Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5511

HelBERT: A BERT-Based Pretraining Model for Public Procurement Tasks in Portuguese

Weslley Emmanuel Martins Lima [**Federal University of Piauí** | ***weslley@ufpi.edu.br***]
Victor Ribeiro da Silva [**Federal University of Piauí** | ***victor.silva@ufpi.edu.br***]
Jasson Carvalho da Silva [**Federal University of Piauí** | ***jasson_jcs@ufpi.edu.br***]
Ricardo de Andrade Lira Rabêlo [**Federal University of Piauí** | ***ricardoalr@ufpi.edu.br***]
Anselmo Cardoso de Paiva [**Federal University of Maranhão** | ***paiva@nca.ufma.br***]

Departamento de Computação, Universidade Federal do Piauí, Campus Universitário Ministro Petrônio Portella, Bairro: Ininga, CEP: 64049-550, Teresina, Piauí, Brazil.

Received: 05 February 2025 • **Accepted:** 18 July 2025 • **Published:** 21 February 2026

Abstract. Deep learning models excel in various tasks but require extensive annotated data for supervised learning. In NLP, limited annotated data hinders deep learning. Self-supervised pretraining addresses this by training models on unlabeled text to learn useful representations. Domain-specific pretraining is crucial for good performance in downstream tasks. Although pretrained BERT models exist for legal documents in some languages, none target public procurement documents in Portuguese. Public procurement documents have terminology that is not found in existing models. In this paper, we propose HelBERT, a BERT-based model pretrained on a large corpus of public procurement documents in the Brazilian Portuguese language, including laws, tender notices, and contracts. The experimental results demonstrate that HelBERT outperforms other models in all analyses. HelBERT surpasses models such as BERTimbau and JurisBERT in classification tasks by achieving improvements of 5% and 4% in the F1 Score, respectively. Furthermore, the model achieves gains that exceed 3% in semantic similarity tasks compared to the baseline models. Moreover, despite using a GPU with reduced memory and processing resources, the proposed approach achieves superior results with fewer and more efficient training epochs than the baseline models. These findings underscore the effectiveness of the proposed model in addressing NLP tasks within the public procurement domain.

Keywords: BERT, NLP, Pretrain, Public procurement

1 Introduction

In recent years, deep learning models have assumed a prominent role in problem-solving within the field of computational intelligence. Their superior performance, when compared to traditional models, establishes them as the state-of-the-art in computer vision, speech recognition, and natural language processing (NLP). However, the effectiveness of these models depends significantly on the availability of comprehensive training datasets, which require precise annotations in the context of supervised learning.

In the domain of NLP, although there are extensive datasets, the availability of annotated data remains limited due to the significant human effort required for annotation. Consequently, deep learning models often encounter challenges in generalizing when trained on insufficient data.

To mitigate this limitation, the researchers focus on self-supervised pretraining, a method that enables models to learn meaningful text representations without human annotation. Studies show that self-supervised pretraining improves the performance of language models by allowing them to capture essential textual features such as syntax, semantics, and polysemy.

Research indicates that the success of language models in downstream tasks depends primarily on the use of domain-specific texts during pretraining and the alignment of the language of the pretrained texts with the target problems. Models pretrained on general-purpose documents often exhibit suboptimal performance due to contextual nuances and potential ambiguities in word usage.

Within the context of pretrained language models based on BERT [Devlin *et al*., 2019], an extensive literature review reveals the absence of models tailored explicitly to public procurement documents.

When the scope of the analysis extends to the legal domain - closely associated with public procurement - pretrained models in several languages are identified [Tagarelli and Simeri, 2022; Bambroo and Awasthi, 2021; Chalkidis *et al*., 2020; Licari and Comandè, 2022]. In Portuguese, which constitutes the focus of this work, the literature highlights JurisBERT [Viegas *et al*., 2023] and LegalBERT-pt [Silveira *et al*., 2023], both pretrained on legal documents in the Brazilian Portuguese language.

Legal documents, such as laws and decrees, serve as the foundational elements governing public procurement processes, establishing a connection between the two domains. Furthermore, public procurement documents frequently incorporate terminology from related fields, including accounting and finance.

Despite these intersections, existing pretrained models fail to integrate the specialized terminology found in public procurement documents adequately. In addition, these models

are typically trained on texts in a single language, which limits their ability to process multilingual datasets.

The lack of a specialized model underscores the necessity of developing a Portuguese language model designed explicitly for public procurement. This model enables enhanced efficiency and accuracy in the analysis of procurement documents and supports the development of NLP-based solutions with superior performance compared to existing models. It further facilitates the processing of data generated daily by electronic procurement systems.

The primary objective of this work is to advance the state-of-the-art in pretrained BERT-based models for addressing challenges in the public procurement domain through NLP.

To this end, this paper introduces HelBERT (Highly efficient language model for public procurements based on BERT), which addresses these gaps and limitations. HelBERT is trained on a diverse corpus of public procurement documents, including laws, tender notices, and public contracts.

The effectiveness of HelBERT is evaluated through fine-tuning for classification and semantic similarity tasks. For benchmarking purposes, the paper employs multilingual BERT (mBERT) and BERTimbau [Souza *et al*., 2020] - both pre-trained with general purpose documents in the Portuguese language - as well as the models JurisBERT and LegalBERT-pt, which are pretrained on legal-domain documents in Portuguese. As demonstrated by the results, HelBERT consistently performs better than the other pretrained models.

The main contributions of this work include:

1. The publication of the first pretrained language model in a large corpus of documents related to public procurement in the Portuguese language.
2. We improve performance in tasks in the public procurement domain that uses long documents, such as red flag classification, by proposing a variant of HelBERT to handle long documents based on the Local-Sparse-Global(LSG) attention mechanism.
3. We demonstrated that HelBERT outperforms general-purpose and specific-purpose models pretrained with texts from a related domain in downstream tasks such as classification and textual similarity.
4. We prune part of HelBERT's layers to create a smaller model with half the layers and two times faster than the standard model. This model is adequate for applications in restricted resource environments and with massive data.
5. We have released the HelBERT models on HuggingFace[1] to aid NLP advancements in public procurement; to the best of our knowledge, these are the first language models trained on an extensive dataset consisting of public procurement-related documents.

2 Background

The diversity of meanings of words and phrases for different domains constitutes one of the main challenges in developing NLP models capable of performing effectively within specific domains [Khurana *et al*., 2023].

This section discusses the definitions relevant to pretraining a model for the public procurement domain.

2.1 Public procurement vocabulary

While there are other models for embedding words in Portuguese, such as Bertimbau, and in related domains, such as JurisBERT, the public procurement domain requires a model that captures the semantic properties of its specialized vocabulary.

Certain terms acquire unique meanings within the public procurement domain when compared to their use in other contexts, including *'encampação'* (a method of unilaterally terminating concession contracts for public works and public services) and *'inexigibilidade'* (a contracting method applied to objects for which competition is unfeasible), *'licitação'* and *'concorrência.'* Consequently, accurately representing these and other domain-specific terms requires pretraining to enable the model to assimilate these contextual nuances.

2.2 Contextualized embeddings

Over time, NLP research has adopted various approaches to textual representation. With the increasing popularity of neural network models, the vector representation known as **word embedding** has become widely used for efficiently storing relevant information about each word. This approach has demonstrated significant relevance in state-of-the-art research in natural language processing [Onan, 2021; Alatawi *et al*., 2021; Alamoudi and Alghamdi, 2021; García-Díaz *et al*., 2021; Fesseha *et al*., 2021; Srinivasan *et al*., 2021]. The core concept involves representing each term in the vocabulary as an n-dimensional vector, facilitating the mapping of similarities and differences between terms. Close vectors represent similar and related words to reflect the meaning relationships between the words.

Earlier methods relied on frequency-based approaches [Salton *et al*., 1975], but were unsuitable for applications sensitive to out-of-vocabulary words. An alternative to overcome this challenge was character-oriented embeddings [Schütze, 1993], which involved segmenting unknown words into smaller and commonly known components. Another alternative involves knowledge-based word embeddings, which incorporate external information beyond the text to enrich word representations [Xu *et al*., 2014].

Cross-lingual embeddings, commonly used in machine translation tasks, map words from multiple languages to a shared vector space [Søgaard *et al*., 2019].

However, the representations discussed thus far exhibit a critical limitation: the **meaning conflation** [Schütze, 1998]. A single vector encodes words within the semantic space, failing to account for the possibility of multiple meanings for a given word. Recent research focuses on storing information related to the word's meaning in the representation to overcome this challenge: sense embeddings.

The meaning of a word derives from the speaker's intent in the communication process. As a result, the sense of a word varies continuously and adapts to the context. These

[1] https://huggingface.co/tcepi

insights align with Firth [1957]'s conclusions, articulated in the **distributional hypothesis**, which states that words in similar contexts tend to have similar meanings. Computational linguistics adopts this principle by using a vector space where each word is represented by an n-dimensional vector that encodes information related to its meaning.

In recent years, a new generation of sense embeddings, known as contextualized embeddings, has gained prominence in NLP tasks. In this approach, the representation of a word dynamically changes according to its context. This feature has led to significant performance gains, often surpassing state-of-the-art results. Several implementations of contextualized embeddings, including ELMo [Peters *et al*., 2018], ULMFit [Howard and Ruder, 2018], and BERT, have been widely adopted, with transformer-based models such as BERT receiving particular attention.

Transformers offer distinct advantages due to their ability to leverage GPUs' parallel processing capabilities and process sentences bidirectionally, unlike recurrent neural network-based models such as ELMo, which generally process input in a unidirectional manner. Furthermore, the self-attention mechanism enables transformers to retain a high capacity for capturing context, thereby allowing the generated embeddings to better reflect the meaning of words in their respective contexts.

3 Related Works

In various fields of knowledge, pretraining BERT with domain-specific corpora has proven to be more effective than pretraining with general-purpose corpora in achieving superior results.

Lee *et al*. [2019] represents one of the first initiatives in the biomedical domain, followed by subsequent efforts in domains such as clinical medicine [Alsentzer *et al*., 2019] and science [Beltagy *et al*., 2019].

These and several other studies have shown that incorporating domain-specific knowledge improves prediction accuracy, language understanding, and contextual representation of domain-specific documents.

Since no prior work on pretraining language models with corpora from the public procurement domain was identified, this study explores models trained with datasets from related fields. Public procurement documents employ different terminology and extensively incorporate terms and definitions from the legal, accounting, and finance domains.

One of the first pretraining efforts in the legal domain was by Chalkidis *et al*. [2020], which developed a model trained on legal texts in English. This initiative achieved notable success, yielding a 2.5% improvement in multilabel classification tasks compared to the general-purpose BERT model. These results encouraged the development of similar models in other languages.

In 2021, JuriBERT, a pretrained model developed with legal texts in French, was introduced by Douka *et al*. [2021], outperforming general-purpose models trained in the same language. Similarly, in the same year, Xiao *et al*. [2021] presented a model pretrained on extensive legal documents in Chinese.

The following year witnessed additional initiatives, including pretrained models with legal texts in Arabic [Al-qurishi *et al*., 2022] and Italian [Tagarelli and Simeri, 2022]. Collectively, these efforts underscored the importance of pretraining with domain-specific corpora to achieve performance improvements over state-of-the-art models.

In the Brazilian context, JurisBERT, proposed by Viegas *et al*. [2023] as an extension of BERT, is a transformer-based model specifically designed for embedding Brazilian Portuguese legal texts and applied to Semantic Textual Similarity (STS) tasks. JurisBERT was trained from scratch using a field-specific legal corpus comprising laws, treatises, and precedents. The study demonstrated that JurisBERT achieved better precision and faster training while requiring fewer computational resources than general-purpose models such as multilingual BERT (mBERT) and BERTimbau. The JurisBERT evaluation used a dataset of 24,000 pairs of court decision summaries with varying degrees of similarity extracted from Brazilian courts' websites. The findings emphasize the effectiveness of pretraining BERT from scratch with domain-specific texts to achieve improved accuracy in the legal domain.

LegalBert-pt, developed by Silveira *et al*. [2023], is another pretrained language model specifically for the Brazilian Portuguese legal domain, aiming to address the challenges posed by its specialized language. Two variations of LegalBert-pt were created: one pretrained from scratch (LegalBert-pt SC) and another by further pretraining BERTimbau (LegalBert-pt FP) on a large and diverse corpus of 1.5 million Portuguese legal documents from Brazilian courts. Both versions were pretrained using the Masked Language Model (MLM) task. Evaluations on named-entity recognition (NER), and text classification indicated that LegalBert-pt outperformed the generic BERTimbau model in domain-specific legal tasks. This highlights the benefit of domain-specific pretraining for improving performance in legal NLP applications.

In the financial domain, Liu *et al*. [2020] stands out as a pretrained model with a corpus from the financial domain. Unlike other works that used masked language modeling (MLM) and next-sentence prediction (NSP) tasks, FinBERT used multitask learning based on six tasks, leading to more effective and robust results.

Lastly, no pretrained models with domain-specific corpora were identified in the accounting domain, highlighting a promising area for future research.

The identified gap in pretrained models for the public procurement domain underscores the need to address this research challenge. Therefore, this work aims to fill the void by proposing a pretrained model developed from scratch using domain-specific documents.

4 HelBERT models

This section introduces HelBERT, a pretrained model specialized in public procurement. It provides a detailed description of the dataset constructed using data extracted from websites specializing in public procurement, the textual preprocessing steps employed to adapt the data, the details of the pretraining process, and the development of a dataset designed to

evaluate the performance of the proposed model.

4.1 Datasets

The dataset comprises 460,000 documents, including tender notices, contracts, and tender laws, collected from four distinct sources and amounting to 47 GB of raw data. The documents used in this research were obtained by scraping data from public websites, ensuring their availability to the scientific community.

The diversity of procurement documents, which reflect the practices of numerous public institutions across a vast country like Brazil and occasionally exhibit distinct regional characteristics, reduced the risk of overfitting the model trained from scratch. This diversity enhances the generalization of the model to various public procurement scenarios.

The documents were segmented into sentences containing a maximum of 128 tokens, using the newline character as a delimiter since most of the sentences in the dataset are of this length (97% of the dataset). Longer sentences were further divided using the `sent_tokenize` method from the NLTK[2] library. Sentences shorter than ten tokens were excluded, as concise sentences typically represent noisier or incomplete text fragments. The final corpus comprises 9,827,103 sentence samples. Table 1 presents more details on the composition of the dataset.

The diversity of the dataset enabled the model to capture the linguistic patterns specific to the public procurement domain and learn the domain's terminology within just a few epochs. Given the constraints on available resources, it was crucial to balance maximizing the model's performance and using resources efficiently. Therefore, we limited the dataset size to train the model in a timely manner.

4.2 Hyperparameters and Preprocessing

Several enhancements were applied to prepare the documents for training. Initially, the documents were processed by Parsr[3], a suite of tools to convert them to plain text.

Parsr also removed irrelevant information commonly found in this document type, such as headers and footers. Additionally, regular expressions were employed to eliminate excess whitespace, special characters, section numbering, and quotations, improving the text's readability.

In a later step, numbers, email addresses, URLs, dates, and times were replaced with standardized tokens, following the approach described by Rodrigues *et al*. [2022], to reduce vocabulary size and optimize processing efficiency. Abbreviations were also expanded to their original forms.

These preprocessing strategies resulted in a dataset with coherent, high-quality sentences, enabling the language model to capture the specific characteristics of the domain within a few epochs.

The training duration was carefully determined. After pretraining over 16 epochs, the model's performance metrics were analyzed during intrinsic and extrinsic evaluations to identify the optimal epoch. The best results were achieved at the sixth epoch. An early stopping strategy with a ten-step patience was implemented to prevent overfitting. Table 2 provides the hyperparameters used for pretraining HelBERT.

[2]https://github.com/nltk/nltk
[3]https://github.com/axa-group/Parsr

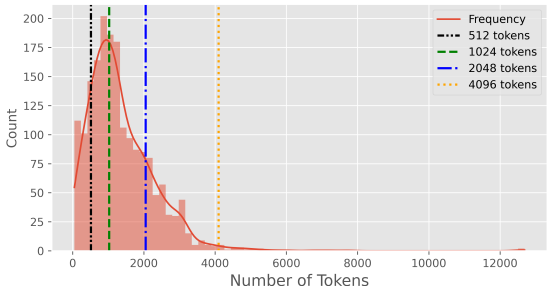

Figure 1. Frequency of samples versus Number of tokens

To construct the tokenizer, we opt for the WordPiece strategy [Schuster and Nakajima, 2012], which leverages the frequency of subwords in the training corpus. The use of subwords enables the model to capture linguistic information better and represent complex text structures.

The conversion of all words into lowercase and without accents reduces vocabulary size and allows more tokens in the corpus. The tokenizer has a total of 33,000 tokens.

The dataset was divided into 9,727,103 sentences for the training set and 100,000 for the validation set.

The pretraining strategy employed was masked language modeling (MLM) proposed by Baevski *et al*. [2019], which involves masking 15% of the tokens in the dataset.

The Next Sentence Prediction (NSP) task, originally introduced by Devlin *et al*. [2019], was not used, as subsequent studies [Conneau and Lample, 2019; You *et al*., 2020; Joshi *et al*., 2020] demonstrated its limited effectiveness.

To minimize discrepancies between the pretraining and fine-tuning phases, as highlighted by Devlin *et al*. [2019], we replace 80% of the tokens with the special token [MASK], substitute 10% with random tokens and leave the remaining 10% as their original tokens. The pretraining was conducted on an RTX 4000 GPU with 8GB memory.

4.3 HelBERT$_{\text{LSG}}$

The maximum input size poses a significant limitation for the models based on the attention mechanisms. This constraint arises from the self-attention layers with quadratic complexity ($O(n^2)$), where n is the size of the input sequence. Consequently, most BERT-based implementations support a maximum input size of 512 tokens.

However, the bid notice dataset consists of long textual sequences, with an average length of 3,080 words per notice. This limitation negatively impacts the performance of classification tasks using such datasets.

One example of a domain-relevant dataset is the red flags dataset, which classifies calls for tenders based on indications of fraud, as described in Section 5.2.2.

To address the input size limitation, the first strategy involved retaining only the Qualification Section for processing, as all red flags identified in this study are located in this section.

An analysis of the resulting dataset revealed that 85% of the samples still exceed 512 tokens, while 98% fall bellow 4,096 tokens, as illustrated in Figure 1.

Table 1: Dataset composition

Type of Documents	Source	Number of Documents	Number of Tokens
Tender notices	National government procurement portal[a]	278,739	1,185,942,256
Tender notices	TCE-PI portal[b]	142,374	277,053,127
Contracts	TCE-PI portal[b]	34,041	22,966,983
Contracts	National Public Procurement Portal[c]	4,861	5,390,659
Laws	Government Portal[d]	21	966,493

[a] https://www.comprasnet.gov.br
[b] https://sistemas.tce.pi.gov.br/licitacoesweb/mural/
[c] https://www.gov.br/pncp/pt-br
[d] https://www.gov.br/compras/pt-br/nllc

Table 2. Hyperparameters configuration for pretraining

Parameter	Value
Batch size	16
Epochs	6
Learning Rate	1e-4
Optimizer	Adam
Beta 1	0.9
Beta 2	0.999
L2 weight decay	0.1

Therefore, most samples remain beyond the input size limit of the original BERT.

Inspired by the work of Pappagari *et al*. [2019], a dimensionality reduction technique was developed for sentences exceeding 512 tokens using the average pooling technique.

First, the text was segmented into chunks of 512 tokens. Then, each slice is tokenized and transformed into contextualized embeddings using a pretrained BERT-based model. The token embeddings within each chunk are combined by averaging, creating an embedding that represents the slice. The chunk embeddings are combined using the same averaging operation, resulting in an embedding representing the document.

The document embeddings and their respective labels are then used to tune a classifier. Figure 2 details the strategy.

However, this approach has several limitations. First, it can lead to the loss of relevant local contextual information, as detailed interactions between tokens may be crucial to understanding meaning that can be hidden in the average. Second, averaging can significantly reduce the dimensionality of the information captured. Third, averaging can oversimplify complex linguistic patterns and dependencies, particularly in lengthy texts, which may not be adequately represented.

To overcome these limitations, a modified version of the model was created using Local-Sparse-Global (LSG) attention [Condevaux and Harispe, 2023], which adapts the traditional self-attention mechanism to process long texts while reducing the associated computational cost.

In the traditional attention mechanism, each token in the sentence attends to all other tokens (full attention). In LSG, however, attention is divided into three distinct types: local attention, in which the token attends to a subset of the input

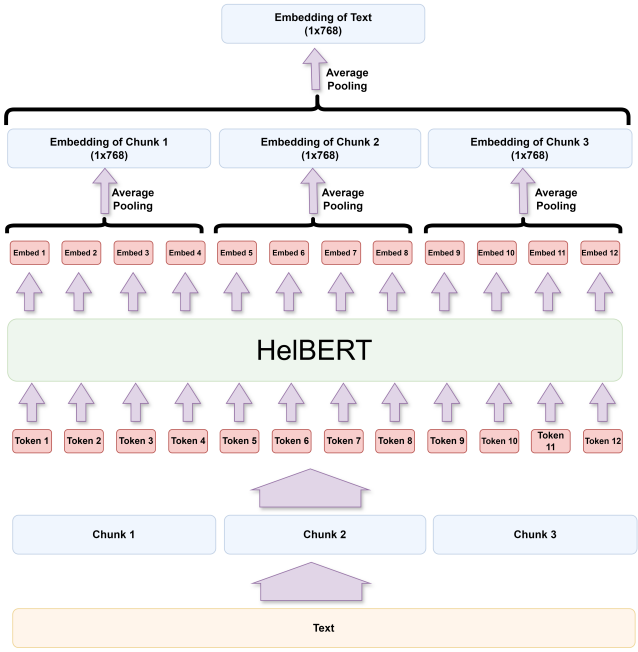

Figure 2. Average pooling technique for long texts

tokens; sparse attention, which extends attention to other selected tokens further away, based on predefined rules; and global attention, in which a small set of tokens in the sentence attends to all the tokens.

This approach demonstrated performance comparable to methods that utilize full attention while achieving linear complexity.

To implement the LSG attention mechanism, a script [GitHub, 2024] was used to adapt the traditional attention mechanism of the HelBERT model. The LSG hyperparameters were defined based on the best results obtained by the authors with the maximum sequence size of 4096, 7 global tokens, 128 local and sparse token block sizes, sparsity factor 2, and the ”norm” sparse selection pattern, which selects tokens with the highest norm values.

4.4 HelBERT$_{\text{PRUNED}}$

Transformer-based language models, such as BERT, have significantly advanced natural language processing but pose challenges related to computational resource consumption

and environmental impact.

Depending on their size and the volume of training data, models like BERT can require days or weeks of training on high-performance computing clusters. In addition, they demand equipment with substantial processing power and storage capacity. As a result, training large language models requires considerable electricity consumption, contributing to the carbon emissions associated with data center operations.

In alignment with current research trends focused on developing cost-effective models without compromising performance, we propose a compressed version of HelBERT. This version aims to achieve more efficient and faster inference while requiring fewer computational resources. Consequently, it can be deployed on devices with limited memory and processing power, making it suitable for real-time applications and environments where rapid responses are essential.

According to Sajjad *et al*. [2023], methods for reducing pretrained models can be categorized into architecture-invariant compression, knowledge distillation, and pruning. Specifically, structured pruning techniques remove less relevant parts of the network for the given task, eliminating the need for retraining.

In this study, we used the model's layers as the pruning unit. We evaluated four alternatives: Even alternate-dropping layers, Odd alternate-dropping layers, Top-layer dropping, and Bottom-layer dropping. As seen in Table 3, removing the even layers yielded the best metrics for the red flag detection task.

Using this structured pruning process and based on dropping even layers, we developed a lightweight and efficient model with six Transformer layers. This model can generate embeddings similar to those of the original HelBERT while being faster and more resource-efficient.

5 Experiments and evaluation

This section presents the evaluations demonstrating HelBERT's superior performance through intrinsic and extrinsic analysis. We have compared it with four pretrained models derived from the BERT architecture:

- **BERTimbau** - BERT-based model trained on general-purpose texts in Portuguese;
- **mBERT** - Multi-language version of BERT;
- **JurisBERT** - Model trained in Portuguese legal texts;
- **LegalBERT-pt** - Another model trained on Portuguese legal texts.

5.1 Intrinsic analysis

5.1.1 Perplexity

As a metric for evaluating the pretrained model, we employed Perplexity (denoted by PPL), which intuitively measures how "surprised" a model is when faced with new data. The lower its value, the better the model has been trained to capture the particularities of the dataset.

This metric is commonly used to evaluate autoregressive language models that generate the next token using only the previous tokens, such as GPT. However, Miaschi *et al*. [2021] demonstrated its effectiveness in evaluating models that generate a token from all other tokens in the sentence, such as BERT. The perplexity is based on the probability of a token according to the following formula:

$$p(S) \approx \prod_{i=1}^{k} p(w_i|context)$$

where *context* corresponds to all the tokens surrounding w_i in the sentence. The perplexity is given by:

$$PPL_S = e^{(\frac{p(S)}{N})}$$

where N corresponds to the length of sentence S.

To compute Perplexity, 15% of the tokens in the evaluation dataset was replaced with a special token `[MASK]`. The model then predicted the original tokens based on the surrounding context. By comparing the predicted token and the original token, we calculated the cross entropy loss for the masked tokens, quantifying the probability of the language model estimating the masked tokens based on the context.

The evaluation dataset chosen was LipSet [O. Silva *et al*., 2024], a public procurement dataset created from 9,761 documents from procurement processes in the Brazilian state of Minas Gerais. Each document was divided into sentences of up to 128 tokens and underwent the same preprocessing steps as for the pretraining dataset.

Selecting a dataset different from the one used in pretraining is essential to assess the model's ability to generalize to unseen samples. Analyzing the ability to predict masked tokens allows us to verify the model's effectiveness in capturing linguistic patterns related to the problem domain.

The perplexity values for HelBERT and the other pretrained models used as baselines are presented in Table 4.

The results demonstrate that HelBERT achieved considerably better perplexity scores than other pretrained models tailored to the legal domain (JurisBERT and LegalBERT-pt) and general-purpose pretrained models (mBERT and BERTimbau).

5.1.2 Masking of domain words

Unlike the perplexity analysis, which performs random masking of tokens, we evaluated the model's ability to predict words based on the context of public procurement. For this analysis, we selected sentences directly related to the problem domain and deliberately masked the significant words. As shown in Table 5, the proposed model demonstrated superior performance to a general-purpose model such as BERTimbau.

In three examples, the word with the highest probability predicted by the HelBERT model matched the original word in the sentence. In Example 3, although the model did not select the original word, the word with the highest probability (`aquisição`) is a synonym and fits the context perfectly. In contrast, for the BERTimbau model, none of the examples yielded the original word, and the word with the highest probability did not align correctly with the context.

5.1.3 Fertility

In language models, tokenizers play a crucial role in dividing text into smaller units called tokens. The tokenization

Table 3. Performance of language models for pruning methods

Pretrained Model	Metrics			
	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Even Alternate	**90.38**	**95.78**	93.63	**94.36**
Odd Alternate	85.98	94.84	91.17	92.13
Bottom-layers	86.20	92.52	**94.56**	92.85
Top-layers	88.29	93.65	93.52	93.06

Table 4. Perplexity of models

Pretrained Model	PPL
$mBERT_{BASE}$	96.56
$BERTimbau_{BASE}$	49.89
JurisBERT	41.48
LegalBERT-pt	48.16
HelBERT	**10.37**

method commonly adopted by BERT-based language models is WordPiece, which segments words into subwords. This approach allows for a more effective representation of out-of-vocabulary words.

The tokenizer is fundamental for the performance of language models, as tokenization quality directly impacts the accuracy and efficiency of the model. Effective tokenization, such as the one adopted and described in Section 4.2, captures linguistic nuances and enables the model to handle variations, rare words, and neologisms more effectively.

A standard metric used for intrinsic analysis is the fertility of the tokenizer [Ács, 2024; Rust *et al*., 2021; Stollenwerk, 2023], which measures the average number of subwords produced per word. An optimal tokenizer has a fertility value of 1, indicating that the vocabulary contains all words in the text.

We analyzed the fertility of the HelBERT tokenizer compared to the tokenizers of the baseline models for datasets of object types and red flags, as shown in Table 6.

The HelBERT tokenizer achieved a fertility value close to one, indicating that the vocabulary encompasses almost all words in the public procurement context. This enables the model to capture the linguistic subtleties of these documents effectively. Notably, the proposed model outperformed the legal domain models (JurisBERT and LegalBERT-pt) with vocabularies similar to the public procurement domain.

5.2 Extrinsic Analysis

The improvements observed in the intrinsic analysis of perplexity, fertility, and masking of domain words do not guarantee the best performance in extrinsic tasks, i.e., tasks related to the public procurement domain (downstream tasks). To obtain a complete analysis of HelBERT, we performed a series of experiments on downstream tasks that allow a more comprehensive evaluation of performance compared to other models based on BERT. The extrinsic analysis allows us to verify the usefulness of HelBERT in solving real problems in public procurement.

HelBERT and the baseline models were evaluated in three specific tasks in the public procurement domain. The first task involved fine-tuning the models to classify procurement objects according to the nature of the contract. The second task focused on fine-tuning the models to identify and categorize the types of red flags present in tender notices. Lastly, the third task assessed the models' ability to detect semantic similarity between the texts of procurement objects.

5.2.1 Classification of procurement objects

An essential part of a bidding notice is the section that details the subject of the purchase [Santos and Souza, 2024]. This section must clearly and precisely describe the goods or services to be contracted.

The categorization of the object section plays a crucial role in the supervision of tender calls, directing them to be analyzed by specialized sectors. An essential classification in the context of Brazilian public administration is the nature of the contract, which refers to the type of good or service to be procured. The nature of a contracted object can be categorized into four categories:

- **Engineering Works** - Activities related to the construction, maintenance, repair, or improvement of infrastructure and works and specialized technical engineering services.
- **Services** - Hiring services such as cleaning, maintenance, and technical support.
- **Consumer Goods** - Goods used in routine activities that are consumed quickly and must be replaced frequently, such as office supplies, cleaning products, and food.
- **Permanent Goods** - Long-lasting goods intended for continuous use, such as machinery, equipment, furniture, and other assets that do not wear out or need to be replaced in the short term.

To evaluate the classification performance of the HelBERT model, we used a dataset annotated on the nature of the contract, comprising 2,137 samples. This and all the other datasets used in this work are detailed in Lima *et al*. [2025]. Table 7 details the number of objects labeled for each category.

We fine-tuned HelBERT and the baseline models to compare their performance in the classification task. Each model undergoes 30 training epochs using the *Cross-Validation* approach with five *folds* and Early Stopping [Lawrence and Giles, 2000], with a *patience* parameter set to three, to prevent overfitting.

Table 5. Performance of models for masking domain words (*In Portuguese*)

ID	Sentence	BERTimbau $_{\text{BASE}}$	HelBERT
1	A regularidade do cadastramento do **licitante** será confirmada por meio de consulta ao Portal COMPRASGOV.	1. ”consumidor” (13.10%) 2. ”usuário” (11.94%) 3. ”veículo” (6.77%) 4. ”imóvel” (6.14%) 5. ”interessado” (5.22%)	1. ”**licitante**” (98.28%) 2. ”fornecedor” (0.9%) 3. ”proponente” (0.4%) 4. ”licitantes” (0.4%) 5. ”sicaf” (0.02%)
2	O critério de julgamento do lance será o **MENOR** LANCE POR LOTE.	1. ”o” (53.28%) 2. ”:” (16.29%) 3. ”um” (5.34%) 4. ”de” (4.79%) 5. ”sempre” (2.18%)	1. ”**menor**” (87.77%) 2. ”maior” (11.64%) 3. ”melhor” (0.02%) 4. ”o” (0.02%) 5. ”de” (0.002%)
3	**Contratação** de NUMERO veículos tipo camionete e um veiculo popular capacidade NUMERO passageiros, destinado ao transporte na secretaria de saúde do municipio.	1. ”Frota” (33.05%) 2. ”frota” (21.42%) 3. ”Além” (6.70%) 4. ”Grupo” (4.23%) 5. ”Transportes” (2.15%)	1. ”aquisicao” (51.54%) 2. ”locacao” (28.35%) 3. ”**contratacao**” (2.90%) 4. ”compra” (2.40%) 5. ”aluguel” (1.74%)
4	Deverá ser fornecida garantia de NUMERO anos e o atestado de **responsabilidade** técnica pela instalação.	1. ”assistência” (30.64%) 2. ”culpa” (21.37%) 3. ”qualidade” (8.44%) 4. ”conformidade” (8.03%) 5. ”responsável” (5.79%)	1. ”**responsabilidade**” (88.71%) 2. ”capacidade” (6.95%) 3. ”garantia” (0.73%) 4. ”assistencia” (0.62%) 5. ”capacitacao” (0.55%)

Table 6. Fertility of language model tokenizers in different datasets

Pretrained Model	Datasets	
	Object Types	Red Flags
mBERT$_{\text{BASE}}$	1.99	1.75
BERTimbau$_{\text{BASE}}$	1.83	1.50
JurisBERT	1.36	1.33
LegalBERT-pt	1.33	1.26
HelBERT	**1.17**	**1.18**

Table 7: Dataset of Nature of contracting objects

Category	# Samples
Engineering works	650
Permanent goods	370
Consumer goods	623
Services	494

The *batch_size* is set to four, and the *learning rate* is fixed at $1e-5$. The classification network consists of a *Transformer* layer with a dimension of 768 and two attention heads, followed by two linear layers.

The first linear layer has an input dimension of 768 and an output dimension of 30. The second linear layer consists of 30 input and four output units corresponding to each class. The adopted loss function is *CrossEntropyLoss*.

The performance of the models is evaluated using the accuracy, F1-macro, and F1-weighted metrics. The results indicate that the accuracy and F1-macro metrics, which are more sensitive to class imbalance, achieve values similar to those of the F1-weighted metric, demonstrating that imbalance did not interfere with the models' performance. Table 8 presents the metrics obtained by each model for the classification task, with values representing the average of the five runs. The best results are highlighted in bold.

The proposed models achieve better F1 weighted scores, outperforming all baseline models. In particular, the pruned version of HelBERT demonstrates better performance across all evaluated metrics despite having fewer layers.

As expected, the long-text version (HelBERT$_{\text{LSG}}$) did not exceed the base version in this task, as most samples contain texts with fewer than 512 tokens.

5.2.2 Red flag classification

Another classification task used to evaluate the performance of the model involves the classification of red flags in tender notices. Red flags are warning signs or indicators that suggest potential problems, irregularities, or fraudulent activities within the procurement process.

According to [Santos and Souza, 2024], in the context of Brazilian public procurement, **red flags** are typically associated with clauses that broadly restrict competition. Reduced competition can lead to inflated bids and the supply of inferior goods and services, causing damage to the contracting entity. This study considers the red flags in Table 9. All identified red flags correspond to prohibited requirements during the qualification phase of a public tender.

For this task, a dataset comprising 1,818 tender notices was created, with the *Qualification* section extracted. This section outlines the criteria and requirements for a company to be deemed eligible to participate in the bidding process.

Public procurement inspection experts annotated the samples in the dataset concerning the presence or absence of seven red flags, using the annotation tool provided by Argilla.

Table 8. Performance of language models for classifying procurement objects

Pretrained Model	Metrics		
	Accuracy (%)	F1-macro (%)	F1-weighted (%)
mBERT $_{\text{BASE}}$	94.15	93.92	94.19
BERTimbau $_{\text{BASE}}$	95.79	95.57	95.80
JurisBERT	94.66	94.35	94.68
LegalBERT-pt	93.26	92.86	93.34
HelBERT$_{\text{BASE}}$	96.58	96.46	96.59
HelBERT$_{\text{PRUNED}}$	**96.96**	**96.92**	**96.96**
HelBERT$_{\text{LSG}}$	95.97	95.68	95.98

Table 9: Red flags description

Red Flags	Description
Certificate of location	Requirement of an operating license or other proof of the bidder's location for qualification.
Certificate of protest	Requirement of negative protest certificate for qualification.
Paid-in Equity	Requirement of proof of paid-up share capital or shareholders' equity for qualification.
Financial suitability	Requirement of financial or banking suitability for qualification.
Number of certificates	Requirement of a minimum number, maximum, or limitation of the sum of certificates of technical capacity for qualification.
Good Practices Certificate	Requirement of certificates of good practices for qualification.
Environmental license	Requirement of an environmental license to qualify for the event.

[2024]. Table 10 details the number of samples corresponding to each type of red flag. This constitutes a multilabel classification problem, as a single tender notice can simultaneously contain more than one red flag.

Table 10: Distribution of labels for the red flags in Tender Documents dataset

Type	# Samples
Certificate of location	700
Certificate of protest	122
Paid-in Equity	330
Financial suitability	40
Number of certificates	40
Good Practices Certificate	103
Environmental license	284
No red Flags	199
Total	**1,818**

Using this dataset, the proposed and baseline models were fine-tuned. Each model underwent 60 training epochs using *Cross-Validation* with 5 *folds* and *Early Sopping* with a *patience* value of 3. The adopted *batch_size* was set to 4, and the *learning rate* was $1e-5$. The classification network's structure mirrored that used in the contract object classification task, with the sole modification being the final layer, which contains seven output units corresponding to the types of red flags in the training dataset. The loss function *Distribution-Balanced Loss* [Wu *et al*., 2020] was used to address the class imbalance. This function accounts for label imbalance, label co-occurrence, and the predominance of negative labels.

The metrics for the Red Flag Classification task are presented in Table 11. These values represent the average of five runs, with the best results highlighted in bold.

The results indicate that the best-performing model, (HelBERT$_{\text{LSG}}$), achieved an F1 score of 94.91%, surpassing HelBERT$_{\text{PRUNED}}$, the second-best model for this task, by approximately 2.13%. Moreover, all three versions of HelBERT outperformed the baseline models in this classification task.

The superior performance of Helbert$_{\text{LSG}}$ compared to other models that employed the average pooling strategy to handle long texts highlights the effectiveness of the Local-Sparse-Global (LSG) attention mechanism in addressing this specific challenge.

5.2.3 Semantic similarity

We also used a dataset to evaluate the models' ability to detect the similarity of meaning between two hiring objects.

Identifying similar objects enables categorizing contracts, which is essential to optimize resource allocation. For textual documents, the natural approach involves detecting similarities between the descriptions of the contracted objects.

Detecting similarities between tender notice objects is crucial in analyzing public procurement. Facilitates the managerial analysis of the goods contracted by public entities and aids in uncovering evidence of fraud related to the ob-

Table 11. Performance of language models for the Red Flag Classification task

Pretrained Model	Metrics				
	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	Hamming Loss
$mBERT_{BASE}$	81.03	88.39	91.71	86.56	0.032
$BERTimbau_{BASE}$	85.70	92.06	93.65	92.15	0.022
JurisBERT	86.75	92.40	92.78	92.82	0.022
LegalBERT-pt	86.97	92.12	93.03	92.38	0.023
$HelBERT_{BASE}$	87.57	92.46	93.62	92.46	0.020
$HelBERT_{PRUNED}$	87.57	92.78	94.57	92.73	0.019
$HelBERT_{LSG}$	**90.97**	**94.91**	**95.97**	**94.80**	**0.014**

ject purchased, such as split purchase. The illegal practice of splitting an expense into several more minor contracts can be identified and mitigated with the help of semantic similarity techniques.

However, applying semantic similarity techniques to the full text of contract objects yielded disappointing results. After a detailed dataset analysis, we hypothesized that the low performance was related to many terms and phrases in the object's text that were not directly related to the contracted good or service and which only enhanced the reader's understanding.

To address this issue, we used a named entity recognition (NER) model based on BERT contextualized embeddings to extract keyphrases from the object description, excluding words and expressions that might confuse models.

Using a NER model, we extracted the keyphrases of the object descriptions and constructed a dataset consisting of 1,754 pairs, extracted from the text of the object section, which experts manually annotated as similar or not similar.

We evaluated the performance of the models using cosine similarity as a scoring metric. A threshold was defined to determine whether a pair of sentences is similar. The threshold, a numerical value, was compared with similarity scores to assign each sample to a class.

Different thresholds yield varying F1 scores when compared to human annotations. Therefore, the optimal threshold for maximizing the F1 score varies by model. Accordingly, the threshold that maximizes the F1 score was calculated for each model. Table 12 presents the best thresholds and the corresponding F1 scores.

Table 12. F1-Score and Thresholds of similarity task

Pretrained Model	Metrics	
	Threshold	F1 Score(%)
$mBERT_{BASE}$	0.81	93.96
$BERTimbau_{BASE}$	0.88	93.53
JurisBERT	0.79	95.70
LegalBERT-pt	0.86	94.98
$HelBERT_{BASE}$	0.82	95.96
$HelBERT_{LSG}$	0.79	96.45
$HelBERT_{PRUNED}$	0.76	**97.20**

As observed in other tasks, the three versions of HelBERT outperformed the baseline models in semantic similarity tasks, with the pruned version achieving the best results.

All samples in the test dataset are fewer than 512 tokens in length. Consequently, results may differ using a dataset that contains longer texts, mainly for $HelBERT_{LSG}$.

5.3 Comparative analysis

Additional analyses compare HelBERT to baseline models concerning certain pretraining parameters and computational resources. Furthermore, the analyses evaluate the gains in resource consumption across the three HelBERT versions.

Table 13 compares several configurations adopted to pretraining the proposed and baseline models.

The HelBERT model was trained with fewer samples than most other models, except JurisBERT. Moreover, the model was trained for the fewest epochs of those available. Consequently, the model achieves improved results while consuming relatively few resources.

The hardware used limited the choice of batch size values, the number of samples, the maximum sentence length in tokens, and the number of pretraining epochs.

The pretraining time of the HelBERT was 72 hours, one of the shortest among the models evaluated, as shown in Table 14, which highlights the resources used and the pretraining time.

Although the RoBERTa model exhibits a shorter training duration, it requires significantly more powerful hardware. The achievement of excellent results with minimal computational resources underscores the importance of the quality and specificity of the training data for the model's performance.

5.3.1 $HelBERT_{PRUNED}$

To evaluate the performance of the pruned model, the analysis employs the metrics described by Xu and McAuley [2023], which are suitable for analyzing the compression of pretrained language models.

Floating point operations (FLOP) quantify the number of floating point computations needed to process a sample.

Carbon Footprint refers to the total carbon dioxide (CO_2) emissions generated due to the energy consumption required for training and deploying these models.

[4]The exact number of samples used during pretraining is not available, and with an average sample size maxing out at 512 tokens, the total number of samples remains unknown.

Table 13. Comparison of language model pretraining configurations

Pretrained Model	Parameters			
	Samples	Batch size	Tokens	Epochs
$mBERT_{BASE}$	3.3 billions of words[4]	256	512	40
$RoBERTa_{BASE}$	-	8.000	512	40
$BERTimbau_{BASE}$	587 millions	128	512	8
JurisBERT	1.5 milion	128	384	20
LegalBERT-pt	12 milions	-	512	-
$HelBERT_{BASE}$	9.7 millions	128	128	6

Table 14. Processing time and computational resources used to pretrain the language models

Pretrained Model	Values	
	Time(h)	Resources
$mBERT_{BASE}$	96	4x Cloud TPUs (256 GB)
$RoBERTa_{BASE}$	24	1024x V100 (32,768 GB)
$BERTimbau_{BASE}$	96	TPU-v3-8 (128 GB)
JurisBERT	168	2x RTX 3080 (24 GB)
LegalBERT-pt	-	-
$HelBERT_{BASE}$	72	RTX Quadro 4000 (8 GB)

Number of parameters is a metric representing the total count of learnable weights of the model.

Model Size reflects the storage cost of a model. Deploying an NLP model on mobile devices can be necessary due to limited storage. It can also indicate the memory footprint and computational cost for training and inference.

Inference time assesses the duration it takes to run an algorithm during its inference stage.

Speed-up ratio is the comparative metric between the inference duration of the standard model - in our context, the HelBERT model - and that of the optimized model.

We used the red flags dataset to analyze the resource consumption of the models. Table 15 shows that the $HelBERT_{PRUNED}$ model was almost twice as fast in terms of inference time, using less storage (model size) and processing (number of floating point operations) resources, as well as generating a smaller carbon footprint than the other models.

6 Conclusion

This paper introduces HelBERT, a language model explicitly designed to enhance performance in natural language processing (NLP) tasks within the public procurement domain in Portuguese. Using the Masked Language Modeling (MLM) task, the pretraining process enabled HelBERT to generate effective representations by training the model from scratch on domain-specific public procurement data.

A model variant incorporates the Local-Sparse-Global (LSG) attention mechanism, allowing HelBERT to process lengthy sequences of up to 4,096 tokens. This approach significantly improved downstream tasks, such as identifying red flags in bidding notices.

Additionally, the pruned version of HelBERT was developed, offering nearly twice the speed in training and inference compared to the base version. Surprisingly, this variant outperformed the original model in specific downstream tasks.

These three variations of HelBERT provide flexibility for different use cases. The pruned version balances performance and speed, while the LSG variant excels in handling long texts.

The creation of HelBERT and its variations has led to improvements in several downstream tasks in the public procurement domain, including sentence classification and similarity.

Despite the significant results obtained, some limitations should be considered. Firstly, all the downstream tasks used for the evaluation only involved tender documents. Future work should include downstream tasks involving other document types, such as awarded contracts and procurement laws, for a more comprehensive analysis of the public procurement domain.

Furthermore, we observed that the results obtained by the proposed model in the named entity recognition tasks were not superior to those obtained by general-purpose models, such as mBERT. Such findings suggest that there are scenarios in which pretraining with domain-specific documents does not produce better results than fine-tuning existing models.

Another limitation is related to a common problem in solutions that use deep learning: the potential bias of the model concerning training data. A small percentage of the original data set was annotated and used for training (approximately 1%). Future work will involve expanding the volume of annotated data to mitigate such biases.

The limited computing resources for this work made it impossible to evaluate other hyperparameter values, such as batch size and sentence length, which can provide significant performance gains. In addition, the increase in the variety of document types related to the domain in question must ensure a better representation of the words by the model.

Overall, HelBERT demonstrated significant improvements in the classification and semantic similarity tasks compared to general-purpose models and domain-specific models in related areas. These results highlight the potential for applying HelBERT to other NLP tasks, such as text summarization.

Finally, making HelBERT available to the scientific community promises substantial advances in public procurement research and solutions to various open challenges in the field.

Table 15. Resource consumption of HelBERT variants

Metric	HelBERT		
	BASE	**LSG**	**PRUNED**
FLOPS(Giga)	87.02	87.02	**43.51**
Carbon Footprint(Kg)	1.31e-4	1.33e-4	**6.73e-5**
Number of parameters (Millions)	109	114	**68**
Model Size(MB)	416.11	435.44	**262.71**
Inference Time(Samples/s)	52.85	52.20	**102.59**
Speed-up Ratio	1.0	0.99	**1.94**

Declarations

Authors' Contributions

WL: Conceptualization, Methodology, Investigation, Data Curation, Software, Visualization, Writing - Original Draft, Writing - Review & Editing, Project administration. VS: Methodology, Investigation, Data Curation, Software. JS: Data Curation, Software. RR: Writing - Review & Editing, Supervision. AP: Writing - Review & Editing, Supervision. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The models generated during the current study are available on Hugging Face.

References

Al-qurishi, M., Alqaseemi, S., and Souissi, R. (2022). AraLegal-BERT: A pretrained language model for Arabic legal text. In Aletras, N., Chalkidis, I., Barrett, L., Goanță, C., and Preoțiuc-Pietro, D., editors, *Proceedings Of The Natural Legal Language Processing Workshop 2022*, pages 338–344. Association for Computational Linguistics. DOI: 10.18653/v1/2022.nllp-1.31.

Alamoudi, E. and Alghamdi, N. (2021). Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal Of Decision Systems*, 30:259–281. DOI: 10.1080/12460125.2020.1864106.

Alatawi, H., Alhothali, A., and Moria, K. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374. DOI: 10.1109/ACCESS.2021.3100435.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics. DOI: 10.18653/v1/W19-1909.

Argilla., S. L. U. (2024). Open-source data curation platform for llms. Available at:`https://argilla.io/`,2024.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. (2019). Cloze-driven pre-training of self-attention networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings Of The Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China. Association for Computational Linguistics. DOI: 10.18653/v1/D19-1539.

Bambroo, P. and Awasthi, A. (2021). Legaldb: Long distilbert for legal document classification. In *2021 International Conference On Advances In Electrical, Computing, Communication And Sustainable Technologies (ICAECT)*, pages 1–4. DOI: 10.1109/ICAECT49130.2021.9392558.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics. DOI: 10.18653/v1/D19-1371.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.findings-emnlp.261.

Condevaux, C. and Harispe, S. (2023). Lsg attention: Extrapolation of pretrained transformers to long sequences. In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25–28, 2023, Proceedings, Part I*, page 443–454, Berlin, Heidelberg. Springer-Verlag. DOI: 10.1007/978-3-031-33374-3_35.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pre-training. In *Proceedings Of The 33rd International Conference On Neural Information Processing Systems*, page 11, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.5555/3454287.3454921.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and

Solorio, T., editors, *Proceedings Of The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long And Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.

Douka, S., Abdine, H., Vazirgiannis, M., El Hamdani, R., and Restrepo Amariles, D. (2021). JuriBERT: A masked-language model adaptation for French legal text. In Aletras, N., Androutsopoulos, I., Barrett, L., Goanta, C., and Preotiuc-Pietro, D., editors, *Proceedings Of The Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic. Association for Computational Linguistics. DOI: 10.18653/v1/2021.nllp-1.9.

Fesseha, A., Xiong, S., Emiru, E. D., Diallo, M., and Dahou, A. (2021). Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. *Information*, 12(2). DOI: 10.3390/info12020052.

Firth, J. (1957). *A Synopsis of Linguistic Theory,1930-1955*. Oxford Blackwell. Book.

García-Díaz, J., Cánovas-García, M., Colomo-Palacios, R., and Valencia-García, R. (2021). Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518. DOI: 10.1016/j.future.2020.08.032.

GitHub, C.-A. (2024). ccdv-ai/convert_checkpoint_to_lsg: Efficient attention for long sequence processing. Available at:`https://github.com/ccdv-ai/convert_checkpoint_to_lsg,2024`.

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 (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics. DOI: 10.18653/v1/P18-1031.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. DOI: 10.1162/tacl_a_00300.

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools And Applications*, 82(3):3713–3744. DOI: 10.1007/s11042-022-13428-4.

Lawrence, S. and Giles, C. (2000). Overfitting and neural networks: conjugate gradient and backpropagation. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 1, pages 114–119 vol.1. DOI: 10.1109/IJCNN.2000.857823.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: a pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. DOI: 10.1093/bioinformatics/btz682.

Licari, D. and Comandè, G. (2022). Italian-legal-bert: A pretrained transformer language model for italian law. In Symeonidou, D., Yu, R., Ceolin, D., Poveda-Villalón, M., Audrito, D., Caro, L. D., Grasso, F., Nai, R., Sulis, E., Ekaputra, F. J., Kutz, O., and Troquard, N., editors, *Companion Proceedings Of The 23rd International Conference On Knowledge Engineering And Knowledge Management*, Bozen-Bolzano, Italy. CEUR. ISSN: 1613-0073. DOI: 10.1016/j.clsr.2023.105908.

Lima, W., Silva, V., Silva, J., Lira, R., and Paiva, A. (2025). Bidcorpus: A multifaceted learning dataset for public procurement. *Data in Brief*, 58:111202. DOI: 10.1016/j.dib.2024.111202.

Liu, Z., Huang, D., Huang, K., Li, Z., and Zhao, J. (2020). Finbert: A pretrained financial language representation model for financial text mining. In Bessiere, C., editor, *Proceedings Of The Twenty-Ninth International Joint Conference On Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization. Special Track on AI in FinTech. DOI: 10.24963/ijcai.2020/622.

Miaschi, A., Brunato, D., Dell'Orletta, F., and Venturi, G. (2021). What makes my model perplexed? a linguistic investigation on neural language models perplexity. In Agirre, E., Apidianaki, M., and Vulić, I., editors, *Proceedings Of Deep Learning Inside Out (DeeLIO): The 2nd Workshop On Knowledge Extraction And Integration For Deep Learning Architectures*, pages 40–47, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.deelio-1.5.

O. Silva, M., P. Oliveira, G., Hott, H., D. Gomide, L., M. A. Mendes, B., A. Bacha, C., L. Costa, L., A. Brandão, M., Lacerda, A., and L. Pappa, G. (2024). Lipset: A comprehensive dataset of labeled portuguese public bidding documents. *Journal of Information and Data Management*, 15(1):196–205. DOI: 10.5753/jidm.2024.3460.

Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency And Computation: Practice And Experience*, 33(23):e5909. e5909 CPE-20-0130.R1. DOI: 10.1002/cpe.5909.

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. (2019). Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition And Understanding Workshop (ASRU)*, pages 838–844. DOI: 10.1109/ASRU46091.2019.9003958.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. DOI: 10.18653/v1/N18-1202.

Rodrigues, R. B. M., Privatto, P. I. M., de Sousa, G. J., Murari, R. P., Afonso, L. C. S., Papa, J. P., Pedronette, D. C. G., Guilherme, I. R., Perrout, S. R., and Riente, A. F. (2022). Petrobert: A domain adaptation language model for oil and gas applications in portuguese. In Pinheiro, V., Gamallo,

P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 101–109. Springer International Publishing. DOI: 10.1007/978-3-030-98305-5_10.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. (2021). How good is your tokenizer? on the monolingual performance of multilingual language models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings Of The 59th Annual Meeting Of The Association For Computational Linguistics And The 11th International Joint Conference On Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.243.

Sajjad, H., Dalvi, F., Durrani, N., and Nakov, P. (2023). On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language*, 77:101429. DOI: 10.1016/j.csl.2022.101429.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620. DOI: 10.1145/361219.361220.

Santos, F. and Souza, K. (2024). *Como combater a corrupção em licitações: detecção e prevenção de fraudes*. Fórum, fourth edition. Book.

Schuster, M. and Nakajima, K. (2012). Japanese and korean voice search. In *2012 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pages 5149–5152. DOI: 10.1109/ICASSP.2012.6289079.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123. Available at:`https://aclanthology.org/J98-1004/`.

Schütze, H. (1993). Word space. In *Advances In Neural Information Processing Systems 5, [NIPS Conference]*, pages 895–902. DOI: 10.5555/645753.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 268–282, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-45392-2_18.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-61377-8_28.

Srinivasan, S., Ravi, V., Alazab, M., Ketha, S., Al-Zoubi, A., and Kotti Padannayil, S. (2021). Spam emails detection based on distributed word embedding with deep learning. *Machine Intelligence And Big Data Analytics For Cybersecurity Applications*, pages 161–189. DOI: 10.1007/978-3-030-57024-8_7.

Stollenwerk, F. (2023). Training and evaluation of a multilingual tokenizer for GPT-SW3. DOI: 10.48550/arXiv.2304.14780.

Søgaard, A., Vulić, I., Ruder, S., and Faruqui, M. (2019). *Cross-lingual word embeddings*. Synthesis Lectures on Human Language Technologies. Springer Cham, 1 edition. DOI: 10.1007/978-3-031-02171-8.

Tagarelli, A. and Simeri, A. (2022). Unsupervised law article mining based on deep pretrained language representation models with application to the italian civil code. *Artificial Intelligence And Law*, 30(3):417–473. DOI: 10.1007/s10506-021-09301-8.

Viegas, C., Costa, B., and Ishii, R. (2023). Jurisbert: A new approach that converts a classification corpus into an sts one. *Computational Science And Its Applications – ICCSA*, 2023:349–365. DOI: 10.1007/978-3-031-36805-9_24.

Wu, T., Huang, Q., Liu, Z., Wang, Y., and Lin, D. (2020). Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, page 162–178, Berlin, Heidelberg. Springer-Verlag. DOI: 10.1007/978-3-030-58548-8_10.

Xiao, C., Hu, X., Liu, Z., Tu, C., and Sun, M. (2021). Lawformer: A pretrained language model for chinese legal long documents. *AI Open*, 2:79–84. DOI: 10.1016/j.aiopen.2021.06.003.

Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., and Liu, T.-Y. (2014). Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings Of The 23rd ACM International Conference On Conference On Information And Knowledge Management*, pages 1219–1228, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2661829.2662038.

Xu, C. and McAuley, J. (2023). A survey on model compression and acceleration for pretrained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10566–10575. DOI: 10.1609/aaai.v37i9.26255.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. (2020). Large batch optimization for deep learning: Training bert in 76 minutes. In *Proceedings of Eighth International Conference On Learning Representations*. DOI: 10.48550/arxiv.1904.00962.

Ács, J. (2024). Exploring bert's vocabulary. Available at:`https://juditacs.github.io/2019/02/19/bert-tokenization-stats.html`.