GovBERT-BR: A BERT-based Language Model for Brazilian Portuguese Governmental Data

Mariana O. Silva[1][0000−0003−0110−9924], Gabriel P. Oliveira[1][0000−0002−7210−6408], Lucas G. L. Costa[1][0009−0002−8898−4237], and Gisele L. Pappa[1][0000−0002−0349−4494]

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. Given the growing interest in natural language processing (NLP) for governmental applications, particularly in Brazil, where vast amounts of governmental data are processed daily, the need for specialized NLP models tailored to the nuances of Brazilian Portuguese and the legal and administrative domains has become increasingly apparent. However, existing models may struggle to accurately interpret the complexities of governmental texts, often leading to suboptimal performance in document classification and analysis tasks. To address these challenges, we introduce GovBERT-BR, a pre-trained language model tailored to the Brazilian governmental context, covering legal and administrative domains. Leveraging insights from diverse governmental texts, GovBERT-BR addresses the challenges of accurately interpreting Brazilian Portuguese and the unique legal and bureaucratic terminology prevalent in governmental documents. We present the pre-training process and experimental evaluation of GovBERT-BR, comparing its performance against baseline models across various text classification tasks relevant to the Brazilian public sector. Our findings demonstrate that GovBERT-BR outperforms existing models in document and short-text classification tasks, showcasing its efficacy in accurately analyzing governmental text data. Furthermore, our analysis reveals insights into the convergence behavior of GovBERT-BR during fine-tuning, highlighting its rapid adaptation to downstream tasks.

Keywords: Natural Language Processing · Brazilian Governmental Data · Legal Texts · BERT · Language Models

1 Introduction

In recent years, the application of natural language processing (NLP) tools to government data has surged significantly. This trend is largely driven by the promise of NLP to enhance public administration's efficiency, transparency, and accountability. In Brazil, where a vast volume of governmental data is generated and processed daily [4], NLP tools can streamline various administrative

processes by automating the analysis and interpretation of textual data, thereby facilitating improved decision-making [10,20,8].

However, the effective utilization of this data is often constrained by the need for specialized NLP models capable of accurately interpreting the specificities of Brazilian Portuguese and the unique legal and bureaucratic terminology. Existing NLP models are predominantly generic [19], trained on datasets that do not adequately reflect the particularities of Brazilian governmental documents. This mismatch usually leads to suboptimal performance when these models are applied to the Brazilian public sector [20].

In this context, some efforts have been made to develop domain-specific models focused on tackling legal [20,18,8] and administrative [10] tasks. Although these endeavors represent important steps forward, they are often limited in scope, addressing only specific sub-contexts of governmental data. The interdisciplinary nature of governmental documents requires models capable of seamlessly integrating insights from multiple domains. Legal documents, for example, frequently intersect with administrative procedures, and understanding these interrelations is crucial for effective decision-making within public administration.

Therefore, we introduce GovBERT-BR, a pre-trained language model specifically tailored to the Brazilian governmental context. It integrates both administrative and legal domains, providing a multifaceted perspective to better align with the diverse nature of governmental text data. In this paper, we present the pre-training process and the experimental setup of GovBERT-BR. Our evaluation involves a series of text classification tasks relevant to the Brazilian governmental context, comparing the performance of GovBERT-BR against several baseline models. Our main contributions are summarized as follows.

- We introduce a novel pre-trained language model designed explicitly for the Brazilian governmental domain, incorporating insights from both administrative and legal domains to enhance its adaptability and effectiveness.
- We provide a comprehensive evaluation framework, including a diverse set of text classification tasks pertinent to the Brazilian governmental context. This allows for a thorough assessment of GovBERT-BR's performance across different types of governmental text data.
- Our experimental results demonstrate GovBERT-BR's superior or competitive performance compared to baseline models across various datasets. This underscores its efficacy in accurately classifying governmental text, highlighting its potential for improving text understanding and decision-making within the Brazilian public sector.

2 Related Work

Using language models in government applications has changed how governments interact with citizens and manage information. Indeed, such models offer advanced capabilities in NLP, enabling automation in understanding public inquiries, analyzing policies, and other applications using Open Government Data

(OGD) [7,14]. They can be applied in several areas, including consulting services [9], topic modeling [10], and sentiment analysis related to government policies [23]. In addition, language is crucial in these models, as the model's language must align with the languages used in the target domain to ensure more efficient adaptation. Therefore, recent works have proposed distinct models to tackle government-related tasks in several languages [1,3,11,15,22].

In the Brazilian context, most models that address government tasks are built from BERTimbau, a general-purpose model pre-trained on a large corpus of Portuguese texts scraped from the Web [19]. Such government-related models can be divided into two distinct categories according to their purpose: legal domain and administrative domain models. The first group includes models focusing mainly on juridical and legislative tasks, including texts from the Brazilian legal system into their pre-training phase to handle legal language and concepts effectively.

For example, LegalBERTPT-br [18] is a model for mining topics within comments on Brazilian draft laws. LegalBert-pt [20], JurisBERT [21] and CLSJUR.BR [12] use texts exclusively from the Brazilian judiciary branch for tasks such as named entity recognition (NER), semantic textual similarity (STS), and document summarization. Finally, RoBERTaLexPT [8] uses a large corpus from the judiciary and legislative bodies over NER and classification tasks. The main difference between such models is that RoBERTaLexPT's corpus contains texts in both European and Brazilian varieties of the Portuguese language, whereas the others consider only texts in Brazilian Portuguese.

On the other hand, administrative domain models are designed to support various governmental functions outside the legal sphere, such as managing bidding processes and processing information from the official gazettes. The development of solutions for this specific domain is relatively recent, with few models dedicated exclusively to this context. One notable example is LiBERT-SE [10], a model developed using a corpus of Brazilian official gazette segments and evaluated in topic detection tasks within bidding documents.

In light of the aforementioned related work, our goal is to propose a model that addresses the legal and administrative domains in Brazilian Portuguese. GovBERT-BR bridges the gap between these two domains, providing a comprehensive solution for handling diverse governmental text data. By integrating insights from both legal and administrative contexts, GovBERT-BR aims to enhance the efficiency and effectiveness of NLP applications within the Brazilian public sector, ultimately contributing to improved decision-making.

3 GovBERT-BR

Governmental data is a rich and diverse source of information that contains many documents, including public bidding documents, legal texts, legislative bills, administrative reports, and more. However, the complexity and specialized nature of this data pose significant challenges for generic NLP models, which often fail to capture the nuances and specific terminology used in governmental texts. Moreover, existing domain-specific models tend to focus on narrow sub-

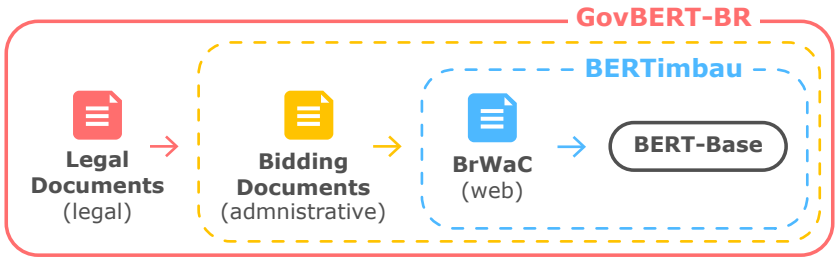

Fig. 1. Overview of GovBERT-BR pre-training methodology.

contexts, failing to address the interdisciplinary nature inherent in governmental documents comprehensively.

To overcome these challenges, we introduce GovBERT-BR, a pre-trained language model specifically tailored to the Brazilian governmental context.[1] It considers both legal and administrative domains, providing a multifaceted perspective that can better align with the diverse nature of governmental text data. In this section, we describe the pre-trained model used as the foundation for GovBERT-BR (Section 3.1) and detail the methodologies employed in its development (Section 3.2).

3.1 Pre-trained Model

GovBERT-BR is built upon the architecture of BERTimbau [19], the first large-scale Portuguese language model. BERTimbau, based on the BERT (Bidirectional Encoder Representations from Transformers) [6] architecture, has been fine-tuned on various tasks to enhance its performance in understanding Portuguese text. GovBERT-BR inherits this architecture and extends it to focus specifically on the nuances of Brazilian governmental and legal documents.

3.2 Pre-training

The GovBERT-BR pre-training process involves further training BERTimbau using two distinct corpora, each tailored to capture the intricacies of governmental and legal texts in Brazilian Portuguese. We believe such a multi-domain approach allows GovBERT-BR to provide a comprehensive perspective that captures the interplay between legal and administrative language, offering a powerful resource for interpreting Brazilian governmental documents.

Figure 1 shows the overview of the GovBERT-BR pre-training methodology, which involves a two-phase approach. Initially, the model is pre-trained using data from the administrative domain to equip the model with an understanding of bureaucratic terminology and administrative procedures commonly found in governmental texts. Next, the model undergoes further pre-training using a

[1] Available at https://huggingface.co/dccmpmgfinalisticas/GovBERT-BR.

dataset comprised of legal documents related to extraordinary appeals received by the Brazilian Supreme Court (STF).

Administrative Data. The initial pre-training corpus includes public bidding documents and segments from the official gazette, which contain summarized texts published to provide information regarding the bidding process [5]. It comprises 1,578,107 sentences and encompasses various document types and administrative data, including announcements, notices, and bidding terms.

Legal Data. The second pre-training corpus includes legal documents related to extraordinary appeals received by the STF. It contains a subset of 1,760,862 pages sourced from the VICTOR dataset [13].

Task. The pre-training task employed for GovBERT-BR is the Masked Language Model (MLM) task. In this task, a percentage of words in a sequence (15%) are masked, and the model is trained to predict these masked tokens based on the surrounding context. This method enables the model to learn contextual relationships and develop a deep understanding of the language patterns specific to the domain.

Setup. Each pre-training session is limited to 10 epochs, with checkpoints saved at the end of each epoch. We use a batch size of 64 sequences, each containing a maximum of 300 tokens. Both pre-training steps involved over 245,000 steps, ensuring thorough exposure to the diverse dataset.

4 Experimental Design

To evaluate the effectiveness of the pre-trained GovBERT-BR language model, we conduct a series of experiments across various text classification tasks relevant to the Brazilian governmental context. These experiments aim to assess GovBERT-BR's performance compared to baseline models and existing NLP approaches, demonstrating its ability to analyze and interpret governmental documents accurately. In this section, we describe the baselines considered (Section 4.1), the downstream tasks (Section 4.2) and datasets (Section 4.3), and the evaluation metrics used (Section 4.4).

4.1 Baselines

Our main objective is to assess whether integrating legal and administrative pre-training data into GovBERT-BR leads to better performance on governmental text classification tasks. Therefore, we compare our model against several baseline models, including **general-purpose** and **domain-specific** models (Table 1). Regarding the general domain, we consider the BERTimbau-Base model, a widely used general-purpose language model pre-trained on a large-scale Portuguese corpus and used as the basis for the GovBERT-BR model.

In addition to BERTimbau, we include domain-specific models trained on legal and administrative data separately. For the **legal domain**, we consider the

Table 1. Overview of the baseline models.

Model	Domain	Corpus	Size
[19] BERTimbau	General	brWac	3.5M
[10] LiBERT-SE	Admin	Official Gazette Segments	300K
[20] LegalBERT-PT	Legal	Legal Documents	1.5M
GovBERT-BR	Admin Legal	Public Bidding Documents Official Gazette Segments VICTOR dataset	3.4M

LegalBERT-PT model [20], which is also built upon BERTimbau and pre-trained on a vast corpus of legal documents, enabling it to capture the nuances of legal language and terminology. As GovBERT-BR builds upon BERTimbau, comparing both models allows us to evaluate the impact of additional pre-training on legal and administrative texts.

For the **administrative domain**, we consider LiBERT-SE [10], a pre-trained language model tailored for the public bidding context. LiBERT-SE is built upon BERTimbau using the MLM task and a dataset of 300,000 official gazette segments sourced from various municipalities in Minas Gerais, as previously extracted in [5]. As it focuses on administrative language and procedures, such a model is a relevant benchmark for evaluating GovBERT-BR's performance in administrative tasks.

4.2 Downstream Tasks

As downstream tasks, we focus on two distinct text classification tasks within the governmental domain: document classification and short text classification.

Document Classification. This task involves categorizing documents, such as legal texts, legislative bills, administrative reports, and public bidding documents, into predefined classes. Document classification provides a high-level understanding of the content and purpose of each document, facilitating efficient document retrieval and organization within governmental systems.

Short Text Classification. Unlike document classification, this task focuses on classifying shorter text segments, such as item descriptions, summaries, or motions. Short text classification is particularly useful for extracting relevant information from large volumes of textual data, enabling quick decision-making and analysis in governmental processes.

4.3 Datasets

For each domain, three datasets are selected to evaluate the performance of GovBERT-BR and the baseline models. Table 2 summarizes each dataset's key characteristics, further described as follows.

Table 2. Datasets used for the downstream tasks.

Ref.	Dataset	Year	Domain	Downstream Task	Size
[17]	LiPSet	2022	Administrative	Document Classification	9,761
-	NaPEx	2024	Administrative	Short text Classification	583,174
-	ProdServ	2024	Administrative	Short text Classification	583,174
[13]	SVic	2020	Legal	Document Classification	339,478
[16]	Motions	2021	Legal	Short text Classification	6,449
[2]	RRIoP	2021	Legal	Short text Classification	10,784

LiPSet [17]. Contains 9,761 labeled public bidding documents from Minas Gerais, Brazil. Such documents are classified into 12 classes: *Ata Registro Preços* (price registration), *Ata Dispensa* (minutes of waiver), *Ata Pregão Presencial* (face-to-face auction), *Outras Atas* (other minutes), *Edital* (public notice), *Contrato* (contract), *Aditamento* (amendment), *Aviso* (notice), *Ratificação* (ratification), *Errata* (erratum), *Homologação* (homologation), and *Outros* (others).

NaPEx. Contains 583,174 labeled public expenditure items from Minas Gerais, Brazil.[2] Each item presents a short text description and is classified into five expenditure nature classes: *Obras* (construction), *Serviços* (services), *Material de Consumo* (consumables), *Material Permanente* (permanent materials), and *Locação* (rent).

ProdServ. Represents the same public expenditure items as NaPEx, but in this dataset, each item description is further classified as either *products* or *services*.

SVic [13]. It is a subset of the VICTOR dataset, containing 94,267 legal documents with 339,478 labeled pages. 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.

Motions [16]. Contains 6,449 legal proceedings, each with an individual and variable number of motions, labeled by lawyers. Legal proceedings refer to formal legal actions or cases brought before a court of law. Motions, in turn, are formal requests made to a judge during legal proceedings, typically seeking a decision or ruling on a specific aspect of the case. Instances in this dataset are labeled as *Arquivado* (archived), *Ativo* (active), or *Suspenso* (suspended), indicating the status of the legal proceedings.

RRIoP [2]. Contains rhetorical annotations within the legal domain, focusing on sentences extracted from judicial sentences from the Court of Justice of Mato Grosso do Sul (TJMS), Brazil. It comprises 10,784 petitions from 70 civil lawsuits filed in TJMS court and was manually labeled with rhetorical roles specifically tailored for petitions. Instances in this dataset are labeled as *Identificação das partes* (identification), *Fatos* (facts), *Argumentos* (arguments), *Fundamentação*

[2] https://dadosabertos.tce.mg.gov.br/

Legal (legal basis), *Jurisprudência* (precedents), *Pedidos* (requests), *Valor da causa* (remedy), and Others.

4.4 Evaluation Setup

In our experimental setup, each dataset is partitioned into distinct subsets for training, validation, and testing, maintaining the distribution of label values across 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 training, 20% to validation, and the remaining 10% to testing. However, it is important to note that we do not use cross-validation in this study, which may introduce bias into the results as the model's performance is evaluated on a single split rather than multiple folds. This choice limits the ability to generalize the findings across different data partitions.

Moreover, the class imbalance present across all datasets poses a challenge. While we refrain from using oversampling or undersampling techniques to reflect real-world scenarios, balancing techniques could be applied during training without compromising the integrity of the imbalanced test scenario. Such techniques might help mitigate the impact of class imbalance on model performance, leading to more reliable and interpretable outcomes. We leave the exploration of these balancing strategies for future work, as it could further enhance the robustness of the models in handling skewed data distributions.

All evaluated language models are trained for a fixed number of five epochs without extensive hyperparameter search. Models are trained until convergence on the validation set loss is achieved, ensuring stable performance. Following convergence, or at the last epoch, model performance is evaluated using the F1-Macro metric. To ensure the robustness of our evaluations, we conduct five experiments for each model, varying the random seeds. Next, the results are averaged to produce a more reliable estimation of model performance.

5 Experimental Results

In this section, we present the experimental results obtained from evaluating GovBERT-BR across various text classification tasks relevant to the Brazilian governmental domain. We analyze GovBERT-BR's performance in terms of overall effectiveness (Section 5.1), the impact of multi-domain pre-training (Section 5.2), and its convergence across different governmental domains (Section 5.3).

5.1 Overall Performance

We first assess how GovBERT-BR's performance compares to other state-of-the-art NLP models. Tables 3 and 4 present the results of document and short text classification tasks in the administrative and legal domains, respectively. The statistical significance is reported using the Wilcoxon rank-sum test to compare the distributions of F1-Macro scores. In document classification, GovBERT-BR

Table 3. Performance comparison on document and short text classification tasks in the **administrative** domain. F1-Macro scores are reported with standard deviations.

	Document	*Short Text*	
Model	**LiPSet**	**NaPEx**	**ProdServ**
BERTimbau	0.902 ± 0.026	0.839 ± 0.003	0.816 ± 0.003
LiBERT-SE	0.937 ± 0.006	0.836 ± 0.002	0.815 ± 0.003
LegalBERT-PT	0.945 ± 0.009	0.831 ± 0.008	0.814 ± 0.004
GovBERT-BR	0.948 ± 0.008	0.839 ± 0.003	0.814 ± 0.004

Table 4. Performance comparison on document and short text classification tasks in the **legal** domain. F1-Macro scores are reported with standard deviations.

	Document	*Short Text*	
Model	**SVic**	**Motions**	**RRIoP**
BERTimbau	0.779 ± 0.008	0.782 ± 0.004	0.825 ± 0.010
LiBERT-SE	0.774 ± 0.010	0.777 ± 0.005	0.814 ± 0.006
LegalBERT-PT	0.787 ± 0.007	0.781 ± 0.002	0.839 ± 0.005
GovBERT-BR	0.794 ± 0.012	0.784 ± 0.003	0.838 ± 0.005

achieves the highest F1-Macro on the LiPSet dataset, with an average score of 0.948, outperforming significantly the general-purpose BERTimbau ($W = 25$, $p < 0.05$) and the domain-specific LiBERT-SE models ($W = 22$, $p \leq 0.05$). However, regarding LegalBERT-PT, the difference in performance is not statistically significant ($W = 15$, $p > 0.05$).

Similarly, in the SVic dataset, GovBERT-BR achieves an F1-Macro score of 0.794, demonstrating its competitive performance. It significantly outperforms LiBERT-SE ($W = 23$, $p < 0.05$), but the difference in performance compared to BERTimbau ($W = 21$, $p > 0.05$) and LegalBERT-PT ($W = 20$, $p > 0.05$) is not statistically significant. These results demonstrate that pre-training on both legal and administrative data provides GovBERT-BR with a solid understanding of the nuances of Brazilian governmental text, enabling it to perform well across diverse datasets and domains.

Regarding short text classification, GovBERT-BR shows competitive performance across all datasets. It achieves the highest F1-Macro score on the NaPEx dataset, tying with BERTimbau at 0.839, and performs competitively on the Motions and RRIoP datasets, indicating its robustness in handling varied short-text data within governmental contexts. Although LegalBERT-PT slightly outperforms GovBERT-BR on the RRIoP dataset, GovBERT-BR's balanced performance across all datasets highlights its versatility. Finally, our results reveal no statistical difference between all models in the ProdServ dataset.

Overall, GovBERT-BR demonstrates superior or competitive performance compared to baseline models across multiple datasets, underscoring its efficacy in accurately classifying both document-level and short-text data within the

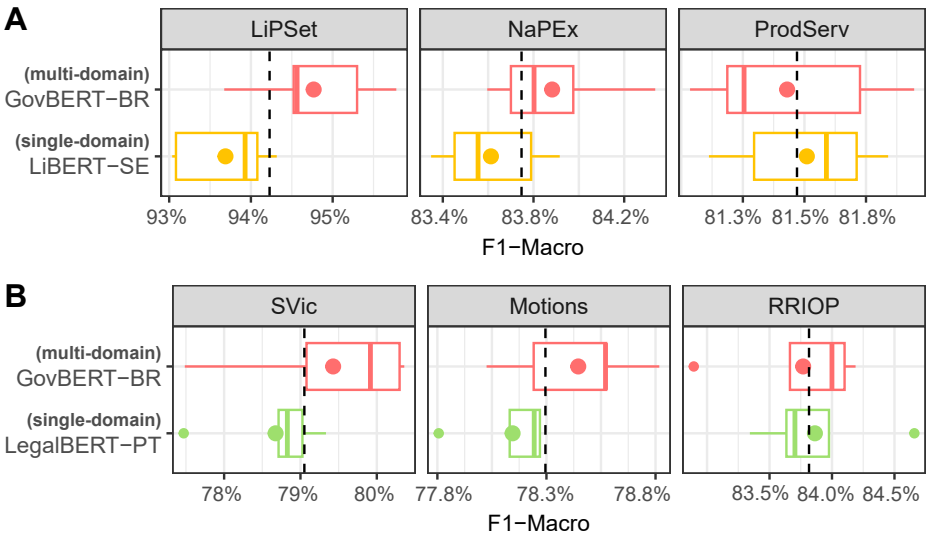

Fig. 2. Distribution of F1-Macro scores on single domains (LiBERT-SE and LegalBERT-PT) and multi-domains (GovBERT-BR) across administrative and legal text classification tasks. (**A**) Administrative and (**B**) Legal domains. The horizontal dashed line represents the mean F1-Macro score for each dataset.

governmental domain. Its robust performance highlights its versatility and effectiveness in handling various types of textual governmental data.

5.2 Multi-Domain Pre-training

To assess the impact of incorporating both administrative and legal domain pre-training data, we compare GovBERT-BR's performance to models pre-trained on single domains. Figure 2 shows the distribution of F1-Macro scores for models pre-trained on a single domain only (LiBERT-SE and LegalBERT-PT) and multi-domains (GovBERT-BR), for text classification tasks in (**A**) administrative and (**B**) legal domains. The horizontal dashed line represents the mean F1-Macro score for each dataset.

In the administrative domain (Figure 2**A**), for most datasets, GovBERT-BR achieves a high mean F1-Macro score compared to the single-domain LiBERT-SE model. The exception is the ProdServ dataset, where the performance difference between GovBERT-BR and LiBERT-SE is minimal, indicating that the benefit of multi-domain pre-training may vary depending on the dataset. This suggests that while multi-domain pre-training generally enhances performance, its impact can be less pronounced in certain administrative contexts.

Similarly, in the legal domain (Figure 2**B**), GovBERT-BR outperforms the single-domain LegalBERT-PT model across most datasets, except for the RRIoP. In this particular dataset, LegalBERT-PT shows a slight edge, possibly due to

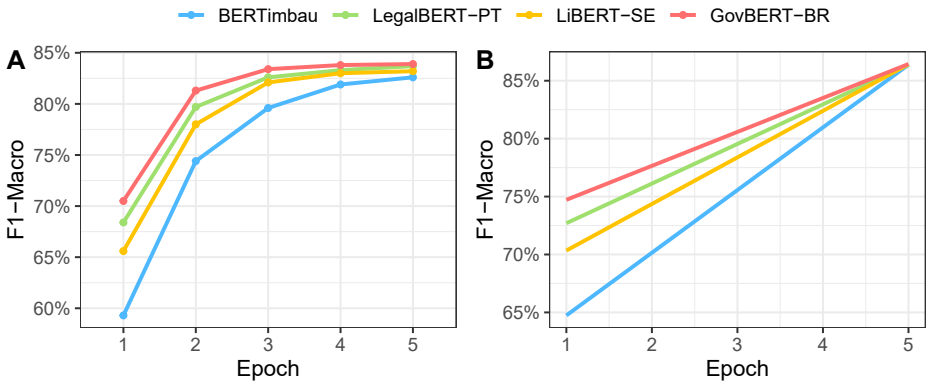

Fig. 3. Convergence analysis. (**A**) Average performance, in terms of F1-Macro, of each model across all datasets, highlighting convergence over training epochs. (**B**) Regression lines of the average performance trajectory during fine-tuning sessions.

its specialized training in legal texts, which are more aligned with the content of RRIoP. However, the difference in performance is not statistically significant, indicating that GovBERT-BR's inclusion of administrative domain data does not detract from its legal domain capabilities.

Overall, the results highlight the advantages of multi-domain pre-training. GovBERT-BR's consistent performance across various datasets, covering both administrative and legal texts, demonstrates the robustness and versatility of multi-domain pre-training. This approach leverages a broader contextual understanding, making the model more adaptable to different types of text classification tasks within governmental contexts. This versatility underscores the potential of GovBERT-BR as a valuable resource for interpreting diverse governmental documents, ultimately supporting more efficient and accurate information processing in governmental operations.

5.3 Convergence Analysis

To further evaluate the models' performance and stability, we analyzed the F1-Macro scores across epochs during fine-tuning on all downstream tasks. Figure 3**A** presents the average performance, in terms of F1-Macro, of each model across the six datasets, highlighting how each model converges over time. Figure 3**B** shows the regression lines for the average performance of each model across all datasets during fine-tuning sessions. The linear regression lines provide insights into the overall performance improvement or degradation trend throughout the fine-tuning process over epochs.

In Figure 3**A**, we observe that BERTimbau initially exhibits lower average performance than the domain-specific models, indicating a slower convergence rate in the early epochs. Such a result is expected as the model undergoes significant adjustments in the initial epochs, transitioning from broad pre-training

to more specialized tasks. As training progresses, BERTimbau's performance steadily increases, eventually reaching comparable levels to the domain-specific models, albeit with a slower rate of improvement. This suggests that BERTimbau may need more epochs to achieve full adaptation.

Overall, all domain-specific models demonstrate rapid convergence compared to the BERTimbau baseline, with GovBERT-BR notably achieving the highest F1-Macro scores early in the training process. LiBERT-SE and LegalBERT-PT also converge quickly, demonstrating the effectiveness of domain-specific pre-training. This allows the models to quickly adjust to the nuances of the target tasks and achieve high-performance levels early in the training process. However, GovBERT-BR consistently maintains a higher F1-Macro score, suggesting that multi-domain pre-training provides a significant advantage across various governmental text classification tasks.

Figure 3**B** corroborates these observations, providing a straightforward visual representation of the trajectory of performance improvement over epochs. The steeper the slope of the line, the faster the model's performance improves during fine-tuning. Conversely, a shallower slope indicates a slower rate of improvement. Although BERTimbau shows the steepest slope among the models, indicating relatively rapid improvement, it starts from a lower performance level compared to the domain-specific models. Conversely, GovBERT-BR, LiBERT-SE, and LegalBERT-PT, while exhibiting slightly gentler slopes, start from higher initial performance levels, reflecting the benefits of domain-specific pre-training.

6 Conclusions and Future Work

This paper introduces GovBERT-BR, a specialized pre-trained language model tailored explicitly for the Brazilian governmental context, encompassing legal and administrative domains. Through pre-training on diverse governmental texts, GovBERT-BR effectively addresses the challenges of interpreting Brazilian Portuguese and governmental documents' intricate legal bureaucratic terminology. Our experimental evaluation demonstrates GovBERT-BR's superior performance compared to existing models across various document and short-text classification tasks relevant to the Brazilian public sector, highlighting its efficacy in accurately analyzing governmental text data.

Moreover, our analysis of GovBERT-BR's convergence behavior during fine-tuning reveals its rapid adaptation to downstream tasks, highlighting its versatility and effectiveness in handling diverse text classification challenges within governmental contexts. By offering a comprehensive solution for processing governmental text data, GovBERT-BR significantly improves the efficiency and effectiveness of NLP applications within the Brazilian public sector, directly contributing to more informed decision-making and governance processes.

In summary, GovBERT-BR represents a significant advancement in applying natural language processing to the Brazilian governmental context. Our model bridges the gap left by more generic models by addressing the specific linguistic and terminological nuances of Brazilian Portuguese and the distinct domains of

legal and administrative documents. Our comprehensive pre-training approach, incorporating diverse governmental texts, ensures that GovBERT-BR can handle the complexities inherent in public administration tasks more effectively than previously available models.

Looking ahead, future research could explore the potential extensions and applications of GovBERT-BR in other domains within the Brazilian governmental landscape. Moreover, ongoing efforts to refine and optimize GovBERT-BR's performance could further enhance its utility and impact in real-world governmental applications. Finally, exploring ways to integrate GovBERT-BR with other existing NLP models or frameworks could lead to synergistic effects, potentially improving its overall performance and expanding its capabilities in addressing complex governmental challenges.

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 Amazon Web Services (AWS) and CNPq.

References

1. Al-Qurishi, M., AlQaseemi, S., Souissi, R.: Aralegal-bert: A pretrained language model for arabic legal text. In: NLLP@EMNLP. pp. 338–344. Association for Computational Linguistics (2022)
2. Aragy, R., Fernandes, E.R., Cáceres, E.N.: Rhetorical role identification for portuguese legal documents. In: Britto, A., Delgado, K.V. (eds.) Intelligent Systems - 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29 - December 3, 2021, Proceedings, Part II. Lecture Notes in Computer Science, vol. 13074, pp. 557–571. Springer (2021). https://doi.org/10.1007/978-3-030-91699-2_38, https://doi.org/10.1007/978-3-030-91699-2_38
3. Bogdanovic, M., Kocic, J., Stoimenov, L.: Srberta - A transformer language model for serbian cyrillic legal texts. Inf. **15**(2), 74 (2024)
4. Brandão, M.A., et al.: PLUS: A semi-automated pipeline for fraud detection in public bids. Digit. Gov. Res. Pract. **5**(1), 5:1–5:16 (2024). https://doi.org/10.1145/3616396
5. Constantino, K., et al.: Segmentação e Classificação Semântica de Trechos de Diários Oficiais Usando Aprendizado Ativo. In: SBBD. pp. 304–316. SBC (2022). https://doi.org/10.5753/sbbd.2022.224656
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT. pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
7. Gao, S., Gao, L., Li, Q., Xu, J.: Application of large language model in intelligent q&a of digital government. In: CNCIT. pp. 24–27. ACM (2023)
8. Garcia, E.A.S., da Silva, N.F.F., Siqueira, F., Albuquerque, H.O., Gomes, J.R.S., Souza, E., de Lima, E.A.: Robertalexpt: A legal roberta model pretrained with deduplication for portuguese. In: PROPOR. pp. 374–383. ACL (2024)
9. Han, J., Lu, J., Xu, Y., You, J., Wu, B.: Intelligent practices of large language models in digital government services. IEEE Access **12**, 8633–8640 (2024)

10. Hott, H.R., Silva, M.O., Oliveira, G.P., Brandão, M.A., Lacerda, A., Pappa, G.L.: Evaluating contextualized embeddings for topic modeling in public bidding domain. In: BRACIS. LNCS, vol. 14197, pp. 410–426. Springer (2023). https://doi.org/10.1007/978-3-031-45392-2_27
11. Licari, D., Comandé, G.: ITALIAN-LEGAL-BERT models for improving natural language processing tasks in the italian legal domain. Comput. Law Secur. Rev. **52**, 105908 (2024)
12. Lins, A.A., Carvalho, C.S., das Chagas Jucá Bomfim, F., de Carvalho Bentes, D., Pinheiro, V.: CLSJUR.BR - A model for abstractive summarization of legal documents in portuguese language based on contrastive learning. In: PROPOR. pp. 321–331. ACL (2024)
13. Luz de Araujo, P.H., de Campos, T.E., Braz, F.A., da Silva, N.C.: VICTOR: a Dataset for Brazilian Legal Documents Classification. In: LREC. pp. 1449–1458. ELRA (2020)
14. Mamalis, M.E., Kalampokis, E., Karamanou, A., Brimos, P., Tarabanis, K.A.: Can large language models revolutionalize open government data portals? A case of using chatgpt in statistics.gov.scot. In: PCI. pp. 53–59. ACM (2023)
15. Miyazaki, K., Yamada, H., Tokunaga, T.: Cross-domain analysis on japanese legal pretrained language models. In: AACL/IJCNLP (Findings). pp. 274–281. Association for Computational Linguistics (2022)
16. Polo, F.M., Ciochetti, I., Bertolo, E.: Predicting legal proceedings status: approaches based on sequential text data. In: ICAIL. pp. 264–265. ACM (2021). https://doi.org/10.1145/3462757.3466138
17. Silva, M.O., et al.: LiPSet: Um Conjunto de Dados com Documentos Rotulados de Licitações Públicas. In: DSW. pp. 13–24. SBC (2022). https://doi.org/10.5753/dsw.2022.224925
18. da Silva, N.F.F., Silva, M.C.R., Pereira, F.S.F., Tarrega, J.P.M., Beinotti, J.V.P., Fonseca, M., de Andrade, F.E., de Carvalho, A.C.P.L.F.: Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In: BRACIS. LNCS, vol. 13074, pp. 104–120. Springer (2021)
19. Souza, F., Nogueira, R.F., de Alencar Lotufo, R.: Bertimbau: Pretrained BERT models for brazilian portuguese. In: BRACIS. LNCS, vol. 12319, pp. 403–417. Springer (2020). https://doi.org/10.1007/978-3-030-61377-8_28
20. de V. Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., Furtado, V.: Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In: BRACIS. LNCS, vol. 14197, pp. 268–282. Springer (2023). https://doi.org/10.1007/978-3-031-45392-2_18
21. Viegas, C.F.O., Costa, B.C., Ishii, R.P.: Jurisbert: A new approach that converts a classification corpus into an STS one. In: ICCSA. LNCS, vol. 13956, pp. 349–365. Springer (2023)
22. Xiao, C., Hu, X., Liu, Z., Tu, C., Sun, M.: Lawformer: A pre-trained language model for chinese legal long documents. AI Open **2**, 79–84 (2021)
23. Yulita, I.N., Wijaya, V., Rosadi, R., Sarathan, I., Djuyandi, Y., Prabuwono, A.S.: Analysis of government policy sentiment regarding vacation during the COVID-19 pandemic using the bidirectional encoder representation from transformers (BERT). Data **8**(3), 46 (2023)