A Comprehensive Exploitation of Instance Selection Methods for Automatic Text Classification

Washington Cunha
UFMG - Minas Gerais - Brazil
washingtoncunha@dcc.ufmg.br

Leonardo Rocha
UFSJ - Minas Gerais - Brazil
lcrocha@ufsj.edu.br

Marcos André Gonçalves
UFMG - Minas Gerais - Brazil
mgoncalv@dcc.ufmg.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 Large Language Models. However, training (or fine-tuning) large dense models for specific applications usually requires significant amounts of computing resources. This **Ph.D. dissertation** focuses on an under-investigated NLP data engineering (DE) technique, whose potential is enormous in the current scenario known as Instance Selection (IS). The IS goal is to reduce the training set size by removing noisy or redundant instances while maintaining the effectiveness of the trained models and reducing the training process cost. We provide a comprehensive and scientifically sound comparison of 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. We also propose two novel IS solutions that are noise-oriented and redundancy-aware, specifically designed for large datasets and transformer architectures. Our final solution achieved an average reduction of 41% in training sets, while maintaining the same levels of effectiveness in all datasets. Importantly, our solutions demonstrated speedup improvements of 1.67x (up to 2.46x), making them scalable for datasets with hundreds of thousands of documents. This thesis strongly aligns with WebMedia's objectives by addressing key challenges in processing vast web and social media data through innovative, scalable, and cost-effective strategies, falls under the (1) Document Engineering, Models and Languages; (2) AI, Machine Learning, and Deep Learning; and (3) NLP topics of the WebMedia call for papers.

KEYWORDS

Instance Selection, Automatic Text Classification, Deep Learning

1 INTRODUCTION

The rapid growth of 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 into predefined semantic categories. Accurate ATC models have become crucial for many emerging applications [6], such as spam, fake news and hate speech detection, relevance feedback, sentiment and product analysis, among others. As a supervised task, ATC benefits from applications generating large volumes of *labeled data*, such as social networks (e.g., X and Facebook). Thus, labeling has become less of an issue, while the abundance of labeled data is.

According to Andrew Ng, the success of Small and Large Language Models (SLMs and LLMs) such as RoBERTa and Llama 4 is due to extensive pre-training on massive datasets (e.g., 45PB for GPT-4) and the adaptability of pre-trained models via fine-tuning. Despite being faster than full training, fine-tuning still requires significant computational power. For instance, fine-tuning the SLM XLNet in the MEDLINE dataset took 80 GPU hours in our experiments. Resource limitations in companies and research groups also restrict experimentation with such models. In our PhD, we ran over 5,000 experiments that took approximately 5,600 hours. Reducing financial, computational, and environmental costs is crucial, given the significant energy consumption and carbon emissions associated with generating and using (large) language models.

Objective. Given increasing data volumes, re-training demands, and environmental concerns, proposing scalable and cost-effective NLP and ATC strategies has become essential. The recent success and real-world impact, including financial, of DeepSeek, which matched or surpassed the effectiveness of SOTA LLMs while reducing computational demands, highlights the importance of the trade-off between effectiveness and cost to the research and practitioner communities. This PhD dissertation focuses exactly on this trade-off, one of the SBC 2025-2035 Grand Challenges on Computer Science on AI Sustainability, from a *data engineering perspective*, by reducing training computational costs and carbon footprint without compromising performance. In particular, we focus on Instance Selection (IS), an understudied (in NLP and ATC at least), yet promising, set of techniques and growing research area [1, 3, 5], focused on selecting the most representative instances (documents) for a training set. IS aims to remove noisy or redundant instances from the training set to improve overall performance. IS methods have three main simultaneous 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, from preprocessing to model training to deployment.*

WebMedia Adherence. This PhD strongly aligns with WebMedia's objectives by addressing key challenges in processing vast web and social media data through innovative, scalable, and cost-effective NLP strategies. By proposing SOTA IS methods for ATC, it improves the efficiency and sustainability of SLMs and LLMs used in media analysis, reducing computational cost, energy use, and carbon footprint while maintaining or enhancing performance. These contributions benefit core WebMedia applications, including fake news detection, sentiment analysis, relevance feedback, and recommendation systems, influencing **five** WebMedia papers across Data Engineering, Machine Learning, Deep Learning, and NLP topics.

In: VII Concurso de Teses e Dissertações (CTD 2025). Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web (CTD'2025). Rio de Janeiro/RJ, Brasil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2596-1683

2 METHODOLOGY

The main contributions of our PhD dissertation are fourfold: (i) a comprehensive survey of the IS applied to ATC; (ii) an extended IS taxonomy; and (iii-iv) two novel SOTA IS approaches applied to NLP/ATC. Due to space limitations, we focus on the latter two, noticing that our article on the ACM Comp. Surveys [6], derived from the dissertation, has been highly cited (60 citations as of July/2025). First, we proposed **E2SC** [2], a two-step IS framework aimed at large datasets. E2SC's first step assigns a probability to each instance being removed from the training set. We exploit cheap and calibrated methods for that (KNN). Our first hypothesis (H1) was that *high confidence (if the model is calibrated) positively correlates with redundancy for building a stronger model.* Next, we estimate a near-optimal reduction rate that does not degrade the SLM's effectiveness. Our second hypothesis (H2) was that *we can estimate the effectiveness of a robust model through the variation of selection rates in a weaker model.* Again, we explored KNN to gather evidence for this hypothesis by introducing an iterative method that statistically compared, using a validation set, the effectiveness of the weak model without data reduction against the model with iterative data reduction rates.

Despite excellent results regarding the trade-off effectiveness-efficiency-reduction, other aspects such as noise – defined as instances incorrectly labeled by humans as well as outliers that do not contribute to model learning – were not explored in our first solution. To fill this gap, we proposed **biO-IS** [4], built on top of E2SC, aimed at simultaneously removing redundant and noisy instances. biO-IS has three main parts: (i) a weak classifier; (ii) a redundancy-based step; and (iii) an entropy-based step. We departed from E2SC, considering the Logistic Regression (LR) as the weak classifier instead of KNN, which proved a better effectiveness-calibration trade-off. To address the noise removal objective, we proposed a new step to be combined with E2SC based on entropy, as well as a novel iterative process to estimate near-optimum reduction rates. Considering wrongly predicted instances by the weak classifier, the main objective is to assign a probability to each of them being removed from the training set based on the probability of the instance being noisy. For this, we proposed using entropy as a proxy to determine the reduction behavior for the sake of training a stronger model.

3 EVALUATION

We compared our proposals with 13 IS baselines in the ATC domain found on our systematic literature review, considering 22 datasets and 7 SOTA classifiers (BERT, RoBERTa, Llama, among others). Our experiments showed that **E2SC** was able to reduce training sets by 29% while maintaining the same levels of effectiveness in almost all datasets, with speedups of 1.37x on average. It scaled to large datasets, reducing them by up to 40% while maintaining the same effectiveness with speedups of 1.70x. E2SC focused only on redundancy, however. **biO-IS**, in turn, extended E2SC, being capable of removing, besides redundancy, also noisy instances in up to 66.6%. biO-IS managed to reduce up to 60% of the training sets while maintaining the same quality levels in **all** of the considered datasets. biO-IS was also capable of consistently producing speedups up to 2.46x. No baseline, not even E2SC, was able to achieve results with this level of quality, considering all tripod criteria. biO-IS improved over E2SC by 41% regarding reduction rate and from 1.42x to 1.67x (on average) regarding speedup, being the current SOTA IS applied to NLP.

4 STATE-OF-THE-ART ADVANCEMENT

In the Ph.D. dissertation, we show SLMs and LLMs often require representative - not large - data to perform well in ATC. Overall, IS techniques effectively reduced training set sizes without compromising effectiveness. The previous SOTA in IS fell short of meeting all tripod criteria simultaneously, underscoring the need for more efficient, scalable IS solutions, especially in big data scenarios. To address these challenges and fill the gaps found in the literature, we proposed two novel IS methods focused on redundancy (only) and noise (in conjunction with redundancy). Extensive experiments confirmed our hypotheses: *SLMs and LLMs can be trained with less data without sacrificing effectiveness.* This not only enables cost savings but also contributes to reducing carbon emissions. Such experimental evaluation established our solutions as the current SOTA IS applied to NLP. Such promising results instill hope for a more sustainable (green) and efficient NLP future, where advancements in IS techniques can produce environmental and economic benefits.

Scientific Production. This PhD directly resulted in four journal papers (2 IP&M, ACM TOIS, and ACM Computing Surveys), two conference papers (SIGIR and ICTIR), and open-source software releases, as well as contributions as co-author to other 8 journals, including ACM CSUR, Neurocomputing, IP&M, Value in Health, JMIR Med, Comp. Linguistics, OSNEM, and JIS. Our ideas, insights, and methods also contributed to other 22 conference papers, including: GDCOMP, ACL'25 and ACL'20, SIGIR'25, CIKM, WSDM, ECIR'25, CoNLL, WebSci'25, ENIAC'24, **five WebMedia**, IIR, SBRC'25, and SBBD. The combined h5-index of all mentioned publications[1] is **2109**. The respective papers have received so far more than **689** (according to Google Scholar[2]). The h-index of the PhD is 13, which is high for someone who just obtained his PhD title.

Awards and Achievements. During the Ph.D., the student has received important awards, including: (1) 2nd place on **CTD-CSBC'25**; (2) Best reviewer on **SIGIR**'24 and **ACL**'24; (3) **CTIC'24** and **CTIC'19** as co-advisor; (4) **SIGIR Student Travel Awards**; (5) **Honorable Mention** in WFA - **WebMedia**'23; (6) CAPES-PRINT grant to spend a semester abroad (ISTI-CNR Italy) (Sandwich PhD).

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

REFERENCES

[1] Washington Cunha, Sérgio Canuto, Felipe Viegas, Thiago Salles, C. Gomes, V. Mangaravite, E. Resende, Thierson Rosa, Marcos Gonçalves, and Leonardo Rocha. 2020. Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *IP&M* (2020).

[2] Washington Cunha, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2023. An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *SIGIR'23.*

[3] Washington Cunha, V. Mangaravite, C. Gomes, S. Canuto, E. Resende, Cecilia Nascimento, F. Viegas, C. França, Jussara M Almeida, et al. 2021. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *IP&M* 58, 3 (2021), 102481.

[4] Washington Cunha, Alejandro Moreo Fernández, Andrea Esuli, Fabrizio Sebastiani, Leonardo Rocha, and Marcos André Gonçalves. 2025. A Noise-Oriented and Redundancy-Aware Instance Selection Framework. *ACM TOIS* 43, 2 (2025).

[5] Washington Cunha, Andrea Pasin, Marcos Goncalves, and Nicola Ferro. 2024. A Quantum Annealing Instance Selection Approach for Efficient and Effective Transformer Fine-Tuning. In *Proceedings of the 2024 ACM SIGIR ICTIR.*

[6] Washington Cunha, Felipe Viegas, Celso França, Thierson Rosa, Leonardo Rocha, and Marcos Gonçalves. 2023. A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM CSUR* (2023).

[1] For more details and publications, we refer the reader to http://bit.ly/3WvX1T5

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