United Experts: Efficient Multi-Task Sentiment Analysis in Portuguese via Model Merging

João Assaoka[1], Evangelos Milios[2], Vlado Keselj[2], José Vitor Sousa Cardoso Requena[1], Diego de Almeida Miranda[1], Lilian Berton[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brazil

[2]Faculty of Computer Science – Dalhousie University
Halifax – NS – Canada

{joao.assaoka, lberton}@unifesp.br

Abstract. *Building robust multi-task models for sentiment analysis traditionally relies on joint training, which can be computationally prohibitive and susceptible to dataset imbalance. As an efficient alternative, we evaluate the application of model merging to combine independently trained Portuguese Small Language Models (SLMs). By fusing five distinct domain experts, we aggregate capabilities at zero additional optimization cost. Empirical results indicate that while joint training serves as a competitive empirical baseline, model merging successfully preserves individual domain expertise and offers a significantly superior compute-performance tradeoff. Additionally, we highlight the utility of merged weights as powerful prior initializations, significantly improving cross-domain transfer performance and reducing the cost of adapting the model to new, unseen domains.*

1. Introduction

Sentiment and emotion analysis remain foundational tasks in Natural Language Processing (NLP), enabling the extraction of subjective information across diverse domains such as financial news, book reviews, and social media [Zhang et al. 2018, Liu 2020]. Recently, Small Language Models (SLMs) have emerged as highly efficient and deployable alternatives to massive Large Language Models (LLMs), offering competitive capabilities on specialized tasks with a fraction of the computational and memory footprint [Maddigan et al. 2024]. However, in the context of the Portuguese language, developing robust SLMs capable of generalizing across highly varied textual landscapes remains particularly challenging, often restricted by the availability of high-quality, diverse corpora compared to English [Souza et al. 2020]. Traditionally, practitioners have relied on either fine-tuning isolated SLM domain experts or training joint multi-task learning (MTL) models [Crawshaw 2020]. While isolated experts perform exceptionally well within their native distributions, they suffer from catastrophic forgetting and fail to generalize across external domains [Kirkpatrick et al. 2017, French 1999]. Conversely, joint multi-task training often serves as a strong empirical multi-task baseline, but it introduces prohibitive computational costs and lengthy processing times [Houlsby et al. 2019], particularly when the model needs to be updated with new data.

To bypass the heavy computational burden of joint training, recent advancements in parameter-efficient techniques have introduced model merging—a process that com-

bines the weights of multiple fine-tuned expert models directly in weight-space without the need for further gradient optimization [Goddard et al. 2025]. Techniques such as TIES-Merging [Yadav et al. 2023] and DARE [Yu et al. 2024] allow for the aggregation of domain-specific knowledge into a single unified backbone. This fusion approach mitigates the interference often seen in multi-task training while retaining the specialized capabilities of the individual experts.

However, the most significant advantage of this fusion approach lies in its operational efficiency when encountering new data. Training a joint model from scratch every time a new dataset or domain is introduced is highly inefficient. In this paper, we demonstrate that a fused model acts as a highly regularized, multi-domain prior that drastically accelerates the processing and adaptation time for new, unseen tasks. By applying a leave-one-dataset-out methodology across five distinct Portuguese sentiment and emotion datasets, we show that initializing a training run with a fused backbone rather than a raw base model leads to much faster convergence. This approach significantly cuts down the time required to achieve high classification performance when transferring to unseen target domains. Ultimately, this work provides a scalable, time-efficient framework for expanding Portuguese sentiment analysis models, proving that model merging is not just a tool for knowledge aggregation, but a critical mechanism for accelerating future learning.

The main contributions of this work are summarized as follows:

1. Comprehensive Multi-Domain Evaluation in Portuguese: We provide a systematic comparison of isolated fine-tuning versus joint multi-task learning (MTL) across five diverse Portuguese sentiment and emotion datasets. We empirically demonstrate the effects of catastrophic forgetting in isolated experts and highlight how data volume asymmetry and semantic alignment allow joint training to act as a strong empirical baseline for synergistic tasks.
2. Parameter-Efficient Multi-Tasking via Model Merging: We introduce weight-space model merging (utilizing TIES and DARE algorithms) to the Portuguese sentiment analysis domain. We show that aggregating domain-specific experts into a single unified backbone is a viable, parameter-efficient alternative to computationally expensive joint training.
3. Accelerated Cross-Domain Transfer Learning: Through a leave-one-dataset-out methodology, we demonstrate that a merged multi-expert model serves as a superior initialization prior for transferring to entirely new target domains. This merged backbone significantly accelerates processing and adaptation times while yielding higher final F1-scores compared to starting from the raw base model.
4. Open-Source Artifacts for Portuguese NLP: To foster future research and reproducibility, we make our complete framework publicly available, including the source code and prompts on GitHub[1], and the fine-tuned model weights and metrics on Hugging Face Hub[2].

2. Related Work

Sentiment and emotion analysis inherently span diverse domains, text styles, and annotation granularities. To build models capable of handling this diversity, Multi-Task Learning

[1]https://github.com/Assaoka/ENIAC_26--United_Experts_Multi_Task_Sentiment_Analysis_in_Portuguese_via_Model_Merging

[2]https://huggingface.co/Assaoka

(MTL) has been extensively explored as a method to mitigate data scarcity and improve generalization capabilities [Crawshaw 2020]. By jointly optimizing a shared representation across multiple related tasks, models can leverage auxiliary data to improve performance on low-resource domains. For instance, massive multi-task frameworks have demonstrated that sharing latent representations across sentiment, natural language inference, and text classification significantly boosts out-of-domain robustness in English benchmarks [Liu et al. 2019]. However, these joint training paradigms are notoriously resource-intensive and prone to negative transfer when task synergies are asymmetric, requiring careful data balancing and gradient routing.

To circumvent the computational overhead of joint MTL, the field has recently pivoted towards weight-space model merging. Rather than training a unified model from scratch, model merging aggregates the learned weights of isolated, domain-specific experts into a single cohesive backbone without requiring additional gradient updates [Goddard et al. 2025]. Foundational strategies such as Model Soups [Wortsman et al. 2022] demonstrated that averaging the weights of multiple fine-tuned configurations could improve out-of-distribution performance. This was further advanced by Task Arithmetic [Ilharco et al. 2023], which showed that task-specific capabilities could be added or subtracted in weight space. Advanced fusion algorithms like TIES [Yadav et al. 2023] and DARE [Yu et al. 2024] now dominate the landscape by resolving parameter interference and sign conflicts, enabling the seamless integration of highly specialized language models. While these techniques are well-documented for English tasks, their application to cross-domain sentiment analysis remains underexplored.

While English dominates NLP research, significant efforts have been made to develop robust sentiment analysis frameworks in other languages, highlighting the universal need for domain-adaptable models. In Spanish, the introduction of language-specific models like BETO established strong baselines for sentiment classification across diverse linguistic dialects and text sources [Cañete et al. 2020]. Similarly, AraBERT significantly advanced Arabic sentiment analysis by providing a contextualized understanding of complex morphological structures and dialectal variations across social media and news domains [Antoun et al. 2020]. For Chinese, models like MacBERT have been extensively benchmarked on massive, multi-domain sentiment and emotion datasets, proving the efficacy of native-language pre-training over multilingual distillation [Cui et al. 2020].

Despite these advancements in other languages, the Portuguese NLP ecosystem still lacks an efficient, scalable framework for rapidly adapting models to new sentiment domains. Existing Portuguese models, such as BERTimbau [Souza et al. 2020], rely heavily on standard single-task fine-tuning. Our work bridges this gap by introducing weight-space model merging to the Portuguese language, offering a computationally efficient alternative to traditional MTL that accelerates cross-domain transfer learning for varied sentiment and emotion tasks.

3. Methodology

In this section, we present the methodology employed to build and evaluate multi-domain sentiment and emotion analysis models in Portuguese.

3.1. Datasets

To validate our approach across diverse domains and tasks, we selected five distinct sentiment analysis datasets for Portuguese. These datasets span varying text domains (e.g., book reviews, financial news, and social media comments) and different target formats (polarity and multi-label emotion classification). The specific characteristics of each dataset are detailed below:

- **ReLi-SA** (hereafter **ReLi**) [Freitas et al. 2014]: A native Portuguese corpus consisting of 12,511 annotated book reviews. It requires models to predict a categorical sentiment label (4 classes: positive, negative, neutral, or mixed). We preserve its original partition of 7,875 training, 1,348 validation, and 3,288 test examples.
- **BRIGHTER** (hereafter **BRGT**) [Muhammad et al. 2025]: A native Portuguese dataset containing 7,078 instances annotated for six fundamental emotion categories (anger, disgust, fear, joy, sadness, and surprise) in a multi-label classification setup. We preserve its original partition consisting of 2,226 training, 400 validation, and 4,452 test examples.
- **Sentiments-FinBERT-PT-BR** (hereafter **FinPT**) [Santos et al. 2023]: A native Portuguese sentiment analysis benchmark specifically curated for financial news written in Brazilian Portuguese, containing 503 annotated instances, with out-of-scope examples filtered out, structured as a polarity classification task. It lacks a train/test split, so we apply an 80/20 stratified split (402 training and 101 test examples) to benchmark its performance.
- **Financial Phrasebank (ptbr)** (hereafter **FinPB**) [Malo et al. 2013]: A Portuguese translation[3] (manually validated) of the classic Financial Phrase Bank sentiment dataset, consisting of 4,845 phrases annotated for polarity classification. It is very similar to FinPT, differing mainly in the news source. Lacking a predefined split, we divided it into 3,876 training and 969 test examples using an 80/20 division.
- **Go Emotions PT-BR** (hereafter **GoEmo**) [Demszky et al. 2020]: A Portuguese translation[4] of the granular multi-label emotion dataset Go Emotions, consisting of 211,225 social media comments mapped to 28 emotion categories alongside a neutral class. We split it into 168,980 training and 42,245 test examples using an 80/20 division. Only 50% of the training data was used for fine-tuning due to computational constraints.

3.2. Model Fine-Tuning and Training Protocol

To adapt the models for Portuguese sentiment and emotion analysis tasks, we employ the *Polygl0t/Tucano2-qwen-0.5B-Instruct* model [Corrêa et al. 2026], a 0.5-billion parameter causal language model tailored for the Portuguese language. A primary advantage of Tucano2 is its fully open-source nature, extending beyond open weights to provide full transparency regarding its pre-training corpora. This ensures that the datasets used in this study were not seen during pre-training, mitigating the risk of invisible data leakage and bolstering the reliability of our evaluations. Furthermore, Tucano2 demonstrates superior intrinsic proficiency in Portuguese compared to broadly multilingual models.

[3] https://www.kaggle.com/datasets/mateuspicanco/financial-phrase-bank-portuguese-translation/

[4] https://www.kaggle.com/datasets/antoniomenezes/go-emotions-ptbr/

A system prompt provides the model with the candidate classes and instructs it to generate predictions strictly within a valid JSON structure, matching the specific attributes required by each task. ReLi, FinPT, and FinPB return categorical class labels, whereas BRGT and GoEmo return lists of emotions present in the text.

We trained one model for each dataset and one joint model with all datasets. Both configurations were evaluated on all datasets' test splits. These models were optimized through full parameter fine-tuning for three epochs with the AdamW optimizer. The learning rate is maintained at 2×10^{-5} with a weight decay of 0.01 and an effective batch size of 64. These hyperparameters follow standard empirical defaults for causal transformer fine-tuning, maintaining a uniform setup across all tasks without per-task tuning sweeps.

3.3. Model Merging

Beyond fine-tuning each expert model, we investigate model merging as a parameter-efficient strategy for multi-task learning (MTL) on sentiment analysis for portuguese, aggregating domain-specific knowledge into a single unified model.

Model merging aggregates the weights of multiple fine-tuned experts into a single model without additional training. This approach bypasses the immense computational cost of re-training a joint model and allows for seamless, modular integration of new tasks.

All merges are performed using the `mergekit` library [Goddard et al. 2025], which provides a unified interface for a variety of model merging methods. We evaluate several merging strategies: Model Soup [Wortsman et al. 2022], Task Arithmetic [Ilharco et al. 2023], TIES-Merging (Trim, Elect Sign, and Merge) [Yadav et al. 2023], DARE (Drop and Rescale) [Yu et al. 2024].

For the merging configurations, we evaluated two variants of DARE (Linear and TIES) to consolidate the remaining weights, testing parameter retention densities of 0.3, 0.5, and 0.7 across both DARE and TIES methods.

3.4. Cross-Domain Transfer Learning via Merged Priors

A primary objective of multi-task learning in sentiment analysis is providing a superior initialization backbone for adapting to new, unseen sentiment tasks in Portuguese. Specifically, while contactless zero-shot evaluation tests unadapted out-of-domain performance, practical workflows adapt models to new domains via fine-tuning. We evaluate this cross-domain transfer learning capacity using a leave-one-dataset-out methodology.

In each iteration of this leave-one-out protocol, we select four out of the five expert models (e.g., ReLi, BRGT, FinPT, and GoEmo) and merge their parameters using the TIES-Merging algorithm to produce a consolidated backbone model. This merged checkpoint serves as a multi-domain initialization prior. We then fine-tune and evaluate this merged prior on the fifth, completely unseen target dataset (e.g., FinPB). This transfer learning protocol simulates a real-world scenario where a practitioner inherits a pre-merged sentiment model and adapts it to a new target domain, allowing us to verify if merged weights offer a superior initialization prior compared to starting from the raw base model.

3.5. Iterative Task Addition

In practical scenarios, new datasets and tasks often arrive sequentially rather than all at once. To evaluate our approach under these conditions, we implement an iterative task addition and merging protocol. Instead of training independent experts from a clean base model and merging them in a single step, the models are trained and merged sequentially. We start with the expert model trained on GoEmotions (Step 1). For each subsequent task, the process consists of: (1) **Pre-train Merge**, where models from previous steps are merged using TIES-Merging to form a consolidated prior; and (2) **Post-train Fine-tuning**, where this merged prior is fine-tuned on the new target dataset.

The sequential pipeline consists of fine-tuning BRIGHTER (Step 2) on top of Step 1, and then sequentially merging and training on ReLi-SA (Step 3), Financial Phrasebank (Step 4), and FinBERT (Step 5), culminating in a final merge of all 5 models (Step 6). To balance old and new knowledge, we evaluate two weight distribution strategies at each merge step N: (i) **Equal Weights** (v1), where each participating model receives an equal weight of $\frac{1}{N}$; and (ii) **Balanced Weights** (v2), which allocates 50% weight to the newest model to protect its domain expertise and splits the rest equally among older models, yielding $w_{\text{new}} = \frac{N+1}{2N}$ and $w_{\text{old}} = \frac{1}{2N}$ each.

3.6. Evaluation Protocol

All the comparatives in this paper are evaluating Macro F1-score, but we also report Micro F1-score, Accuracy, and Hamming Loss for all tasks. As aggregate metrics for the MTL framework, we employ the simple average $\mu = \frac{1}{N}\sum_{i=1}^{N} M_i$ and the min-max normalized average $\mu_n = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{M_i - M_i^{\min}}{M_i^{\max} - M_i^{\min}}\right)$ across all tasks, where N is the total number of tasks ($N = 5$), M_i represents the raw metric score achieved by a model on task i, while $M_i^{\min}$ and $M_i^{\max}$ correspond to the minimum and maximum performance bounds observed across all evaluated models for that specific task.

The key difference lies in how each metric handles task scale imbalances. The simple average (μ) treats raw scores directly, meaning tasks with naturally higher absolute values or higher variance disproportionately dominate the aggregate score. Conversely, the normalized average (μ_n) applies a column-wise Min-Max scaling, bound between $[0, 1]$. This standardizes the contribution of each task, giving more weight to relative improvements in difficult tasks where the performance gap between the baseline and the best expert model is compressed.

The source code, scripts, and task-specific prompts are publicly available on GitHub, and the fine-tuned models, merged checkpoints, and evaluation metrics are hosted on Hugging Face Hub. Model availability is guaranteed for at least one year, after which easily reproducible checkpoints may be archived as they can be deterministically reconstructed using the provided codebase.

To ensure clarity and consistency across all empirical evaluations, tables follow this formatting convention. In-domain tasks are explicitly highlighted with an underline. Within each competitive grouping the highest performing score for each task is typeset in bold, while the second-highest score is denoted in italics. For the comparative analysis, superscripts $^{+}$ and $^{-}$ are employed to indicate an absolute performance gain or loss ≥ 0.005, respectively, in comparison to the corresponding to Table 1.

4. Results and Discussion

In this section, we present and analyze the empirical performance of the base model, isolated expert models, the joint multi-task baseline, and the model merging strategies.

4.1. Cross-Task and Multi-Domain Evaluation

Table 1. Cross-task evaluation results (Macro F1) of base, expert, and fusion models.

Model / Task	ReLi	BRGT	FinPT	FinPB	GoEmo	μ	$\mu_{\mathbf{n}}$
Base	0.3066	0.3347	0.5477	0.3813	0.0759	0.3292	0.3129
ReLiSA	**0.6420**	0.3526	0.4132	0.4255	0.0739	0.3814	0.4423
Brighter	0.3288	0.3512	0.7295	0.6164	0.0601	0.4172	0.4502
Finbert	0.4172	0.3357	*0.8081*	0.6711	0.0643	0.4593	0.5169
Phrasebank	0.4481	0.3251	0.7919	**0.8584**	0.0702	*0.4988*	*0.5825*
GoEmotions	0.3994	*0.3683*	0.1447	0.4203	**0.3000**	0.3265	0.4678
Joint	*0.5951*	**0.4083**	**0.8591**	*0.8218*	*0.2999*	**0.5968**	**0.9114**
Model-Soup	0.4636	0.3890	**0.7505**	*0.7352*	0.1454	0.4967	0.6536
Task-Arithmetic	**0.4644**	0.3888	**0.7505**	**0.7359**	0.1459	*0.4971*	*0.6544*
TIES	*0.4638*	*0.3900*	0.7182	0.7017	*0.2160*	**0.4979**	**0.6935**
DARE-Linear	0.4637	**0.3902**	**0.7505**	0.7335	0.1450	0.4966	0.6538
DARE-TIES	0.4361	0.3729	0.2786	0.5186	**0.2249**	0.3662	0.4934

The experimental results reveal a stark contrast between local domain specialization and multi-domain generalist capabilities. As shown in Table 1, isolated fine-tuning yields strong domain-specific experts, usually picking first place in their respective datasets. Exceptions are made on BRGT and FinPT, where the Joint model picks first place.

We hypothesize that this behavior is driven by two main factors: (i) **Financial Domain Synergy**, where the small *FinPT* training split (402 samples) co-learns from *FinPB* ($3,876$ samples) during joint training as high-affinity data augmentation; and (ii) **Emotion Class Overlap**, where *BRGT* ($2,226$ samples, 6 emotions) benefits from *GoEmo* ($84,490$ samples, 28 emotions) because GoEmo's 28 emotions natively encompass the 6 categories of *BRGT*, enabling the Joint model to outperform isolated experts.

However, the performance of all experts drops when evaluated on out-of-domain datasets due to catastrophic forgetting. We can observe this pattern on the ReLi-SA and GoEmotions experts, which perform worse than the Tucano model with base weights (hereafter Base).

While the Joint model achieves strong raw metrics on specific datasets, it acts primarily as an empirical baseline. In real-world applications, re-training a joint model from scratch for every new dataset incurs prohibitive computational and financial costs. Model merging disrupts this paradigm by democratizing multi-task learning. By mathematically fusing expert weights post-training, merging achieves highly competitive performance at virtually zero optimization cost. This makes it a vastly superior and sustainable (Green AI) approach for scaling sentiment analysis.

Comparing the merging strategies in Table 1, flat merging methods yield closely matched raw averages ($\mu \approx 0.4966$–0.4979), indicating that algorithms such as TIES,

Table 2. Density sensitivity analysis (Macro F1).

Model / Task	ReLi	BRGT	FinPT	FinPB	GoEmo	μ	$\mu_{\mathbf{n}}$
DARE-Linear							
Density 0.3	0.4631	0.3876	**0.7505**	*0.7346*	0.1445	0.4960	0.6512
Density 0.5	0.4637	0.3902	**0.7505**	0.7335	0.1450	0.4966	0.6538
Density 0.7	**0.4642**	0.3899	**0.7505**	**0.7356**	0.1462	0.4973	0.6554
DARE-TIES							
Density 0.3	0.2437	0.3714	0.1447	0.2807	0.2092	0.2500	0.2846
Density 0.5	0.4361	0.3729	0.2786	0.5186	*0.2249*	0.3662	0.4934
Density 0.7	0.4627	0.3800	0.5696	0.6256	**0.2275**	0.4531	0.6281
TIES							
Density 0.3	0.4636	**0.3908**	*0.7182*	0.6986	0.2159	0.4974	0.6929
Density 0.5	*0.4638*	0.3900	*0.7182*	0.7017	0.2160	**0.4979**	*0.6935*
Density 0.7	0.4636	*0.3903*	*0.7182*	0.7017	0.2159	*0.4979*	**0.6935**

Task Arithmetic, Model Soup, and DARE-Linear perform virtually indistinguishably within empirical variance on unweighted raw scores. Nevertheless, TIES maintains the highest overall consistency across both raw ($\mu = 0.4979$) and normalized metrics ($\mu_n = 0.6935$), demonstrating robust retention across heterogeneous tasks. TIES and DARE share a hyperparameter, the density (k), which governs the proportion of parameters retained prior to fusion. In Table 2, we evaluate sensitivity to this parameter. While TIES and DARE-Linear remain stable across density thresholds (0.3, 0.5, and 0.7), DARE-TIES exhibits high sensitivity, requiring higher density values to avoid performance degradation. We hypothesize this sensitivity stems from the compact 0.5B parameter scale of the backbone model, where aggressive parameter dropping can disrupt crucial representation paths.

Table 3. Comparison between training execution times and model merging times.

Dataset	Train Size	Training Time
ReLi	7,875	5 min 17 s
BRGT	2,226	1 min 42 s
FinPT	402	28.13 s
FinPB	3,876	2 min 32 s
GoEmo	84,490	1 h 11 min 40 s
Joint	98,869	1 h 17 min 45 s

Method	Merge Time
Model-Soup	34.81 s
Task-Arithmetic	20.92 s
TIES	25.30 s
DARE-Linear	22.60 s
DARE-TIES	24.67 s
Avg.	25.66 s

Table 3 juxtaposes the training execution times of the expert models against the model merging times. While fine-tuning isolated experts or a centralized joint model requires hours of GPU compute on high-end accelerators, the merging process executed via `mergekit` completes in seconds (≈ 25s). This near-instantaneous weight fusion enables modular knowledge integration without costly joint retraining runs. Furthermore, when adapting to new target domains, initializing from a merged prior yields higher initial task performance and rapid adaptation within standard 3-epoch runs compared to starting from the raw base model. For merging methods that utilize weight pruning (TIES and DARE), we report the execution times using the standard 0.5 density configuration, as other density levels (0.3 and 0.7) yielded virtually identical execution durations.

4.2. Cross-Domain Transfer Learning via Merged Priors

Table 4. Cross-task merging initialization evaluation results (Macro F1).

Model / Task	ReLi	BRGT	FinPT	FinPB	GoEmo	μ	μ_{n}
ReLiSA	**0.6975**$^{+}$	0.3772^{+}	0.2245^{-}	0.3073^{-}	0.1973^{+}	0.3608^{-}	0.4980^{+}
Brighter	0.4346^{+}	**0.4335**$^{+}$	0.7104^{-}	*0.6941*$^{+}$	0.1608^{+}	0.4867^{+}	0.6674^{+}
Finbert	0.4023^{-}	*0.3947*$^{+}$	*0.7972*$^{-}$	0.6610^{-}	0.2148^{+}	*0.4940*$^{+}$	*0.6795*$^{+}$
Phrasebank	*0.4544*$^{+}$	0.3653^{+}	**0.8117**$^{+}$	**0.8532**$^{-}$	*0.2193*$^{+}$	**0.5408**$^{+}$	**0.7470**$^{+}$
GoEmotions	0.2192^{-}	0.3902^{+}	0.1447	0.2959^{-}	**0.3000**	0.2700^{-}	0.3709^{-}

A primary advantage of model merging is the creation of a robust prior for subsequent fine-tuning—a process we refer to as Knowledge Accumulation. Table 4 shows the impact of using a merged multi-task backbone as the initialization point for training new experts.

Table 5. Comparison of Prior-Merge models against the TIES baseline (Macro F1).

Model / Task	ReLi	BRGT	FinPT	FinPB	GoEmo	μ	μ_{n}
ReLiSA	**0.4881**$^{+}$	0.3904	0.6610^{-}	0.6604^{-}	**0.2337**$^{+}$	0.4867^{-}	0.6893
Brighter	0.4539^{-}	**0.4184**$^{+}$	0.7086^{-}	0.6868^{-}	*0.2302*$^{+}$	0.4996	*0.7176*$^{+}$
Finbert	0.4366^{-}	0.3969^{+}	**0.8128**$^{+}$	*0.7311*$^{+}$	0.2245^{+}	**0.5204**$^{+}$	**0.7309**$^{+}$
Phrasebank	0.4571^{-}	0.3816^{-}	*0.7359*$^{+}$	**0.7491**$^{+}$	0.2282^{+}	*0.5104*$^{+}$	0.7142^{+}
GoEmotions	*0.4780*$^{+}$	*0.3996*$^{+}$	0.6404^{-}	0.6817^{-}	0.2091^{-}	0.4817^{-}	0.6735^{-}

With the exception of the GoEmotions expert, all models exhibited significant improvements in cross-domain transfer evaluation when initialized from the merged prior, with notable boosts on BRGT and GoEmo datasets leading to a higher μ_n. The ReLi and BRGT experts also benefited from this prior on In-Domain tasks.

Most importantly, the Phrasebank expert fine-tuned from the merged prior outperforms our best flat merging strategy on both μ and μ_n. This demonstrates that while flat merging might lag behind joint training, using the merged model as a strong initialization for new target data yields superior generalization. It effectively transfers accumulated knowledge into the new domain.

4.3. Iterative Task-Addition

The results of the sequential task addition protocol, detailed in Table 6, highlight the dynamics of incremental learning and weight fusion.

First, sequentially fine-tuning on top of a merged checkpoint consistently yields high local performance on the newly introduced task, matching or even exceeding the performance of isolated experts trained directly from the clean base model. For instance, in Step 3 Post-train, the model achieves a Macro F1 score of 0.6681 on ReLi, outperforming the isolated ReLiSA expert (0.6420). Similarly, in Step 5 Post-train, the model reaches 0.8338 on FinPT, compared to 0.8081 for the isolated Finbert expert. This indicates that sequentially accumulating sentiment and emotion knowledge provides a beneficial initialization that aids in learning new aligned domains.

Second, the intermediate pre-train merges (Steps 3, 4, and 5 Pre-train) generally suffer from performance degradation on the newer tasks compared to their post-train

Table 6. Sequential task adding and merging evaluation results (Macro F1).

Model / Task	ReLi	BRGT	FinPT	FinPB	GoEmo	μ	$\mu_{\mathbf{n}}$
Brighter	0.4297	**0.4413**	0.3242	0.5381	0.2094	0.3885	0.5545
ReLiSA$_e$	*0.6681*	0.4094	0.3137	0.3117	0.2170	0.3840	0.5554
Phrasebank$_e$	0.4687	0.4120	0.7836	*0.8579*	0.2328	0.5510	0.7973
Finbert$_e$	0.4352	0.4247	0.8338	0.7373	0.2566	0.5375	0.7883
Final Merge$_e$	0.4641	0.4175	0.6750	0.7486	**0.2638**	0.5138	0.7600
ReLiSA$_w$	**0.7029**	0.4088	0.1786	0.2533	0.2193	0.3526	0.5147
Phrasebank$_w$	0.4890	0.4205	0.7835	**0.8636**	0.2327	0.5579	0.8147
Finbert$_w$	0.4784	0.4234	**0.8678**	0.7941	0.2401	*0.5608*	*0.8193*
Final Merge$_w$	0.4826	*0.4255*	*0.8607*	0.8087	0.2494	**0.5654**	**0.8334**
Final Merge$_{w,\text{eq}}$	0.4816	0.4160	0.6730	0.7636	*0.2579*	0.5184	0.7654

peaks. This behavior is driven by a task-vector dilution effect. Because earlier tasks (such as GoEmotions and BRIGHTER) are present in all models participating in the merge pool, their parameter updates are reinforced. Conversely, the newer tasks (e.g., ReLi-SA in Step 3, or Phrasebank in Step 4) are only present in the latest fine-tuned model. Consequently, their task vectors are scaled down by the TIES-Merging process (e.g., by a factor of $1/K$ where K is the number of merged models).

This dilution is particularly visible in the final Step 6 merge, where the FinBERT task vector only has a 0.2 weight representation, causing FinPT to drop from 0.8338 to 0.6750. However, despite this unequal task-vector scaling, the final Step 6 model achieves an overall average Macro F1 of 0.5138, which still outperforms the flat TIES-Merge of the five independent isolated experts (0.4979). This demonstrates that sequential training on top of merged checkpoints shapes parameter paths that are more structurally compatible and less prone to destructive interference than independent training runs.

5. Conclusion and Future Work

In this work, we investigated model merging as a highly efficient alternative to centralized multi-task learning for sentiment analysis in Portuguese. Using the Tucano2-qwen-0.5B-Instruct backbone, we fine-tuned domain-specific experts on five distinct corpora and combined them using several model merging techniques.

Our empirical results demonstrated a clear compute-performance trade-off. While joint training acts as a strong empirical baseline ($\mu = 0.5968$) for datasets with high domain synergy, its massive computational overhead and vulnerability to data imbalance make it infeasible for continuous scaling. Model merging overcomes this barrier by aggregating expertise at zero additional training cost (≈ 25 s merging duration), with TIES-Merging achieving a strong aggregate Macro F1 of $\mu = 0.4979$ ($\mu_n = 0.6935$).

Crucially, we demonstrated that the benefits of merging extend beyond flat evaluation. By utilizing fused weights as an initialization prior, we proved that merging acts as a powerful mechanism for Knowledge Accumulation. Models initialized from the merged backbone achieved up to $\mu = 0.5408$ on transfer evaluation and demonstrated rapid adaptation during iterative task-addition (reaching $\mu = 0.5654$), effectively eliminating the need to retrain from a clean base model.

For future work, we plan to: (i) Evaluate the merging and knowledge accumulation protocols on wider Portuguese language models (e.g., 1.5B and 3.7B Tucano2 variants). (ii) Explore parameter-efficient merging methods (such as LoRA-Merging) to further optimize cross-domain sentiment synthesis. (iii) Conduct systematic hyperparameter optimization during expert fine-tuning and merging.

6. Acknowledgments

The authors thank FAPESP (Research Grants 2024/17511-2 and 2025/21207-0) and Serasa Experian for financial support, the PARANA (PAttern Recognition And Network Analysis) research group at UNIFESP, and Dalhousie University for academic support and infrastructure. This research was enabled in part by support provided by Acenet and the Digital Research Alliance of Canada.

The authors utilized the Antigravity artificial intelligence tool to assist in drafting this document. The contributions of the tool include: (i) code generation assistance and (ii) refinement of the written text. All generated content was meticulously reviewed, and the authors remain solely and fully responsible for the technical accuracy, veracity, and integrity of the final presented text.

References

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 9–15.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Corrêa, N. K., Sen, A., Fatimah, S., Falk, S., Landgraf, L., Kastner, J., and Flek, L. (2026). Tucano 2 cool: Better open source llms for portuguese.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (2020). Revisiting pre-trained models for chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Freitas, C., Motta, E., Milidiú, R. L., and César, J. (2014). Sparkling vampire... lol! annotating opinions in a book review corpus. In Aluísio, S. and Tagnin, S. E. O., editors, *New Language Technologies and Linguistic Research: A Two-Way Road*, pages 128–146. Cambridge Scholars Publishing.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Goddard, C., Siriwardhana, S., Ehghaghi, M., Meyers, L., Karpukhin, V., Benedict, B., McQuade, M., and Solawetz, J. (2025). Arcee's mergekit: A toolkit for merging large language models.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. (2023). Editing models with task arithmetic.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Liu, B. (2020). *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge university press.

Liu, X., He, P., Chen, W., and Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Maddigan, P. et al. (2024). The rise of small language models: A survey of efficient architectures and applications. *arXiv preprint*.

Malo, P., Sinha, A., Korhonen, P. J., Wallenius, J., and Takala, P. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Muhammad, S. H. et al. (2025). Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Santos, L. L., Bianchi, R. A., and Costa, A. H. (2023). Finbert-pt-br: Análise de sentimentos de textos em português do mercado financeiro. In *Anais do II Brazilian Workshop on Artificial Intelligence in Finance*, pages 144–155. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained language models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–411. Springer.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. (2022). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.

Yadav, P., Tam, D., Choshen, L., Raffel, C., and Bansal, M. (2023). Ties-merging: Resolving interference when merging models.

Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. (2024). Language models are super mario: Absorbing abilities from homologous models as a free lunch.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.