Optimization strategies for BERT-based Named Entity Recognition

Monique Monteiro
mlbm@cin.ufpe.br

Abstract

Transfer learning through language modeling achieved state-of-the-art results for several natural language processing tasks. Despite impressive results, some tasks are still demanding more general solutions. This paper explores different approaches to improve named entity recognition (NER) transformer-based models pretrained at language modeling. We investigate model soups and domain adaptation methods for Portuguese language entity recognition. In particular, we evaluate different BERT-based models trained on different datasets considering general and specific domains. Our results show significant improvements when considering model soup techniques and in-domain pretraining compared to within-task pretraining. **Keywords**: NER, BERT, transfer learning, model soups, domain adapting.

1 INTRODUCTION

Named entity recognition (NER) is a core task for meaning extraction of textual content. Generally, it is a preprocessing step for building semantical representations necessary for more complex tasks, such as relation extraction and mention clustering. In the last years, advancements in transfer learning based on deep neural networks, such as LSTM [Howard and Ruder, 2018], or transformers [Vaswani et al., 2017] enabled substantial performance improvements for NLP models, especially in low-resource languages. In this context, transfer learning enables training by leveraging previous knowledge of a particular language's grammatical structure - that knowledge is embedded in models previously trained on unlabeled data through self-supervision. Language models are the most classic examples among these pretrained models. In particular, for the Portuguese language, the most commonly used model is BERTimbau [Souza et al., 2020], a neural network that replicates with minor changes the architecture and training procedures for BERT [Devlin et al., 2018], by using brWaC [Filho et al., 2018] as training corpus. BERTimbau was validated in specific tasks such as named entity recognition, textual similarity, and recognizing textual entailment, surpassing previously published results in the same way it occurs for similar works in other languages. Currently, a classical solution for constructing entity recognizers, classifiers, and other discriminative models for low-resource languages requires fine-tuning a pretrained language model (e.g., BERTimbau for Portuguese) in the target task by adding linear layers and adjusting the network weights by retraining with a new optimization objective. Such approach has achieved satisfactory results for most situations, at a low training cost ([Silva et al., 2021], [Monteiro, 2021b], [Monteiro, 2021a], [Guillou, 2022]).

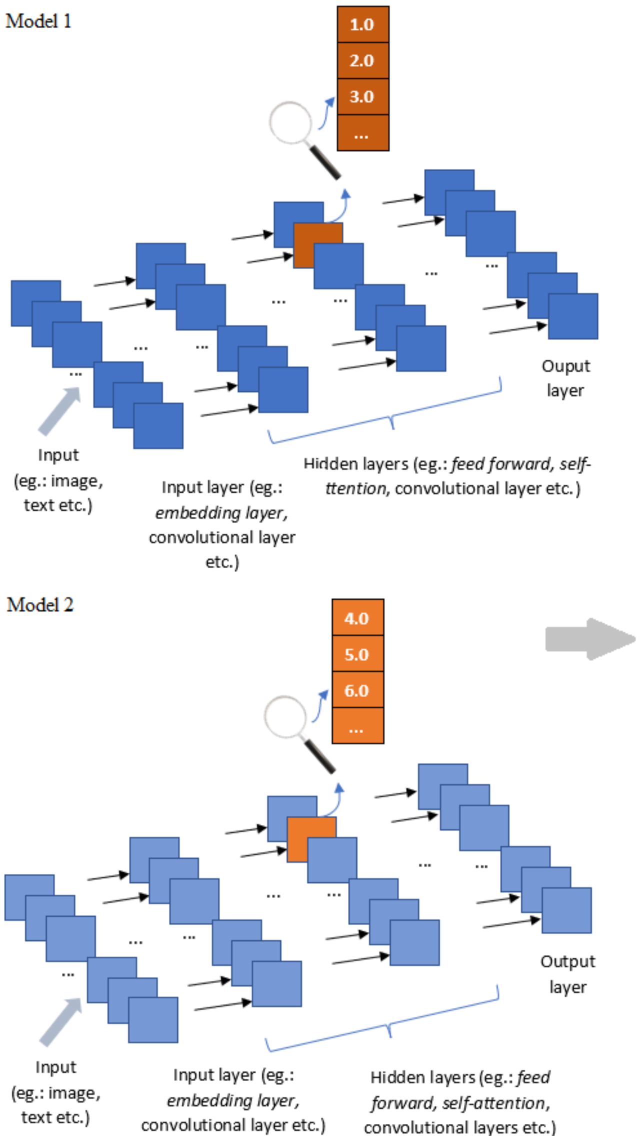
However, there are still not yet explored research fields, at least for entity recognition tasks in the Portuguese language: domain adaptation, one-shot/few-shot learning and recent techniques such as *model soups* [Wortsman et al., 2022].

Therefore, in this paper, we evaluate the applicability and possible improvements for two of the techniques cited in the last paragraph: model soups and additional training of a pretrained language model on an intermediary language model for posterior training in the target task (domain adaptation). To our knowledge, no other work in the literature investigates these techniques in the context of named entity recognition for a language such as Brazilian Portuguese. Therefore, it stresses the importance of research in natural language processing for low-resource languages.

This paper is organized in the following way: 2 introduces the techniques to be analyzed, while 3 and 4 present the detailed experiments and results, respectively.

2 RELATED WORK

Wortsman et al. (2022) propose a technique that consists of generating a model by averaging the weights of two or more trained models, in opposition to the traditional ap-


Model 1
1.0
2.0
3.0
...
Ouput layer
Input (eg.: image, text etc.)
Input layer (eg.: embedding layer, convolutional layer etc.)
Hidden layers (eg.: feed forward, self-ttention, convolutional layer etc.)
Model 2
4.0
5.0
6.0
...
Output layer
Input (eg.: image, text etc.)
Input layer (eg.: embedding layer, convolutional layer etc.)
Hidden layers (eg.: feed forward, self-attention, convolutional layers etc.)

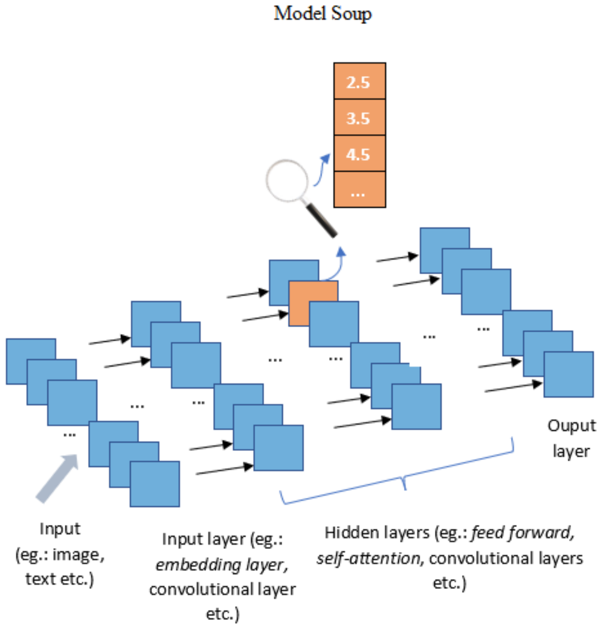

Model Soup
2.5
3.5
4.5
...
Ouput layer
Input (eg.: image, text etc.)
Input layer (eg.: embedding layer, convolutional layer etc.)
Hidden layers (eg.: feed forward, self-attention, convolutional layers etc.)

Figure 1: Mode soup - model creation by averaging weights of multiple models trained with different hyperparameter sets (uniform soup)

proach, which is based on 1) training multiple models with several hyperparameters and 2) choosing the model with the best performance on a validation set.

Also, the idea is to combine multiple models without the additional inference and memory costs related to traditional ensemble learning.

The authors refer to the technique as model soups, which is illustrated in Figure 1. It supports three variations: 1) construction of a model by averaging all models (*uniform soup*); 2) *greedy soup*, in which models are added sequentially to the soup as they improve the model accuracy in the validation dataset; and 3) *learned soup*, in which the interpolation weights for each model are optimized through gradient descent.

According to the authors, the greedy strategy showed the best results. At the same time, learned soups require loading in memory all the models simultaneously, generating more extensive networks and leading to little gain in accuracy. Their experiments were conducted on image and text classification.

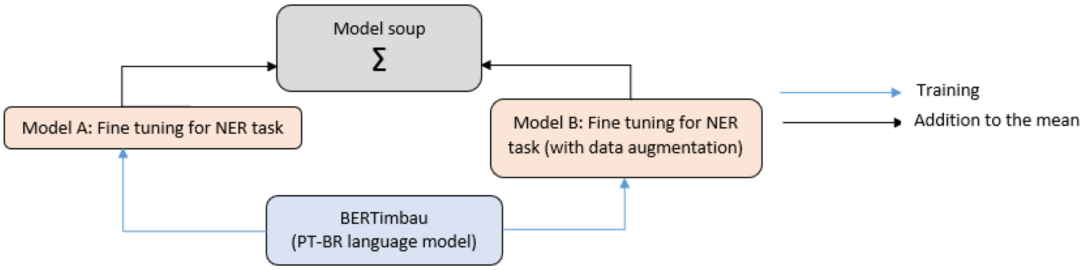

Model soup
Σ
Model A: Fine tuning for NER task
Model B: Fine tuning for NER task (with data augmentation)
BERTimbau
(PT-BR language model)
Training
Addition to the mean

Figure 2: Model soup - the original idea (adapted to the setup of two models trained on a NER task from a common initialization - in this case, the language model).

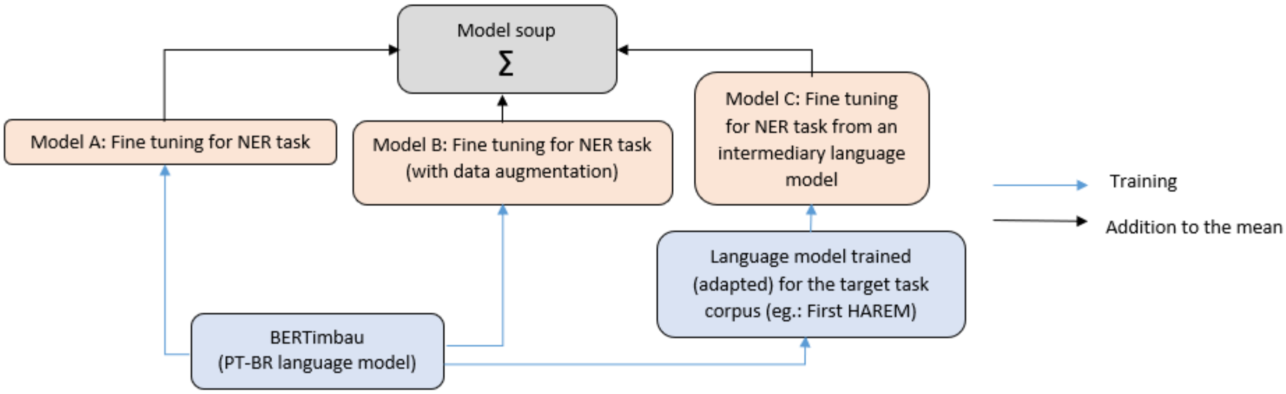

Model soup
Σ
Model A: Fine tuning for NER task
Model B: Fine tuning for NER task (with data augmentation)
Model C: Fine tuning for NER task from an intermediary language model
Language model trained (adapted) for the target task corpus (eg.: First HAREM)
BERTimbau
(PT-BR language model)
Training
Addition to the mean

Figure 3: Model soup - alternative version using two models trained from a common parameter set (BERTimbau language model) and a third model trained on an intermediary language model.

Sun et al. (2019) conducted experiments on fine-tuning BERT pretrained models on document classification. In particular, they analyzed domain adaptation - a different training approach with training data from the target task or the same domain. Additionally, they investigated the variation in learning rate per layer (a technique already used, for example, in the NER model used to validate BERTimbau) and multi-task learning.

3 EXPERIMENTS

3.1 Model Soups

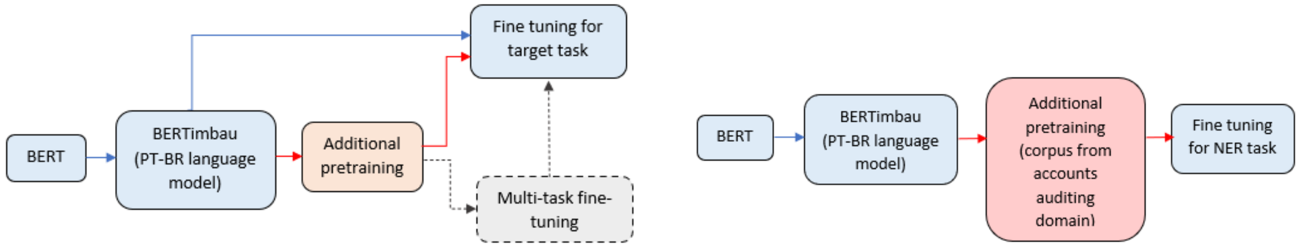
Creating a model by averaging other models' weights was validated on image and document classification tasks [Wortsman et al., 2022].

Formally, let $\theta = FineTune(\theta_0, h)$ the set of parameters obtained by fine-tuning the pretrained initialization θ_0 and the hyperparameters configuration h. The technique uses a mean value for θ_i, i.e., $\theta_S = \frac{1}{|S|}\sum_{i \in S} \theta_i$, where $S \subset \{1, ..., k\}$ and k is the number of hyperparameter configurations (or models) to be used.

Initially, we reproduced the idea using the uniform average strategy. However, the greedy approach was not analyzed due to our low number of available candidate models.

So, we created a model whose weights are the average of the weights of different models:

(A) the entity recognition model developed to validate BERTimbau ([Souza et al., 2020], [Souza et al., 2019]), trained on an adapted ver-


BERT
BERTimbau (PT-BR language model)
Additional pretraining
Fine tuning for target task
Multi-task fine-tuning
BERT
BERTimbau (PT-BR language model)
Additional pretraining (corpus from accounts auditing domain)
Fine tuning for NER task

(a)

(b)

Figure 4: (a) BERT fine-tuning options. Adapted from [Sun et al., 2019]. (b) Training flow for a NER specific to public accounts audit domain.

sion from the First HAREM [1] [Santos et al., 2006];

(B) model analogous to (A), but trained with data augmentation[2] and;

(C) model adapted to the same corpus for the target task (First HAREM), as described in Section 3.2.

We denote this model as the average model or model soup.

For the combinations cited above, we evaluated the BERT base and large variations, which differ in the number of parameters. Also, we evaluated setups with and without an additional layer for conditional random fields (CRF), which we used to reduce the probability of generating invalid label sequences.

Later, we analyzed a variation for the model soup technique, which consists of adding a fine-tuning step on the target task for the model soup instead of using it directly at inference time.

There is a difference between our experiments and the original proposal by Wortsman et al. (2022): the authors assume that the models were independently optimized from the same initialization, which would lead them to the same region or valley on the error surface. Such strategy is represented on Figure 2. But here, according to Figure 3, only models A and B start from the same initialization (BERTimbau language model). In contrast, model C was adapted to a second language model trained with First HAREM textual content.

The experiments' results are shown on Section 4.

3.2 Domain Adaptation

Figure 4a), adapted from Sun et al. (2019), shows the available options for domain adaptation/fine-tuning in our concrete case. Although it shows multi-task fine-tuning, Sun et al. (2019) conclude that its benefits tend to be inferior to those obtained from additional pretraining.

To conduct the experiments on domain adaptation, we use as start points:

- an entity recognition model for (long) documents related to public accounts auditing[3];
- the NER model trained for BERTimbau evaluation ([Souza et al., 2020], [Souza et al., 2019]);
- an entity recognition model [Silva et al., 2021] trained on a media public news dataset [Monteiro, 2021a].

In the first experiment, during domain adaptation, the original language model - BERTimbau - was trained on a dataset different from the one used to train the NER model. However, this dataset came from the same domain and origin (documents related to public accounts auditing), leading to an intermediary language model. As the training task was Masked Language Model (MLM), such dataset does not contain any label and can be considered a superset[4] for the entity recognition dataset. The complete flow is described in Figure 4b).

For the second and third experiments, during the construction of the intermediary language models, the respective datasets used for the NER task training were used: First, HAREM for the original BERTimbau NER model and the

[1] Here, the adapted version refers to a setup called "selective" by the authors, in which only 5 classes are considered (PERSON, ORGANIZATION, LOCAL, VALUE and DATE).

[2] Here, we used label-wise token replacement (LwTR) [Dai and Adel, 2020].

[3] We used during the model training the same hyperparameters for learning rate and batch size used by Silva et al. (2021).

[4] The dataset for the domain-adapted language model has 52,912 documents, against 431 documents for the labeled dataset used during entity recognition model training.

Table 1: Experiments with model soups for NER BERTImbau ("BERT$_{\text{BASE}}$" or "BERT$_{\text{LARGE}}$") trained on First HAREM

MODEL	**PRECISION (%)**	**RECALL (%)**	**F1 (%)**	↑
Baseline: Original BERT$_{\text{BASE}}$	82.1 ($\sigma = \pm 1.3$)	82.4 ($\sigma = \pm 0.6$)	82.3 ($\sigma = \pm 0.9$)	-
M1: BERT$_{\text{BASE}}$: original (best F1)	83.9	83.1	83.5	-
M2: BERT$_{\text{BASE}}$: dom. adapt. HAREM	82.9	83.0	82.9	-
M3: BERT$_{\text{BASE}}$ with data augmentation	82.2	82.4	82.3	-
M1 + M3	**85.5**	82.9	**84.2**	**+1.9**
M1 + M2 + M3	**86.1**	82.0	84.0	+1.7
M1 + M2 + M3 with additional fine-tuning	83.5 ($\sigma = \pm 1.1$)	83.0 ($\sigma = \pm 0.9$)	83.3 ($\sigma = \pm 0.8$)	+1.0
Baseline: Original BERT$_{\text{LARGE}}$	81.8 ($\sigma = \pm 1.2$)	82.4 ($\sigma = \pm 0.5$)	82.1 ($\sigma = \pm 0.5$)	-
M4: BERT$_{\text{LARGE}}$: original (best F1)	82.8	82.3	82.6	-
M5: BERT$_{\text{LARGE}}$: dom. adapt. HAREM	82.7	83.3	83.0	-
M6: BERT$_{\text{LARGE}}$ with data augmentation	84.2	82.8	83.5	-
M4 + M6	**86.0**	82.6	**84.3**	**+2.2**
M4 + M5 + M6	**87.1**	81.1	**84.0**	**+1.9**
Baseline: Original BERT$_{\text{BASE}}$ + CRF	84.1 ($\sigma = \pm 1.2$)	82.0 ($\sigma = \pm 0.6$)	83.0 ($\sigma = \pm 0.8$)	-
M7: BERT$_{\text{BASE}}$ + CRF: original (best F1)	85.0	82.3	83.6	-
M8: BERT$_{\text{BASE}}$ + CRF: dom. adapt. HAREM	82.7	83.3	83.0	-
M9: BERT$_{\text{BASE}}$ + CRF with data augmentation	86.2	81.8	83.9	-
M7 + M9	**87.5**	81.5	84.4	+1.4
M7 + M8 + M9	**87.8**	80.4	83.9	+0.9
Baseline: Original BERT$_{\text{LARGE}}$ + CRF	83.9 ($\sigma = \pm 1.2$)	82.0 ($\sigma = \pm 0.8$)	83.0 ($\sigma = \pm 0.6$)	-
M10: BERT$_{\text{LARGE}}$ + CRF: original (best F1)	84.7	82.8	83.7	-
M11: BERT$_{\text{LARGE}}$ + CRF: dom. adapt. HAREM	84.8	83.0	83.9	-
M12: BERT$_{\text{LARGE}}$ + CRF with data augmentation	86.1	82.0	84.0	-
M10 + M12	**87.2**	81.1	84.1	+1.1
M10 + M11 + M12	86.3	79.3	82.7	-0.3

Experiments with model soups: the strategy shown in Figure 2 (i.e.: two models initialized from the same weights and independently optimized) leads to better results than the alternative strategy shown in Figure 3. The base variant shows better results for precision and F1 metrics without additional fine-tuning, making unnecessary the extra training step. Values shown in bold font are above one standard deviation from the mean value for each metrics.

media news dataset. The labels were discarded for both datasets in the phase of MLM training.

The three resulting language models were used respectively as base models for retraining the three cited entity recognizers, aiming at measuring domain adaptation impact.

Learning rate and batch size hyperparameters for all intermediary language models training were the same as reported by Souza et al. (2020).

The experiments' results are shown on Section 4.

4 RESULTS AND DISCUSSIONS

4.1 Model Soup

In the first experiment, we evaluate the direct application of the model soup technique. We used a uniform average from a set composed of the respective best models (i.e., best training) among the variations described in 3.1.

Later, we evaluated the second setup, in which the model soup received additional fine-tuning for the NER task. As shown in Table 1, for the smaller variant in model size (base), the first model, without additional fine-tuning, shows better results for precision and F1 metrics, making unnecessary the extra training step.

As already described on 3.1, initially, for each model size (base/large), we added the following components to each combination: (A) original BERTimbau NER; (B) BERTimbau NER retrained with data augmentation; and (C) BERTimbau NER retrained after domain adaptation to First HAREM (original BERTimbau language model fine-tuned to First HAREM text set), as described on 3.2).

Later, the third variant (C) was removed from each combination, leading to the original schema shown in Figure 2. Finally, we observed that the combinations based only on (A) (original NER) and (B) (data augmented NER) led to better values for precision and F1 metrics, confirming Wortsman's (2022) original hypothesis of using indepen-

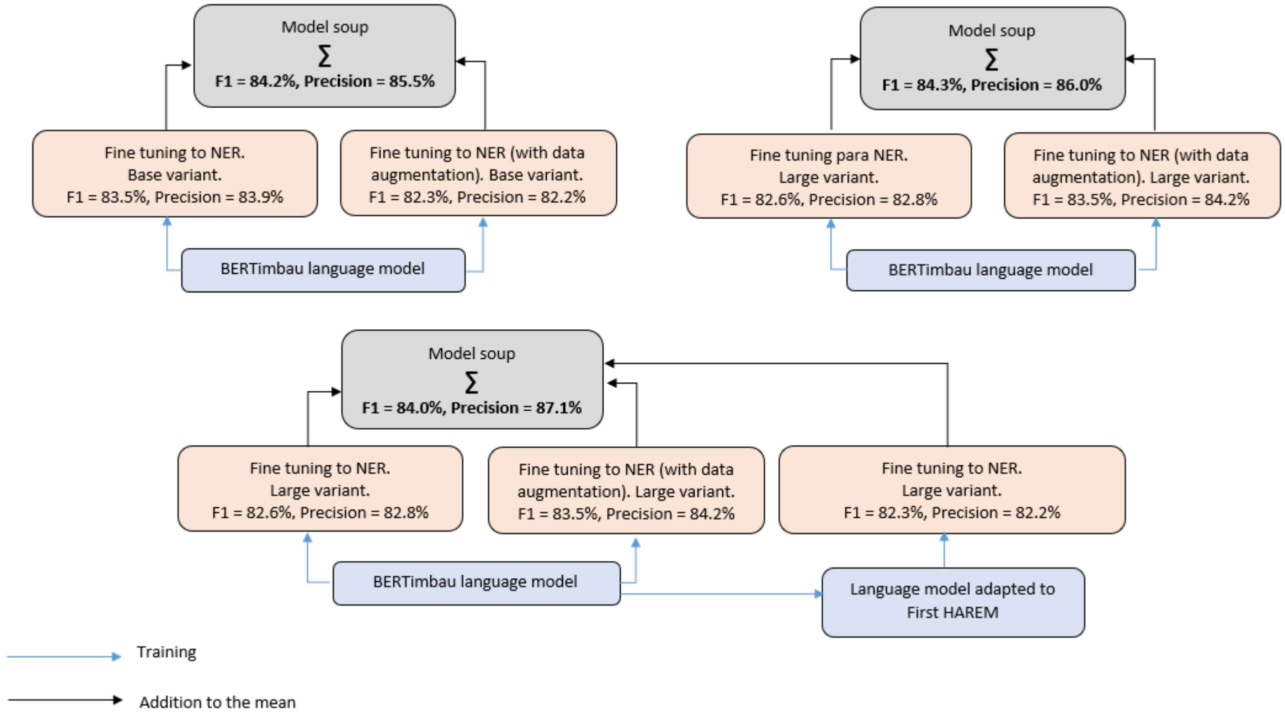

Model soup
∑
F1 = 84.2%, Precision = 85.5%
Fine tuning to NER.
Base variant.
F1 = 83.5%, Precision = 83.9%
Fine tuning to NER (with data augmentation). Base variant.
F1 = 82.3%, Precision = 82.2%
BERTimbau language model
Model soup
∑
F1 = 84.3%, Precision = 86.0%
Fine tuning para NER.
Large variant.
F1 = 82.6%, Precision = 82.8%
Fine tuning to NER (with data augmentation). Large variant.
F1 = 83.5%, Precision = 84.2%
BERTimbau language model
Model soup
∑
F1 = 84.0%, Precision = 87.1%
Fine tuning to NER.
Large variant.
F1 = 82.6%, Precision = 82.8%
Fine tuning to NER (with data augmentation). Large variant.
F1 = 83.5%, Precision = 84.2%
Fine tuning to NER.
Large variant.
F1 = 82.3%, Precision = 82.2%
BERTimbau language model
Language model adapted to First HAREM
Training
Addition to the mean

Figure 5: Summary for the best results obtained with model soup technique. Here are shown the combinations which lead to gains above one standard deviation from baseline - further details on Table 1.

dently optimized models from the same initialization.

The best results are summarized in Figure 5.

When we compare our methodology with the one used by the authors [Wortsman et al., 2022], they use accuracy in most experiments with image and text classification. Here, it does not make sense to use accuracy because it tends to be too high in entity recognition tasks due to the inherently unbalanced datasets (e.g., the high number of “non-entity” tokens). For example, the authors do not refer explicitly to recall, which shows worse results in our experiments.

So, further investigation is needed about the reason recall is worsening by the use of the model soup technique, which at the moment makes us believe that such method could be more suitable to situations in which precision is more important than getting a high number of entities or false positive cases.

Another experiment analyzed the use of F_β metrics during the component models (re)training. It was used on the validation set to modify the weights given to recall and precision. We used 2.0 to β, assigning a higher weight to recall. We changed the metrics instead of the error function, so it has only indirect influence because, during training, the model is saved with the weights for the best epoch according to the target metrics (F1 or F_β) on the validation set. As a result, despite improvement for recall in the component models, the final model soup kept the worsening pattern for recall, leading to worse F_β values. Yet F1 was improved for the model soup. Detailed results are shown in the supplementary material.

On the other hand, given known limitations in the use of precision-recall-F1 for entity recognition, better and more interpretable metrics for this task are a current research topic [Fu et al., 2020].

Finally, according to Wortsman et al. (2022), preliminary experiments show improvement in the textual corpus, although present, is not so profound as in image classification. The authors stress the need for additional investigation into this subject. They used ImageNet and several variations, including illustrations and other representations beyond real photos (e.g.: drawings, arts, cartoons, tattoos, origami, sculptures etc.). But for textual classification, they used general domain datasets for paraphrase detection, coherence vs. contradiction, linguistic acceptability, and sen-

Table 2: Domain adaptation for documents related to public accounts audit.

MODEL	PRECISION (%)	RECALL (%)	MASKED F1 (%)	↑
$\text{BERT}_{\text{BASE}}$: original	82.2 ($\sigma = \pm 0.4$)	89.1 ($\sigma = \pm 0.3$)	86.0 ($\sigma = \pm 0.1$)	-
$\text{BERT}_{\text{BASE}}$: domain adapted	82.4 ($\sigma = \pm 0.3$)	89.4 ($\sigma = \pm 0.8$)	**86.5** ($\sigma = \pm 0.4$)	+0.5

Domain adaptation for documents related to public accounts audit: here, we used in-domain adaptation and achieved the best result among the experiments, in contrast to within-task adaptation. For Masked F1, the improvement was exactly on the standard deviation frontier. The values for precision, recall and Masked F1 are the mean values from 5 runs with random initialization.

Table 3: Domain adaptation for media news.

MODEL	PRECISION (%)	RECALL (%)	MASKED F1 (%)	↑
$\text{BERT}_{\text{BASE}}$: original	85.1 ($\sigma = \pm 0.7$)	87.6 ($\sigma = \pm 1.2$)	86.9 ($\sigma = \pm 0.7$)	-
$\text{BERT}_{\text{BASE}}$: domain adapted	85.3 ($\sigma = \pm 0.7$)	87.7 ($\sigma = \pm 1.0$)	87.2 ($\sigma = \pm 0.6$)	+0.3

Domain adaptation for media news: within-task domain adaptation led to little improvement for Masked F1 (below one standard deviation).

Table 4: Domain adaptation for First HAREM.

MODEL	PRECISION (%)	RECALL (%)	F1 (%)	↑
$\text{BERT}_{\text{BASE}}$: original	82.1 ($\sigma = \pm 1.3$)	82.4 ($\sigma = \pm 0.6$)	82.3 ($\sigma = \pm 0.9$)	-
$\text{BERT}_{\text{BASE}}$ + CRF: original	84.1 ($\sigma = \pm 1.2$)	82.0 ($\sigma = \pm 0.6$)	83.0 ($\sigma = \pm 0.8$)	-
$\text{BERT}_{\text{LARGE}}$: original	81.8 ($\sigma = \pm 1.2$)	82.4 ($\sigma = \pm 0.5$)	82.1 ($\sigma = \pm 0.5$)	-
$\text{BERT}_{\text{LARGE}}$ + CRF: original	83.9 ($\sigma = \pm 1.2$)	82.0 ($\sigma = \pm 0.8$)	83.0 ($\sigma = \pm 0.6$	-
$\text{BERT}_{\text{BASE}}$: domain adapted	81.9 ($\sigma = \pm 1.0$)	82.2 ($\sigma = \pm 0.8$)	82.0 ($\sigma = \pm 0.9$)	-0.3
$\text{BERT}_{\text{BASE}}$ + CRF: domain adapted	85.1 ($\sigma = \pm 0.7$)	81.6 ($\sigma = \pm 2.0$)	83.3 ($\sigma = \pm 1.1$)	+0.3
$\text{BERT}_{\text{LARGE}}$: domain adapted	81.0 ($\sigma = \pm 2.3$)	82.8 ($\sigma = \pm 0.9$)	81.9 ($\sigma = \pm 1.5$)	-0.2
$\text{BERT}_{\text{LARGE}}$ + CRF: domain adapted	84.9 ($\sigma = \pm 0.8$)	81.7 ($\sigma = \pm 0.8$)	83.3 ($\sigma = \pm 0.4$)	+0.3

Domain adaptation for First HAREM: within-task domain adaptation led to little improvement for Masked F1 (below one standard deviation) in some setups, and to worse results in others.

timent analysis.

Preliminary and qualitative analysis of the different data for images vs. texts shows more variability and larger data size for the first case (e.g., ImageNet contains millions of images), which could have led to a more significant impact on image classification.

4.2 Domain adaptation

In this section, we report results achieved by NER models when trained over intermediary language models, as described on 3.2.

The results of the experiment carried out with the NER model trained on documents related to public accounts auditing are shown in Table 2. As comparison metrics, Masked F1 [Silva et al., 2021] was used. This metric is F1 calculated over post-processed output, correcting invalid transitions according to the IOB2 schema instead of using the raw output directly. This setup - based on in-domain adaptation - led to the most pronounced improvements.

On the experiment conducted on the NER model for media news ([Silva et al., 2021], [Monteiro, 2021b], [Monteiro, 2021a]), the results are shown on Table 3. It is important to say that the experiment was conducted only with variants based on a pretrained Brazilian Portuguese language model (BERTimbau) because multi-language models gave an inferior performance, according to Silva et al. (2021). Masked F1 was also used as the main comparison metric.

Table 4 shows the achieved results with the NER model used in BERTimbau evaluation ([Souza et al., 2020], [Souza et al., 2019]).

As can be noted in Tables 3 and 4, experiments conducted with the media news NER and BERTimbau NER did not reveal significant differences after domain adaptation. Such results confirm observations by Sun et al. (2019) because the domain adaptation made on the second and third experiments (media news and First HAREM) is “within-task”, in which the texts used are the same as the training texts for the target task: in general, “in-domain” pretraining, i.e.,

Table 5: Domain adaptation for First HAREM - qualitative analysis.

TARGET CLASSIFICATION	**BERT$_{\text{BASE}}$ + CRF**	**DOM. ADAPT. BERT$_{\text{BASE}}$ + CRF**
"... que costumavam comer na noite de Consoada $_{\text{TIME}}$*"* (...who used to eat at **Consoada** night)	*"... que costumavam comer na noite de Consoada* $_{\text{LOC}}$*"*	*"... que costumavam comer na noite de Consoada* $_{\text{TIME}}$*"*
"Gostei muito da feira do Soajo $_{\text{LOC}}$*"* (I enjoyed **Soajo** Fair a lot)	*"Gostei muito da feira do Soajo* $_{\text{PERSON}}$*"*	*"Gostei muito da feira do Soajo* $_{\text{LOC}}$*"*
"E qual a lembrança mais antiga da cidade? São João do Souto $_{\text{LOC}}$*"* (And what's the oldest city's memory? **São João do Souto**)	*"E qual a lembrança mais antiga da cidade? São João do Souto"*	*"E qual a lembrança mais antiga da cidade? São João do Souto* $_{\text{LOC}}$*"*
"... comecei a fazer trabalhos na Abade da Loureira $_{\text{ORG}}$*"* (... I started to work for **Abade da Loureira**)	*"... comecei a fazer trabalhos na Abade da Loureira* $_{\text{LOC}}$*"*	*"... comecei a fazer trabalhos na Abade da Loureira* $_{\text{ORG}}$*"*
"Tirei o curso de formação no Centro de Formação de Informática do Minho $_{\text{ORG}}$*"* (I took the graduation course at **Minho Informatics Center**)	*"Tirei o curso de formação no Centro de Formação de Informática do Minho* $_{\text{LOC}}$*"*	*"Tirei o curso de formação no Centro de Formação de Informática do Minho* $_{\text{ORG}}$*"*

The second column shows outputs generated by the baseline NER (with CRF), while the third column shown outputs from the NER trained on the intermediary language model (domain-adapted NER). While the first one misclassifies portuguese expressions, the second one labels them correctly. All the examples belong to First HAREM test dataset.

using texts from the same domain which are different from the texts in the training dataset for the target task, gives superior results. We suspect that within-task pretraining could lead to overfitting because it uses the same texts from the target task dataset.

On the other hand, after error analysis, we realized that some Portuguese linguistic expressions, organizations, and local names could be correctly classified after domain adaptation. At the same time, they were misclassified when NER was trained directly from the raw BERTimbau language model. The results are shown on Table 5. It makes sense because First HAREM is a Portuguese corpus, different from the Brazilian corpus used to train BERTimbau. These results show semantic gains from domain adaptation, although quantitative performance differences are not statistically significant. Further details are discussed in the supplementary material.

Further, the results shown here were obtained in the NER task context. For document classification, within-task pretraining has been commonly used ([Howard and Ruder, 2018], [Guillou, 2022]).

A final experiment intended to investigate the overfitting hypothesis through data augmentation to the language model training dataset. Such experiment used only BASE model variants and did not show a significant impact because the data augmentation technique replaced only 10% words with the same entity-type tokens, keeping the texts similar to the ones in the target task dataset. Detailed results are shown in the supplementary material.

Finally, we showed that domain adaptation in the first experiment - by training an intermediary language model with a larger, same-domain dataset - led to a higher impact on F1 metrics when compared to the experiments with within-task domain adaptation (second and third experiments). Furthermore, qualitative analysis for the intermediary language model (presented as supplementary material) shows example outputs for predicting masked term tasks, i.e., the public accounts auditing language model can generate texts related to themes such as contracts and biddings.

5 CONCLUSIONS

Among the analyzed techniques, the model soup technique achieved the best results for the named entity recognition (NER) task.

During experiments conducted on domain adaptation, the best results were achieved with in-domain adaptation. We did not observe quantitatively significant improvements for within-task domain adaptation, at least for the NER task. However, we realized the model could learn domain-specific terms with the First HAREM corpus. We believe that advances in one-shot/zero-shot learning and prompt engineering could bring competitive results compared to domain adaptation.

References

[Dai and Adel, 2020] Dai, X. and Adel, H. (2020). An analysis of simple data augmentation for named entity recognition. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3861–3867. International Committee on Computational Linguistics.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Filho et al., 2018] Filho, J. A. W., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

[Fu et al., 2020] Fu, J., Liu, P., and Neubig, G. (2020). Interpretable multi-dataset evaluation for named entity recognition. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6058–6069. Association for Computational Linguistics.

[Guillou, 2022] Guillou, P. (2022). Modelos e web app para reconhecimento de entidade nomeada (ner) no domínio jurídico brasileiro. https://link.medium.com/IDioCVUZUqb. Last accessed on 2022-05-22.

[Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

[Monteiro, 2021a] Monteiro, M. (2021a). Extrator de entidades mencionadas em notícias da mídia. https://github.com/SecexSaudeTCU/noticias_ner. Last accessed on 2022-05-21.

[Monteiro, 2021b] Monteiro, M. (2021b). Riskdata brazilian portuguese ner. https://huggingface.co/monilouise/ner_news_portuguese. Last accessed on 2022-05-21.

[Santos et al., 2006] Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). HAREM: an advanced NER evaluation contest for portuguese. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 1986–1991. European Language Resources Association (ELRA).

[Silva et al., 2021] Silva, E. H. M. D., Laterza, J., Silva, M. P. P. D., and Ladeira, M. (2021). A proposal to identify stakeholders from news for the institutional relationship management activities of an institution based on named entity recognition using BERT. In Wani, M. A., Sethi, I. K., Shi, W., Qu, G., Raicu, D. S., and Jin, R., editors, *20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021, Pasadena, CA, USA, December 13-16, 2021*, pages 1569–1575. IEEE.

[Souza et al., 2020] Souza, F., Nogueira, R., and de Alencar Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer.

[Souza et al., 2019] Souza, F., Nogueira, R. F., and de Alencar Lotufo, R. (2019). Portuguese named entity recognition using BERT-CRF. *CoRR*, abs/1909.10649.

[Sun et al., 2019] Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune BERT for text classification? In Sun, M., Huang, X., Ji, H., Liu, Z., and Liu, Y., editors, *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 of *Lecture Notes in Computer Science*, pages 194–206. Springer.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

[Wortsman et al., 2022] Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Lopes, R. G., 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. *CoRR*, abs/2203.05482.

Optimization strategies for BERT-based Named Entity Recognition: Supplementary Materials

1 ADDITIONAL EXPERIMENTS

1.1 Model Soups: The Effect of F_β Metrics during the Component Models (Re)training

This experiment analyzed the use of F_β metrics during the component models (re)training. It was used on the validation set to modify the weights given to recall and precision. We set 2.0 to β, assigning a higher weight to recall. During training, the model is saved with the weights for the best epoch according to the target metrics (F1 or F_β) on the validation set. Table 1 below shows the effect of this change on precision, recall, F_β and F1, in comparison to "Baseline: Original BERT_{LARGE} + CRF". As we can see, despite improvement for recall in the component models, the final model soup kept the worsening pattern for recall, leading to worse F_β values. Yet F1 was improved for the model soup. The values for precision, recall, F_β and F1 are the mean values from 5 runs with random initialization (σ stands for standard deviation).

Table 1: Experiments on F_β metrics.

MODEL	PRECISION (%)	RECALL (%)	F_β (%)	F1 (%)	↑
M10: $\text{BERT}_{\text{LARGE}}$ + CRF: original (best F_β)	85.3 ($\sigma = \pm0.3$)	83.1 ($\sigma = \pm1.7$)	83.5 ($\sigma = \pm1.4$)	-	-
M12: $\text{BERT}_{\text{LARGE}}$ + CRF with data augm.	85.9 ($\sigma = \pm1.2$)	83.7 ($\sigma = \pm0.7$)	84.2 ($\sigma = \pm0.7$)	-	-
M10 + M12	86.5	81.5	82.4	84.0	+1.0

1.2 Domain Adaptation on Augmented First HAREM

This experiment intended to investigate the overfitting hypothesis through data augmentation to the language model training dataset. As we can see in Table 2, the impact on F1 metrics was not significant. As the data augmentation technique replaced only 10% of words with the same entity-type tokens, we suspect that it kept the texts very similar to the ones in the target task dataset. The values for precision, recall and F1 are the mean values from 5 runs with random initialization (σ stands for standard deviation).

Table 2: Domain Adaptation on Data Augmented First HAREM.

MODEL	PRECISION (%)	RECALL (%)	F1 (%)	↑
$\text{BERT}_{\text{BASE}}$: original	82.1 ($\sigma = \pm1.3$)	82.4 ($\sigma = \pm0.6$)	82.3 ($\sigma = \pm0.9$)	-
$\text{BERT}_{\text{BASE}}$ + CRF: original	84.1 ($\sigma = \pm1.2$)	82.0 ($\sigma = \pm0.6$)	83.0 ($\sigma = \pm0.8$)	-
$\text{BERT}_{\text{BASE}}$: domain adapted	82.1 ($\sigma = \pm0.4$)	82.9 ($\sigma = \pm0.4$)	82.5 ($\sigma = \pm0.3$)	+0.2
$\text{BERT}_{\text{BASE}}$ + CRF: domain adapted	84.1 ($\sigma = \pm1.0$)	80.7 ($\sigma = \pm1.1$)	82.4 ($\sigma = \pm0.7$)	-0.6

1.3 Samples from Public Accounts Auditing Language Model

The qualitative analysis for the intermediary language model trained on documents about public accounts auditing shows it can generate texts related to themes such as contracts and bidding, as seen in Table 3.

Table 3: Sample outputs by a generic domain LM vs a specific accounts auditing domain LM.

LM TRAINED ON FIRST HAREM	**LM TRAINED ON ACCOUNTS AUDITING DOMAIN**
QUESTION 1: ESTE É UM GRANDE (THIS IS A BIG) <?>	
Este é um grande desafio	*Este é um grande administrador*
(This is a big challenge)	(This is a great administrator)
Este é um grande passo	*Este é um grande empresário*
(This is a big step)	(This is a great enterpreneur)
Este é um grande tesouro	*Este é um grande produtor*
(This is a big treasure)	(This is a great producer)
QUESTION 2: ESTA É UMA GRANDE (THIS IS A BIG) <?>	
Esta é uma grande vitória	*Esta é uma grande empresa*
(This is a big victory)	(This is a big company)
Esta é uma grande oportunidade	*Esta é uma grande economia*
(This is a great opportunity)	(This is a huge economy)
Esta é uma grande conquista	*Esta é uma grande concorrência*
(This is a great achievement)	(This is a big bid)

2 QUALITATIVE ANALYSIS ON DOMAIN ADAPTATION TO FIRST HAREM

As we showed in Section 4.2, Table 5, some Portuguese linguistic expressions, organizations, and local names could be correctly classified only after BERTimbau Brazilian language model domain adaptation on First HAREM, a Portuguese corpus. In this section, we explain the terms shown in Table 5 (main paper).

- *1st row: In Portugal, "Consoada" refers to Christmas Night, which means it should be classified as a temporal (TIME) mention.*
- *2nd row: "Soajo" is the name of a Portuguese village.*
- *3rd row: "Abade da Loureira" refers to both an organization (ORG) and a street (LOC). But given the specific context in the sentence, it should be classified as an organization (ORG).*
- *4th row: "Centro de Formação de Informática do Minho" is a local educational institution.*
- *5th row: "São João do Souto" refers to an extinct Portuguese parish.*