca et al. 2025] Fonseca, G., Cunha, W., Prenassi, G., Gonçalves, M. A., and Da Rocha, L. C. D. (2025). Instance-selection-inspired undersampling strategies for bias

reduction in small and large language models for binary text classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9323–9340.

[Fonseca et al. 2024a] Fonseca, G., Cunha, W., and Rocha, L. (2024a). Análise comparativa de métodos de undersampling em classificação automática de texto baseada em transformers. *Revista Eletrônica de Iniciação Científica em Computação*, 22(1).

[Fonseca et al. 2024b] Fonseca, G., Prenassi, G., Cunha, W., Gonçalves, M. A., and Rocha, L. (2024b). Estratégias de undersampling para redução de viés em classificação de texto baseada em transformers. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web*.

[França et al. 2025a] França, C., Nunes, I., Salles, T., Cunha, W., Jallais, G., Rocha, L., and Gonçalves, M. A. (2025a). Muitas classes desbalanceadas? não classifique-ranqueie! uma abordagem baseada em retrieval-augmented generation (rag)-labels para classificação textual multi-classe. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 264–277. SBC.

[França et al. 2025b] França, C., Rabbi, G., Salles, T., Cunha, W., Rocha, L., and André Gonçalves, M. (2025b). Optimizing tail-head trade-off for extreme multi-label text classification (xmtc) with rag-labels and a dynamic two-stage retrieval and fusion pipeline. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1392–1401.

[França et al. 2024] França, C., Lima, R. C., Andrade, C., Cunha, W., de Melo, P. O. S. V., Ribeiro-Neto, B., Rocha, L., Santos, R. L. T., Pagano, A. S., and Gonçalves, M. A. (2024). On representation learning-based methods for effective, efficient, and scalable code retrieval. *Neurocomputing*.

[Garcia et al. 2012] 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*.

[Gonçalves et al. 2025] Gonçalves, M. A., Rocha, L., Cunha, W., and Dal Bianco, G. (2025). Mais com menos - processamento de linguagem natural inteligente e sustentável baseado em engenharia de dados e inteligência artificial avançada. *Grandes Desafios da Computaçao no Brasil 2025-2035*.

[Júnior et al. 2022] Júnior, A. P. D. S., Cecilio, P., Viegas, F., Cunha, W., Albergaria, E. T. D., and Rocha, L. C. D. D. (2022). Evaluating topic modeling pre-processing pipelines for portuguese texts. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 191–201.

[Li et al. 2022] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41.

[Martins et al. 2021] 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*.

[Mendes et al. 2020] Mendes, L. F., Gonçalves, M. A., Cunha, W., da Rocha, L. C., Rosa, T. C., and Martins, W. (2020). "keep it simple, lazy- metalazy: A new metastrategy

for lazy text classification. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1125–1134. ACM.

[Pasin et al. 2022] Pasin, A., Cunha, W., Gonçalves, M. A., Ferro, N., et al. (2022). A quantum annealing-based instance selection approach for transformer fine-tuning. In *the 14th Italian Information Retrieval Workshop*.

[Pasin et al. 2026a] Pasin, A., Dacrema, M. F., Cremonesi, P., Cunha, W., Gonçalves, M. A., and Ferro, N. (2026a). Quantumclef 2026 the third edition of the quantum computing lab at clef. In *European Conference on Information Retrieval*, pages 306–314. Springer.

[Pasin et al. 2026b] Pasin, A., Dacrema, M. F., Cremonesi, P., Cunha, W., Gonçalves, M. A., and Ferro, N. (2026b). Quantumclef 2026 the third edition of the quantum computing lab at clef. In *European Conference on Information Retrieval*, pages 306–314. Springer.

[Pasin et al. 2025a] Pasin, A., Ferrari Dacrema, M., Cunha, W., André Gonçalves, M., Cremonesi, P., and Ferro, N. (2025a). Overview of quantumclef 2025: The second quantum computing challenge for information retrieval and recommender systems at clef. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 412–435. Springer.

[Pasin et al. 2025b] Pasin, A., Ferrari Dacrema, M., Cunha, W., Gonçalves, M. A., Cremonesi, P., Ferro, N., et al. (2025b). Quantumclef 2025: Overview of the second quantum computing challenge for information retrieval and recommender systems at clef. In *CEUR WORKSHOP PROCEEDINGS*, volume 4038, pages 4086–4104. CEUR-WS.

[Prenassi et al. 2025] Prenassi, G. S., Fonseca, G., Reis, D., Cunha, W., Gonçalves, M. A., and Rocha, L. (2025). Um estudo comparativo de estratégias de seleção de exemplos para in-context learning aplicado à classificação automática de texto com grandes modelos de linguagem. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 921–927. SBC.

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

[Santos et al. 2023] Santos, W., Cunha, W., França, C., Fonseca, G., Canuto, S., Rocha, L., and Gonçalves, M. (2023). Uma metodologia para tratamento do viés da maioria em modelos de stacking via identificação de documentos difíceis. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 408–413. SBC.

[Silva et al. 2024] Silva, N., Reis, D., Cunha, W., Tuler, E., Silva, T., Silva, N., , and Rocha, L. (2024). Integrando avaliações textuais de usuários em recomendação baseada em aprendizado por reforço. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web*.

[Uppaal et al. 2023] 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*.

[Viegas et al. 2023] Viegas, F., Canuto, S., Cunha, W., França, C., Valiense, C., Rocha, L., and Gonçalves, M. A. (2023). Clusent–combining semantic expansion and denoising for dataset-oriented sentiment analysis of short texts. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, pages 110–118.

[Viegas et al. 2024a] Viegas, F., Canuto, S., Cunha, W., França, C., Valiense, C., Fonseca, G., Machado, A., Rocha, L., and Gonçalves, M. (2024a). Pipelining semantic expansion and noise filtering for sentiment analysis of short documents – clusent method. *Journal on Interactive Systems*, 15(1).

[Viegas et al. 2019] Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of WSDM '19*, pages 753–761.

[Viegas et al. 2020] Viegas, F., Cunha, W., Gomes, C., Pereira, A., da Rocha, L. C., and Gonçalves, M. A. (2020). Cluhtm - semantic hierarchical topic modeling based on cluwords. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8138–8150. Association for Computational Linguistics.

[Viegas et al. 2024b] Viegas, F., Pereira, A., Cunha, W., França, C., Andrade, C., Tuler, E., Rocha, L., and Gonçalves, M. A. (2024b). Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics. *Computational Linguistics*, pages 1–41.

[Zanotto et al. 2021a] Zanotto, B., Etges, A., Dal Bosco, A., Cortes, E., Ruschel, R., Martins, S., Souza, A., Valiense, C., Viegas, F., Canuto, S., et al. (2021a). Pcv50 automatic classification of electronic health records for a value-based program through machine learning. *Value in Health*, 24:S76.

[Zanotto et al. 2021b] Zanotto, B. S., Beck da Silva Etges, A. P., dal Bosco, A., Cortes, E. G., Ruschel, R., De Souza, A. C., Andrade, C. M. V., Viegas, F., Canuto, S., Luiz, W., Ouriques Martins, S., Vieira, R., Polanczyk, C., and André Gonçalves, M. (2021b). Stroke outcome measurements from electronic medical records: Cross-sectional study on the effectiveness of neural and nonneural classifiers. *JMIR Med Inform*, 9(11):e29120.

Appendix: Sub-Products

A. Impact Statistics

Figure 1 shows the Google Scholar statistics of the Ph.D. holder [7]. His h-index is **16** – high for someone who just obtained his PhD title – and total number of citations is **944**. In particular, the IP&M paper [Cunha et al. 2021] currently has **122** citations. The combined h5 index of all publications during the Ph.D. period is **2805**.

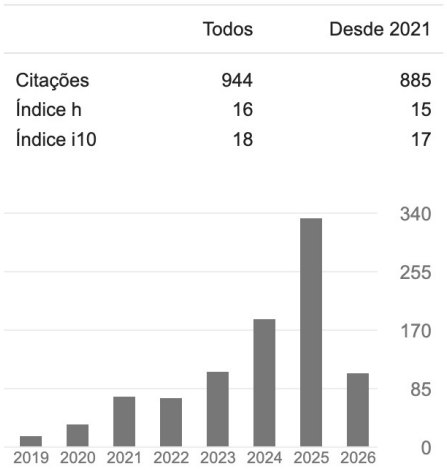


Figura 1. Google Scholar statistics accessed on July 30, 2026.

B. Additional Collaborations and Contributions

The PhD dissertation led to several international collaborations, notably including with: (1) **Fabrizio Sebastiani**, **Andrea Esuli**, and **Alejandro Moreo** (ISTI-CNR Italy) investigating IS strategies through Feature Selection and the proposal of the **biO-IS** framework for ATC. [Cunha et al. 2024a] (**TOIS**); (2) **Nicola Ferro** (UniPd Italy), with whom we had a full paper published in the *International Conference on the Theory of Information Retrieval* (**ICTIR**) [Cunha et al. 2024b], introducing an innovative Quantum Annealing IS approach and the proposal of a new CLEF task on Quantum IS [Pasin et al. 2026b]; (3) **Davide Bacciu** (UniPi Italy) studying Continuous Learning with Instance Selection support for temporal ATC; (4) **Maurizio Ferrari Dacrema** and **Paolo Cremonesi** (PoliMi-Italy) organizing the 2nd edition of the QuantumCLEF to explore the effectiveness of Quantum Annealing in IR with participants accessing real quantum computers from CINECA, a non-profit consortium composed of 69 Italian universities and 27 international public research centers.

The PhD work also allowed us to co-advise undergraduate students. We highlight the work presented in [Fonseca et al. 2024a], presented in the Scientific Initiation Competition (CTIC-CSBC) covering the proposal of IS-inspired **Undersampling** strategies for bias reduction in the context of transformer-based ATC – work co-advised by the PhD holder. An extension of this work was published on ACL 2025 Conference[Fonseca et al. 2025].

[7] https://scholar.google.com.br/citations?user=TiRmr48AAAAJ

For sake of reproducibility, we made the documented code of all compared methods as well as the preprocessed and raw datasets, including the division in folds, available[8] to the community for replication and further comparisons.

C. Contributions as Advisor

Our thesis also resulted in human resources, as I also served as advisor and co-advisor for master's theses, final papers, and research initiations.

Master's Thesis Co-Advisor

1. (2025-Atual) Gabriel Catizani Faria Oliveira (UFMG) - Rotulação Automática com Grandes Modelos de Linguagem
2. (2025-Atual) Guilherme do Nascimento (UFMG) - Instance-Selection-Inspired Undersampling Strategies for Bias Reduction in Small and Large Language Models for Binary Text Classification. ACL (2025).
3. (2025-Atual) Arthur Bittencourt Vasconcelos (UFF) - Copia, mas quão igual?
4. (2023-Atual) Alexandre Cassimiro (UFMG) - IA e Prontuário Médico.
5. (2024-2025) Daniel Carvalho (UFSJ) - Modelagem de Tópicos como Ferramenta na Predição de chamadas Help Desk.

Advisor (or co-advisor) of undergraduate POC and TCC and Scientific Initiation

6. (2024-2025) Marcos Rezende (UFMG)
7. (2024-2025) Igor Eduardo Martins Braga (UFMG)
8. (2022-2023) Thiago Morais Araújo (UFMG)
9. (2022-2024) Guilherme do Nascimento (UFSJ)
10. (2018-2019) Christian Reis Fagundes Gomes (UFSJ)

In particular, Guilherme's research – which the Ph.D. student co-supervise as part of his master's thesis and which was recently accepted at ACL, the premier international conference in Computational Linguistics – proposes novel solutions to open challenges directly connected to the core issues addressed in my Ph.D. dissertation.

D. SBC Great Challenges in Computing - 2025 - 2035

The student is one of the lead proponents of one of the Great Challenges in Computing hosted by the Brazilian Computer Society (SBC) [Gonçalves et al. 2025], centered on Intelligent and Sustainable Natural Language Processing, leveraging Data Engineering and Advanced Artificial Intelligence. This initiative is strongly aligned with WebMedia's objectives and with the United Nations' 2030 Sustainable Development Goals, particularly in promoting inclusive and responsible technology. Finally, our dissertation is also strongly aligned with the Brazilian Artificial Intelligence Plan (PBIA) [9].

E. Dissemination of research in interviews, TV programs, and invited talks

During the Ph.D., the student research has been disseminated through interviews, TV programs, and invited talks, contributing to public engagement and academic discussions:

1. Palestra na UFV - Sustainable and Responsible AI based on Data Engineering for NLP.

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

[9] https://www.gov.br/lncc/pt-br/assuntos/noticias/ultimas-noticias-1/plano-brasileiro-de-inteligencia-artificial-pbia-2024-2028

2. Palestra no IC/Unicamp - IA Sustentável Baseada em Engenharia de Dados para PLN.
3. Mesa-redonda sobre Inteligência Artificial no IV ENAEPE - Encontro Anual de Ensino, Pesquisa, Extensão e Inovação da UEA
4. UFMG cria projeto inédito sobre Inteligência Artificial responsável `CancaoNova`
5. Inteligência artificial faz parte do dia a dia de muitas pessoas. (2023) `Record-MG`
6. MG Tec: destaques tecnológicos, pesquisas e inovações. (2022) `MG-Tech/MGTV`

F. Advisory and Consulting Activities

During my PhD studies, I also engaged in advisory and consulting activities, notably as a Data Science consultant for the Center for Management and Strategic Studies (CGEE), contributing to its Observatory of Science, Technology and Innovation (OCTI). In this role, I conducted a pioneering study on the profile of Brazilian researchers holding CNPq Productivity in Research (PQ) fellowships (levels 1A, 1B, and 1C) over a 12-year period. The results were published in CGEE's annual bulletin and established a new strategic monitoring line focused on human competencies within the national science, technology, and innovation system. My work involved the methodological design, analysis of Lattes Platform data, and interpretation of findings, demonstrating my ability to transform complex scientific data into strategic insights for public policy and national research planning.

G. Highlighted Project

During my PhD, I played a central role in the "Amplify Project - Google/DCC - Development of Datasets for Generative Artificial Intelligence Applications" a collaborative initiative between Google and Federal University of Minas Gerais (UFMG) aimed at creating datasets for generative AI applications. I was recognized as the facilitator of this project, identifying synergies between the research interests of UFMG and Google and helped establish the partnership. My contribution was also essential in formalizing the collaboration and structuring the project as sub-coordinator. The initiative secured funding of US$100,000 (R$627,000 in December 2024), which supports research scholarships for undergraduate, graduate students, and faculty members involved in the project.

H. Program Committee Member and Technical Reviewer

During my PhD, I actively served as a program committee member and technical reviewer for leading journals and conferences in Artificial Intelligence, reflecting my commitment to scientific rigor and excellence. I acted as a regular reviewer for major venues such as ACM SIGIR, ACM RecSys, NeurIPS, The Web Conference, ACL, EMNLP, AAAI Conference on Artificial Intelligence, ECIR, and journals such as Information Sciences and Expert Systems, among others. My recognition with Best Reviewer Awards at both ACM SIGIR 2024 and ACL 2024 highlights my critical evaluation skills and sustained contribution to advancing high-quality research within the international AI community.