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. *Progress in Natural Language Processing (NLP) has been dictated by the "rule of more": more data, more computing power and more complexity, best exemplified by the current Large Language Models (LLMs). Indeed, to properly work (with high accuracy) for (domain-)specific applications, these LLMs have to be fine-tuned, i.e., trained with domain-specific data, which usually requires significant amounts of computational (and natural) resources. This* ***Ph.D. dissertation*** *focuses on a data engineering technique under-investigated in NLP, whose potential is enormous in the current data-intensive scenario, known as* ***Instance Selection (IS)****. The IS goal is to reduce the training set size by removing noisy or redundant training instances while maintaining the effectiveness of the trained models, thus reducing the training process costs. In the PhD dissertation, we provide a comprehensive and scientifically sound comparison of many state-of-the-art (SOTA) IS methods applied to an essential NLP task – Automatic Text Classification (ATC), considering several classification solutions and many datasets. Our findings reveal a significant untapped potential for IS solutions. As a response to the limitations found in the SOTA IS methods when applied to ATC, the dissertation proposes two novel noise-oriented and redundancy-aware IS solutions specifically designed for large datasets and Transformer architectures. Our final solution achieved an average reduction of 41% in training set size while maintaining the same effectiveness levels in* ***all*** *experimented datasets. Our solutions demonstrated average speedup improvements of 1.67x (up to 2.46x), reducing carbon emissions (up to 65%), making them scalable for datasets with hundreds of thousands of documents. All code and datasets produced in the dissertation are available for replication on GitHub. Our results were published in some of the most important Information Retrieval and NLP conferences and journals, as it shall be detailed in this document.*

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

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

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`

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. 2023b] 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) [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 Information Retrieval (IR) and Natural Language Processing (NLP) journals in the last four years, including **two** published papers in the *Information Processing and Management*

[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`

(IP&M) – h5-index: 96; IF: 7.4 [Cunha et al. 2021, Cunha et al. 2020]; a survey paper published in the prestigious *ACM Computing Surveys* **(CSUR) – h5-index: 157; IF: 23.8** [Cunha et al. 2023a], and a paper on the *ACM Transactions on Information Systems* **(TOIS)– h5-index: 48; IF.: 5.4** [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) – h5-index: 103** [Cunha et al. 2023c] - the main worldwide IR conference - and the *ACM Int. Conf. on Theory of IR* **(ICTIR) – h5-index: 24** [Cunha et al. 2024b].

Our dissertation has also indirectly (by inspiration or use of results) contributed to other **7** journal papers during the doctorate period, including: **Neurocomputing** (h5-index: 136; IF: 5.5), **IP&M** (h5-index: 114; IF: 7.4), **Value in Health** (h5-index: 57; IF: 4.9), **JMIR Med Infor.** (h5-index: 52; IF: 3.1), **Computational Linguistics** (h5-index: 38; IF: 3.7), **OSNEM** (h5-index: 28; IF: 4.4), **Journal on Interactive Systems** (h5-index: 9). Similar indirect influence of our dissertation can be found in the following **14** conference papers: **ACL** (h5-index: 215), **CIKM** (h5-index: 91), **WSDM** (h5-index: 77), **CoNLL'24** (h5-index: 39), **WebSci** (h5-index: 34), five **WebMedia** (h5-index: 13), **Italian Information Retrieval Workshop** (h5-index: 7), **SBBD'23** (h5-index: 7), and the **IV Seminário de Grandes Desafios da Computação no Brasil 2025-2035**.

The combined h5-index of all above publications is **1501**. The respective papers have received so far more than **564** (according to Google Scholar[4]). For more details, we refer the reader to the companion "Sub-products" document. The h-index of the PhD is 11, which is high for someone who just obtained his PhD title.

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. 2023c]; (3) **Davide Bacciu** (UniPi Italy) studying Continuous Learning with IS support for temporal ATC, currently submitted as a full paper to SIGIR'25; (4) **Maurizio Ferrari Dacrema** and **Paolo Cremonesi** (PoliMi-Italy), helping with the organization the 2nd edition of the QuantumCLEF [Pasin et al. 2025] 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.

The PhD work also involved advising undergraduate students. We highlight the work presented in [Fonseca et al. 2024], accepted in the Scientific Initiation Paper Competition (CTIC-CSBC 2024) covering the proposal of IS-inspired **Undersampling** strategies for bias reduction in the context of transformer-based ATC – work co-advised by the PhD. This work also resulted in a paper under review for the ACL 2025 Conference.

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. 2023a])

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

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 these 13 most representative traditional IS methods 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. 2023a] (47 citations in Google Scholar in March, 2025).

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 se-

[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.

arch. 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 of the Dissertation), along with the experimental results and discussions, was published in a conference paper in the *ACM SIGIR Conference* (Qualis A1) [Cunha et al. 2023c], the most important worldwide Information Retrieval conference (acceptance rate: 20%).

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.*

Experimental 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 state-of-the-art 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], the most prestigious journal in Information Retrieval and Information Systems.

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. 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.

Referências

Andrade, C., Belém, F. M., Cunha, W., et al. (2023). On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". *IP&M*.

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 *CoNNL'24*.

Cunha, Washington Rosa, T., Rocha, L., Gonçalves, M. A., et al. (2023a). A comparative survey of instance selection methods applied to nonneural and transformer-based text classification. *ACM Comput. Surv.*

Cunha, W., Canuto, S., Viegas, F., Salles, T., et al. (2020). Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *IP&M*.

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

Cunha, W., Mangaravite, V., Gomes, C., et al. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *IP&M*.

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, W., Pasin, A., Goncalves, M., and Ferro, N. (2024b). A quantum annealing instance selection approach for efficient and effective transformer fine-tuning. In *ACM SIGIR ICTIR'24*.

Cunha, W., Rocha, L., Gonçalves, M. A., et al. (2023c). An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *ACM SIGIR'23*.

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

Fonseca, G., Cunha, W., and Rocha, L. (2024). Análise comparativa de métodos de undersampling em classificação automática de texto baseada em transformers. *CTIC*, 22(1).

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*.

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 TIST*, 13(2):1–41.

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*.

Pasin, A., Cunha, W., Dacrema, M. F., Cremonesi, P., Goncalves, M., and Ferro, N. (2025). Quantumclef - quantum computing at clef. In *Advances in Information Retrieval (ECIR)*.

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*.