Evaluating Domain-adapted Language Models for Governmental Text Classification Tasks in Portuguese

Mariana O. Silva[1], **Gabriel P. Oliveira**[1], **Lucas G. L. Costa**[1], **Gisele L. Pappa**[1]

[1] Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

{mariana.santos,gabrielpoliveira}@dcc.ufmg.br
lucas-lage@ufmg.br, glpappa@dcc.ufmg.br

Abstract. *Domain-adaptive pre-training (DAPT) is a technique in natural language processing (NLP) that tailors pre-trained language models to specific domains, enhancing their performance in real-world applications. In this paper, we evaluate the effectiveness of DAPT in governmental text classification tasks, exploring how different factors, such as target domain dataset, pre-trained model language composition, and dataset size, impact model performance. We systematically vary these factors, creating distinct domain-adapted models derived from BERTimbau and LaBSE. Our experimental results reveal that selecting appropriate target domain datasets and pre-training strategies can notably enhance the performance of language models in governmental tasks.*

1. Introduction

In an era where digital transformation permeates all sectors, government institutions are increasingly challenged by a growing volume of data across various tasks in the public sector. Such data enable the development of data-driven solutions, insights, and analyses across various domains, including economics, public health, urban planning, and environmental science [Oliveira et al. 2022, Silva et al. 2023]. Leveraging such data not only promotes transparency, accountability, and public participation but also enriches democratic processes by granting citizens access to insights regarding government activities, expenditures, and decision-making processes [Brandão et al. 2024].

Indeed, the continuous production of new information in the public sector highlights the need for automated processing of such data. For example, governmental institutions have increasingly used natural language processing (NLP) techniques to streamline operations and extract insights from vast repositories of textual data [Luz de Araujo et al. 2020]. In particular, language models have emerged as essential tools for automating tasks such as document classification, information extraction, and topic modeling [Silveira et al. 2021, Constantino et al. 2022, Brandão et al. 2023]. However, the effectiveness of such models relies on their ability to comprehend and accurately interpret the intricacies of the language specific to the domain in which they are applied.

Regarding governmental applications in the Portuguese language, adapting language models to domain-specific tasks is particularly challenging due to the complexities of administrative, legal, and bureaucratic discourse. Traditional language models trained on generic corpora may struggle to comprehend the specialized vocabulary, syntax, and conventions prevalent in government documents, leading to suboptimal performance and unreliable results. Therefore, addressing this challenge requires the development of domain-adapted language models tailored to the unique linguistic characteristics of governmental applications in Portuguese [Silva et al. 2021, Hott et al. 2023].

Domain adaptation usually involves tailoring pre-existing models to perform effectively within a specific context by continued pre-training language models on domain-specific unlabeled data [Gururangan et al. 2020]. However, efficient domain adaptation relies on multiple factors, including choosing an appropriate target domain dataset, the pre-trained model's language composition and the dataset size used [Zhu et al. 2021, Singhal et al. 2023]. Therefore, in this paper, we assess the impact of such factors on the performance of text classification tasks within the governmental domain. Specifically, this work is guided by the following research questions (RQs):

RQ1. Domain Relevance. *How does the target domain dataset used in pre-training impact the performance of text classification tasks within the governmental domain?*

RQ2. Language Composition. *How does the language composition (monolingual vs. multilingual) of the pre-trained model used in pre-training influence the performance of text classification tasks within the governmental domain?*

RQ3. Dataset Size. *How does the size of the target domain dataset used in pre-training impact the performance of text classification tasks within the government domain?*

The remainder of this paper is organized as follows. After discussing related work in Section 2, we delve into domain-adaptive pre-training and its influencing factors in Section 3. Next, the methods and datasets used to create the domain-adapted language models are outlined in Section 4. Our experimental setup and results are detailed in Section 5. Finally, we discuss our work's main conclusions and limitations in Section 6, along with avenues for future work.

2. Related Work

In the governmental domain, natural language processing (NLP) methods have been increasingly used to streamline operations and derive insights from vast repositories of textual data [Brandão et al. 2024]. These methods encompass a range of techniques, with language models standing out as crucial tools for automating tasks, including document classification [Luz de Araujo et al. 2020, Brandão et al. 2023], information extraction [Luz de Araujo et al. 2018], and topic modeling [Constantino et al. 2022, Hott et al. 2023]. Yet, the effectiveness of these models depends on their capacity to comprehend domain-specific language nuances effectively within the governmental context.

This challenge is particularly pertinent for the Portuguese language, where the complexities of governmental jargon and bureaucratic language pose unique obstacles to NLP applications [Oliveira et al. 2022, Brandão et al. 2024, Silva et al. 2023]. Despite advancements in general-purpose language models, such as BERT and GPT, their off-the-shelf performance may not meet the demands of governmental tasks due to the specialized nature of this domain [Gururangan et al. 2020]. Therefore, tailoring language models to understand better domain-specific language has become imperative for achieving optimal performance in specific applications, including governmental ones [Silva et al. 2021, Silveira et al. 2023].

Among the existing techniques in NLP, domain-adaptive pre-training (DAPT) emerges as a promising approach for bolstering the performance of language models in specific domains [Gururangan et al. 2020]. DAPT involves continuing the pre-training process on domain-specific unlabeled data, enabling models to adapt to the linguistic characteristics of a particular domain. This technique has gained traction in various domains, including literary [Silva and Moro 2024], clinical [Schneider et al. 2020], oil and gas [Rodrigues et al. 2022], and legal domains [Silva et al. 2021, Silveira et al. 2023], where specialized language and terminology are prevalent, showcasing its potential for enhancing language models in the governmental domain.

However, efficient domain adaptation may rely on several factors, such as selecting an appropriate target domain dataset, choosing the pre-trained model architecture and language composition, and determining the quantity and quality of data available for pre-training [Singhal et al. 2023]. [Gururangan et al. 2020] investigated several variations for adapting pre-trained language models to new domains and highlighted the importance of domain-specific data in achieving effective adaptation. Their findings suggest that even small amounts of domain-specific data can significantly improve model performance when carefully selected. Similarly, [Feijó and Moreira 2020] explored the impact of multilingual versus monolingual pre-training on model performance, finding that while multilingual models offer broad language coverage, monolingual models can sometimes outperform them in specific language tasks due to their specialized training.

Compared to existing studies, our investigation uniquely focuses on systematically evaluating the impact of these factors in the context of Portuguese governmental applications. We provide a comprehensive analysis of how target domain dataset selection, pre-trained model language composition, and dataset size influence model efficacy, offering a nuanced view of the factors contributing to successful domain adaptation. Our findings emphasize that while DAPT holds significant promise, achieving optimal performance in specialized domains requires carefully balancing various factors, including dataset relevance, model architecture, and data quantity. This study thus contributes to the ongoing efforts to refine NLP techniques for domain-specific applications, particularly within the governmental sector.

3. Domain-Adaptive Pre-training

Domain-adaptive pre-training (DAPT) is a natural language processing (NLP) technique that integrates domain-specific knowledge into pre-trained language models, enhancing their performance in targeted real-world applications [Gururangan et al. 2020]. DAPT involves continued pre-training language models on domain-specific unlabeled data, incorporating domain-specific features or vocabulary. Several critical factors must be considered for efficient domain adaptation, including data domain and quantity, as well as the language of the pre-trained models. Next, we briefly discussed each of these factors.

Domain Relevance. Choosing an appropriate target domain dataset is essential for successful domain adaptation through pre-training [Gururangan et al. 2020]. The dataset should accurately reflect the target domain's linguistic characteristics, vocabulary, and nuances to facilitate effective adaptation.

Language Composition. The pre-trained model's language should align with the language(s) present in the target domain to facilitate more efficient adaptation [Feijó and Moreira 2020]. A multilingual model offers advantages by capturing language-specific features and nuances from multiple languages in the target domain, thus improving its adaptability. However, it is worth noting that a multilingual model may not capture domain-specific nuances as effectively as a monolingual model pre-trained specifically on the target language.

Dataset Size. It is important to have enough representative data from the target domain to enable the model to learn domain-specific patterns effectively. A larger dataset can help the model capture a broader range of language variations and domain-specific contexts, improving its ability to perform well in the target domain [Singhal et al. 2023]. However, it is worth noting that including more data for DAPT does not necessarily guarantee improved performance in downstream tasks [Zhu et al. 2021].

4. Methods and Data

To assess the relative importance of each domain adaptation factor, we create 18 domain-adapted language models by systematically varying these factors and evaluating their performance across two governmental tasks. Specifically, we consider three different-sized domain-specific datasets that reflect the linguistic characteristics of the governmental domain, each differing in scale. Moreover, we continuously pre-trained on two types of language models for each dataset: one configured for a single language and the other with a multilingual composition.

This section outlines the methods and data used to create the different domain-adapted language models. First, in Sections 4.1 and 4.2, we detail the downstream tasks and the pre-training datasets considered in our evaluation. Next, in Section 4.3, we describe the language models selected for domain-adaptive pre-training. Finally, in Section 4.4, we describe the 18 domain-adapted language models designed to explore the importance of critical factors in domain adaptation.

4.1. Downstream Tasks

We focus on two distinct text classification tasks within the governmental domain to define our downstream tasks. These tasks are deliberately selected to represent common scenarios faced in governmental applications, as described below.

Document Classification. Involves categorizing documents into predefined classes based on their content, facilitating efficient organization and retrieval of information within governmental datasets. By automatically classifying documents into relevant categories, governmental agencies can streamline document management processes, improve information accessibility, and enhance overall operational efficiency [Brandão et al. 2024].

Item Classification. In this task, individual items are classified based on their descriptions, distinguishing their nature or type (i.e., product and service). Such classification enables precise analysis and differentiation among various offerings within governmental documents. By automatically classifying items into these categories, governmental agencies can gain insights into their procurement activities, track expenditures, and ensure compliance with procurement regulations [Silva et al. 2023].

4.2. Pre-training Data

As previously stated, we consider three pre-training datasets, each selected to capture the domain-specific characteristics prevalent in diverse governmental documents and product/service descriptions. The first dataset comprises segments extracted from official gazettes, while the second one contains legal documents related to extraordinary appeals the Brazilian Supreme Court received. The third dataset comprises expenditure items from political campaigns. We generate three versions with varying scales (small, medium, and large) for each dataset to evaluate their impact on the domain adaptation process and subsequent task performance. The cut-off thresholds for these corpora size scales are established based on the availability and distribution of data within each dataset. The small versions represent approximately one-third of the total data, the medium versions include two-thirds, and the large versions encompass the entire dataset. Such stratification ensures a systematic comparison of how dataset size influences the effectiveness of domain-adaptive pre-training. Next, each dataset is further described.

Official Gazette Segments [Constantino et al. 2022]. It contains segments related to the public bidding process of several municipalities in Minas Gerais, Brazil. Such segments represent summarized texts published within the official gazette to provide information regarding the bidding process, including announcements, notices, and bidding terms. The complete dataset contains 917,920, with a total of 75,621,362 tokens. To provide a comprehensive analysis, we create three versions of varying sizes: small (313,824 segments), medium (627,649 segments), and large (917,920 segments).

VICTOR [Luz de Araujo et al. 2020]. It contains legal documents related to extraordinary appeals received by the Brazilian Supreme Court (STF). Originally, the VICTOR dataset contained three different versions, varying in size. Here, we only use the Medium VICTOR (MVic), which contains 1,760,862 unlabeled legal texts, totaling 284,032,540 tokens. From MVic, we also generate three versions of varying sizes: small (440,216 texts), medium (880,431 texts), and large (1,760,862 texts).

Electoral Expense Items. It comprises expenditure items from election campaigns in Brazil, published by the Brazilian Superior Electoral Court (TSE). It contains data from the 2018 municipal and the 2020 general elections, describing each item. In total, the complete dataset contains a total of 5,591,325 items with 22,792,429 tokens. Using the complete dataset, we also generate three versions of varying sizes: small (1,471,824 items), medium (2,943,648 items), and large (5,591,325 items).[1]

4.3. Pre-trained Models

We choose two commonly used language models for pre-training, one with a single-language composition tailored to Brazilian Portuguese, and the other adopts a multilingual composition capable of processing multiple languages, including Portuguese.

BERTimbau [Souza et al. 2020]. It is a state-of-the-art pre-trained language model specifically designed for Brazilian Portuguese. Trained on a large corpus of text from various sources, including news articles, social media posts, and web pages, BERTimbau captures the linguistic nuances and complexities of the Portuguese language. It has demonstrated impressive performance across a wide range of natural language processing tasks, making it a suitable choice for fine-tuning in our study.

LaBSE (Language-agnostic BERT Sentence Embedding) [Feng et al. 2022]. It is a multilingual language model trained to generate language-agnostic sentence embeddings.

[1] https://dadosabertos.tse.jus.br/group/prestacao-de-contas-eleitorais?res_format=CSV

Table 1. Overview of the baseline and domain-adapted models.

Pre-trained Model	Corpus Domain		Size
BERTimbau [Souza et al. 2020]	brWac (web texts)	-	3.5M
LaBSE [Feng et al. 2022]	Wikipedia (web texts)	-	22B
BERTimbau LaBSE	Official Gazette Segments (public bidding process)	small	314k
		medium	628k
		large	918k
BERTimbau LaBSE	VICTOR (legal texts)	small	440k
		medium	880k
		large	1.8M
BERTimbau LaBSE	Electoral Expense Items (public expenditure items)	small	1.5M
		medium	2.9M
		large	5.6M

By encoding sentences into a shared semantic space regardless of language, LaBSE facilitates cross-lingual transfer learning and improves performance on downstream tasks. Trained on a diverse, multilingual corpus, LaBSE can process over 100 languages, including Portuguese, making it a versatile option for fine-tuning in multilingual settings. Here, due to resource constraints, we used a smaller version of LaBSE distilled from the original model, focusing on only 15 languages (including Portuguese).[2]

4.4. Domain-adapted Models

To create the domain-adapted models, we conducted continuous pre-training on both BERTimbau and LaBSE models using the Masked Language Modeling (MLM) task. Such a task involves masking a percentage of words in a sequence (15%) and prompting the model to predict the masked words. The pre-training sessions are limited to 10 epochs, with checkpoints saved at intervals corresponding to each epoch. We evaluated each saved checkpoint based on its performance on the two downstream tasks outlined in Section 4.1, rather than directly assessing the MLM task. Given that each pre-trained domain-specific dataset has three different scale versions, a total of nine distinct datasets are considered. Therefore, combining the two pre-trained models with the nine pre-training datasets resulted in 18 domain-adapted models. Table 1 summarizes the key characteristics of the baseline and the domain-adapted language models.

Hyperparameters. During pre-training, the hyperparameters are primarily set to the default values specified in the *Trainer* class of the Hugging Face Transformers library.[3] Specifically, we use a learning rate of 5e-5 (0.00005) along with the AdamW optimizer. The optimization parameters are configured with β_1 set to 0.9 and β_2 set to 0.999, and an L2 weight decay of 0.0 is applied. We do not employ any warm-up steps or linear learning rate decay, while a dropout probability of 0.1 is implemented to help mitigate overfitting.

5. Experimental Evaluation

Our experimental evaluation focuses on assessing the performance of domain-adapted language models in two distinct downstream tasks: document classification and item

[2] https://huggingface.co/setu4993/smaller-LaBSE

[3] https://huggingface.co/docs/transformers/main_classes/trainer

Table 2. Datasets used for the downstream tasks.

Downstream task	Dataset	Domain	Size
Document Classification	SVic [Luz de Araujo et al. 2020]	Legal texts	339,478
	LiPSet [Silva et al. 2022]	Public bidding process	9,761
Item Classification	ProdServ	Public expenditure items	3,76M
	NaPEx	Public expenditure items	583,174

classification. For benchmarking, we consider both language models used during domain-adaptive pre-training as baselines, BERTimbau and LaBSE. In this section, we describe the datasets used for each downstream task (Section 5.1), the evaluation setup (Section 5.2), and the results for each research question (Section 5.3).

5.1. Datasets

For each downstream task, we select a distinct dataset to ensure diverse and representative samples of text data pertinent to the governmental domain. This approach enables us to evaluate the performance of domain-adapted language models across different text classification tasks. Table 2 summarizes each dataset's key characteristics, which are further described as follows.

SVic [Luz de Araujo et al. 2020]. It is the small version of the VICTOR dataset, containing 339,478 labeled legal texts. Note that this subset includes different data samples from those used during pre-training. The instances are labeled as *Acórdão* (lower court decisions under review), *Recurso Extraordinário* (appeal petitions), *Agravo de Recurso Extraordinário* (motions against the appeal petition), *Sentença* (judgments), *Despacho* (court orders), and Others (documents not included in the previous classes).

LiPSet [Silva et al. 2022]. It consists of 9,761 labeled public bidding documents from the Brazilian state of Minas Gerais (i.e., the second most populous in the country). Such documents are characterized by their technical language and are classified into four meta-classes subdivided into 12 classes: (i) *Minutes*: price registration, minutes of waiver, face-to-face auction, others; (ii) *Public Notice*: public notice; (iii) *Homologation*: homologation; and (iv) *Others*: contract, notice, amendment, ratification, erratum, others.

ProdServ. It contains items purchased by the public sector of Minas Gerais, Brazil. In short, such items refer to products or services that are specified in public bidding notices, contracts, or invoices. Each item is detailed in terms of its features, quantity, quality, and other relevant specifications. In total, the dataset comprises 3,755,948 labeled items categorized as (i) *products* or (ii) *services*.[4]

Nature of Public Expenditure (NaPEx). It contains public expenditure items from the state of Minas Gerais, Brazil. Each item presents a text description and is classified into five nature classes: (i) *Obras* (construction), (ii) *Serviços* (services), (iii) *Material de Consumo* (consumables), (iv) *Material Permanente* (permanent materials), and (v) *Locação* (rent). In total, the dataset comprises 583,174 labeled items.

[4] https://portalsicom1.tce.mg.gov.br/

5.2. Evaluation Setup

The fine-tuning setup incorporates several key components to optimize model performance for classification tasks. The final layers of each model include a fully connected layer that maps the hidden states to the number of output classes, followed by a softmax activation function in the output layer. Additionally, to mitigate overfitting during training, dropout is applied, randomly setting a fraction of the input units to zero at each update. We also use the Adam optimizer to minimize the loss function, leveraging its ability to compute adaptive learning rates for each parameter, which aids in faster convergence. The loss function employed is the cross-entropy loss, a common choice for classification tasks, which measures the models' performance by comparing the predicted probabilities with the actual class labels.

In our experimental setup, we partition each dataset into distinct subsets for training, validation, and testing purposes, ensuring that the distribution of label values is preserved across the partitions. Given that all datasets exhibit class imbalance, we perform the split in a stratified manner by class. This stratified partitioning scheme allocates 70% of the data to the training set, 20% to the validation set, and the remaining 10% to the test set. Regarding imbalanced data, which is prevalent across all considered datasets, we deliberately avoid employing oversampling or undersampling techniques to balance the dataset. Instead, we assess the models' performance on the original imbalanced data, thereby ensuring that our assessments accurately reflect real-world scenarios.

All evaluated domain-adapted models are trained for a fixed number of five epochs, striking a balance between computational efficiency and sufficient training time to capture meaningful patterns. Extensive hyperparameter tuning is not conducted, as the primary focus is on comparing and evaluating the efficacy of domain adaptation in governmental applications rather than achieving state-of-the-art performance on downstream tasks. Moreover, the models are trained until they converge regarding the validation set loss, ensuring they reach a stable performance level. After convergence, or at the end of the last epoch, we evaluate the models' performance using the F1-Macro metric, due to its effectiveness in handling class imbalances, which are prevalent in our datasets.

5.3. Results

In this section, we present and discuss the results of our experimental evaluation. We analyze the performance of domain-adapted language models for each research question in document and item classification tasks. We explore the impact of different factors, including the choice of target domain dataset, the language composition of the pre-trained model, and the size of the dataset used in pre-training. Table 3 shows the F1-Macro of the evaluated models in each of the four downstream tasks.

Domain Relevance (RQ1). By comparing the F1-Macro scores across different datasets, we assess how well each model adapts to governmental documents' specific linguistic nuances and characteristics. Regarding the document classification tasks, models pre-trained on segments related to bidding and legal texts showed superior performance compared to those pre-trained on electoral expenditure item descriptions. Both baseline general-domain models (BERTimbau and LaBSE) outperformed models pre-trained on expenditure item descriptions, indicating a lower degree of adaptation to such governmental applications for these domain-adapted models.

Table 3. Performance comparison of domain-adapted language models across different downstream tasks. Best task performance is boldfaced.

Domain	Model	Size	LiPSet	SVic	ProdServ	NaPEx
			Document	Document	Item	Item
Web	BERTimbau	-	0.9574	0.7661	0.9033	0.8457
	LaBSE	-	0.9317	0.7837	**0.9039**	0.8356
Bidding	BERTimbau	small	**0.9583**	0.7873	0.9033	0.8483
		medium	0.9453	0.7908	0.9029	0.8434
		large	0.9462	**0.7928**	0.9012	0.8471
	LaBSE	small	0.9557	0.7718	0.9011	0.8358
		medium	0.9443	0.7702	0.9015	**0.8495**
		large	0.9513	0.7647	0.9002	0.8367
Legal	BERTimbau	small	0.9412	0.6408	0.9027	0.8372
		medium	0.9406	0.6571	0.9025	0.8451
		large	0.9523	0.7869	0.9012	0.8361
	LaBSE	small	0.9410	0.6643	0.9017	0.8415
		medium	0.9421	0.6707	0.9019	0.8396
		large	0.9396	0.7876	0.9024	0.8325
Expend. items	BERTimbau	small	0.8272	0.5861	0.9012	0.8400
		medium	0.8309	0.7642	0.9018	0.8415
		large	0.7900	0.6341	0.8996	0.8354
	LaBSE	small	0.7117	0.5803	0.9007	0.8325
		medium	0.7727	0.7321	0.9023	0.8414
		large	0.5923	0.5440	0.9012	0.8338

As expected, the superior performance of the models pre-trained on the official gazette segments and the MVic dataset suggests a greater linguistic alignment with the context of public bidding and legal documents found in the LiPSet and SVic datasets, thereby facilitating better adaptation. However, for the SVic dataset, models pre-trained on legal texts—the same target domain as the downstream task—showed lower F1-Macro values than the baselines. This underscores the importance of not only the relevance of the target domain dataset but also the quality and representativeness of the pre-training data in ensuring effective domain adaptation [Zhu et al. 2021].

Regarding the baselines, although the *BERTimbau+Bidding* (small) and (large) models showed superior performance for LiPSet and SVic, respectively, the baselines also demonstrated competitive results. Figure 1 illustrates the results grouped by downstream task and pre-trained model, regardless of the size. Compared to the averaged results of the domain-adapted models, BERTimbau and LaBSE maintain competitive performance, suggesting that domain-adapted models offer improvements in specific contexts but may not consistently outperform general-domain models across all tasks and datasets [Zhu et al. 2021]. This highlights that factors other than the target domain may play a crucial role in determining the effectiveness of domain adaptation.

On the other hand, for the item classification tasks, all models showed consistent performance across the datasets, with slight variations in F1-Macro scores (Figure 1). This is likely due to the more generic nature of these tasks, as item descriptions usually

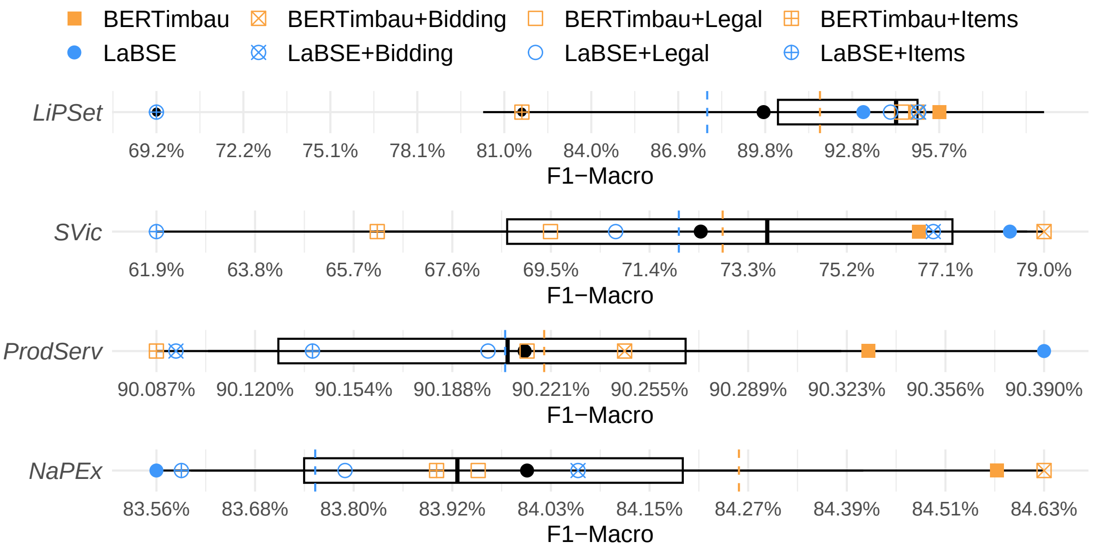

Figure 1. Results by downstream task and pre-trained model. Black points represent mean values for each dataset. Dashed lines represent mean values for the models grouped by the pre-trained base model.

contain a wide range of vocabulary and language patterns, with terms drawn from various domains. Therefore, the performance of the domain-adapted models may not be heavily influenced by the specific domain from which the models were pre-trained. Instead, the performance variations among models in item classification tasks are more likely influenced by factors such as the language composition of the pre-trained models.

Language Composition (RQ2). Figure 1 shows dashed lines representing mean values for the models grouped by the pre-trained model, enabling comparison between models pre-trained on monolingual (BERTimbau) and multilingual (LaBSE) versions. Although the general-domain LaBSE model demonstrated competitive performance, our findings indicate that models based on BERTimbau consistently exhibit slightly higher mean F1-Macro scores than those based on LaBSE across document and item classification tasks. This suggests that pre-training a multilingual model with single-language data may lead to suboptimal performance compared to models explicitly trained on the target language.

Regarding only the baseline models, the multilingual LaBSE has more wins than BERTimbau across most downstream tasks. This could be attributed to LaBSE being pre-trained in multiple languages, which might make it more adept at capturing the linguistic nuances present in diverse datasets. Especially for item classification tasks, where the language tends to be more generic, the multilingual capabilities of LaBSE might offer an advantage over BERTimbau. However, it is important to note that the specific characteristics of each downstream task and dataset could also play a significant role in determining which model performs better.

For instance, in tasks involving highly specific or domain-specific language, such as governmental documents, BERTimbau's training, specifically in Portuguese, may offer an edge over LaBSE's multilingual training. This advantage stems from BERTimbau's tailored understanding of Portuguese language intricacies and nuances, which are crucial for accurately processing and classifying domain-specific textual data. In contrast, while

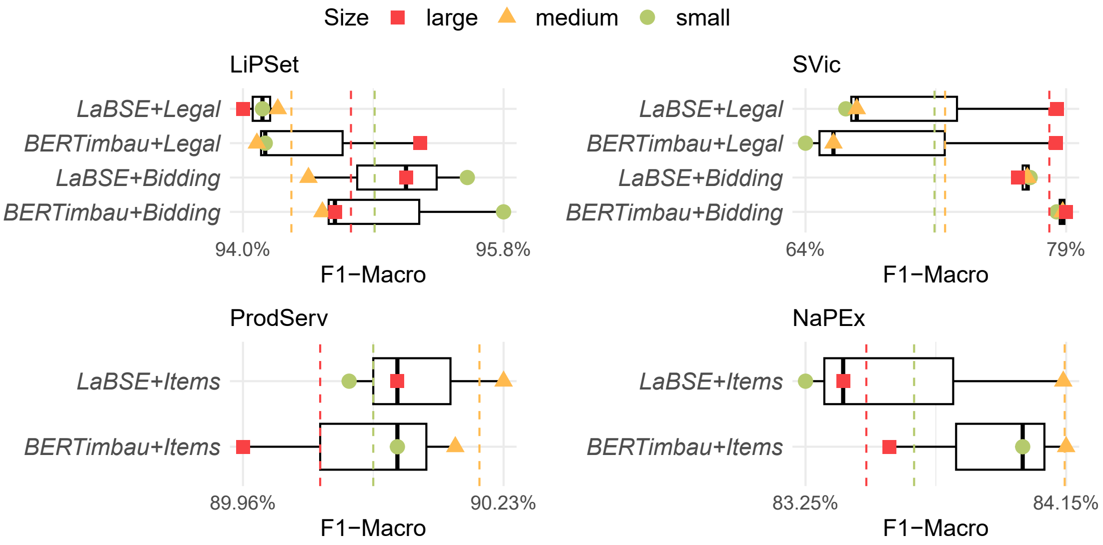

Figure 2. Results by downstream task and domain-adapted model. Dashed lines represent mean values for the models grouped by the size version.

LaBSE's multilingual training provides broader linguistic coverage, it may not capture the subtle contextual nuances unique to Portuguese governmental documents. Therefore, our findings highlight the importance of considering both model capabilities and task domain when selecting a pre-trained language model for downstream applications.

Dataset Size (RQ3). We compare the F1-Macro scores of models pre-trained on different dataset sizes to determine whether increasing the amount of domain-specific data during pre-training leads to improved model efficacy in governmental tasks. Figure 2 shows the results grouped by downstream task and domain-adapted model, with dashed lines representing mean values for models categorized by size version. For each downstream task, we plot only the domain-adapted models pre-trained on domains most related to the task at hand to provide a focused comparison.

For both classification tasks, all domain-adapted models demonstrated consistent performance regardless of the pre-trained dataset size. This suggests that the varying scales of domain-specific data used during pre-training might not have been substantial enough to significantly impact the models' effectiveness in classifying government documents and items [Zhu et al. 2021]. However, it is important to note that other factors, such as the quality and representativeness of the data, could still play a crucial role in model performance. Further exploration and experimentation could provide deeper insights into the optimal dataset size for effective domain adaptation in such tasks.

6. Conclusion

In this paper, we investigated the effectiveness of domain-adaptive pre-training (DAPT) in improving the performance of language models for Portuguese governmental applications. We explored how different factors, including target domain dataset selection, dataset size, and pre-trained model language composition, impact model performance. Our study involves the creation and evaluation of 18 domain-adapted models across four governmental text classification tasks to uncover insights into the optimal strategies for domain adaptation in such a context.

Overall, our experimental evaluation yielded several key findings regarding the effectiveness of domain-adapted language models in governmental text classification tasks. Firstly, regarding the influence of the choice of target domain dataset (RQ1), our results indicate that the linguistic nuances captured in datasets more closely aligned with the target domain contribute to better adaptation. The more related the target domain dataset is to the downstream task, the more effective the domain adaptation appears to be, as evidenced by the superior performance of models pre-trained on segments related to bidding and legal texts compared to those pre-trained on electoral expenditure item descriptions.

Secondly, our investigation into the language composition of pre-trained models (RQ2) revealed that while multilingual models like LaBSE demonstrated competitive performance, pre-training models specifically on the target language may offer particular advantages in domain-specific applications. Finally, regarding the impact of dataset size on model performance (RQ3), we observed that increasing the amount of domain-specific data during pre-training did not consistently lead to improved efficacy in governmental tasks. Although larger datasets are typically presumed to enhance model performance, our findings suggest that other factors, such as dataset quality and representativeness, may be equally or more important.

While our study provides valuable insights, it is important to acknowledge its limitations. For instance, we recognize that our exploration of factors influencing model performance may not be exhaustive, and there may be additional variables or interactions that require further investigation. Additionally, the single-run training of each model limits our ability to calculate statistical significance for observed performance differences. Furthermore, the presence of class imbalance in our datasets could affect the models' performance evaluations, as certain classes may be underrepresented. Such limitations present opportunities for future research to conduct more comprehensive experiments with larger datasets and explore a wider range of tasks and languages, as well as to investigate techniques for addressing class imbalance in governmental text classification.

Acknowledgments. This work was funded by the Prosecution Service of State of Minas Gerais (in Portuguese, *Ministério Público do Estado de Minas Gerais*, or simply MPMG) through its Analytical Capabilities Project and by CNPq, CAPES, FAPEMIG and the partnership project between AWS and CNPq.

References

Brandão, M. A. et al. (2023). Impacto do pré-processamento e representação textual na classificação de documentos de licitações. In *SBBD*, pages 102–114. SBC.

Brandão, M. A. et al. (2024). PLUS: A Semi-automated Pipeline for Fraud Detection in Public Bids. *Digital Government: Research and Practice*, 5(1):1–16.

Constantino, K. et al. (2022). Segmentação e Classificação Semântica de Trechos de Diários Oficiais Usando Aprendizado Ativo. In *SBBD*, pages 304–316. SBC.

Feijó, D. V. and Moreira, V. P. (2020). Mono vs Multilingual Transformer-based Models: a Comparison across Several Language Tasks. *CoRR*, abs/2007.09757.

Feng, F. et al. (2022). Language-agnostic BERT Sentence Embedding. In *ACL*, pages 878–891. Association for Computational Linguistics.

Gururangan, S. et al. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *ACL*, pages 8342–8360. Association for Computational Linguistics.

Hott, H. R. et al. (2023). Evaluating contextualized embeddings for topic modeling in public bidding domain. In *BRACIS*, volume 14197 of *LNCS*, pages 410–426. Springer.

Luz de Araujo, P. H., de Campos, T. E., Braz, F. A., and da Silva, N. C. (2020). VICTOR: a Dataset for Brazilian Legal Documents Classification. In *LREC*, pages 1449–1458. ELRA.

Luz de Araujo, P. H. et al. (2018). LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text. In *PROPOR*, volume 11122 of *LNCS*, pages 313–323. Springer.

Oliveira, G. P. et al. (2022). Detecting Inconsistencies in Public Bids: An Automated and Data-based Approach. In *WebMedia*, pages 182–190. ACM.

Rodrigues, R. B. M. et al. (2022). PetroBERT: A Domain Adaptation Language Model for Oil and Gas Applications in Portuguese. In *PROPOR*, volume 13208 of *LNCS*, pages 101–109. Springer.

Schneider, E. T. R. et al. (2020). BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. In *ClinicalNLP@EMNLP*, pages 65–72. Association for Computational Linguistics.

Silva, M. O. et al. (2022). LiPSet: Um Conjunto de Dados com Documentos Rotulados de Licitações Públicas. In *DSW*, pages 13–24. SBC.

Silva, M. O. et al. (2023). Análise de Sobrepreço em Itens de Licitações Públicas. In *WCGE*, pages 118–129. SBC.

Silva, M. O. and Moro, M. M. (2024). Evaluating Pre-training Strategies for Literary Named Entity Recognition in Portuguese. In *PROPOR*, pages 384–393. Association for Computational Lingustics.

Silva, N. F. F. et al. (2021). Evaluating Topic Models in Portuguese Political Comments About Bills from Brazil's Chamber of Deputies. In *BRACIS*, volume 13074 of *LNCS*, pages 104–120. Springer.

Silveira, R. et al. (2021). Topic Modelling of Legal Documents via LEGAL-BERT. In *Procs. of the 1st International Workshop RELATED - Relations in the Legal Domain*.

Silveira, R. et al. (2023). LegalBert-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain. In *BRACIS*, volume 14197 of *LNCS*, pages 268–282. Springer.

Singhal, P., Walambe, R., Ramanna, S., and Kotecha, K. (2023). Domain Adaptation: Challenges, Methods, Datasets, and Applications. *IEEE Access*, 11:6973–7020.

Souza, F., Nogueira, R. F., and de Alencar Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *BRACIS*, volume 12319 of *LNCS*, pages 403–417. Springer.

Zhu, Q. et al. (2021). When does Further Pre-training MLM Help? An Empirical Study on Task-Oriented Dialog Pre-training. In *Workshop on Insights from Negative Results in NLP*, pages 54–61. Association for Computational Linguistics.