Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*

João Rodrigues[1(✉)], Luís Gomes[1], João Silva[1], António Branco[1], Rodrigo Santos[1], Henrique Lopes Cardoso[2], and Tomás Osório[2]

[1] NLX—Natural Language and Speech Group, University of Lisbon, Faculty of Sciences (FCUL), Dept. Informatics, Campo Grande, 1749-016 Lisboa, Portugal
{jarodrigues,luis.gomes,jsilva,antonio.branco,rsdsantos}@fc.ul.pt
[2] Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC) Faculdade de Engenharia da Universidade do Porto (FEUP), Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
hlc@fe.up.pt
tomas.s.osorio@gmail.com

Abstract. To advance the neural encoding of Portuguese (PT), and a fortiori the technological preparation of this language for the digital age, we developed a Transformer-based foundation model that sets a new state of the art in this respect for two of its variants, namely European Portuguese from Portugal (PT-PT) and American Portuguese from Brazil (PT-BR). To develop this encoder, which we named Albertina PT-*, a strong model was used as a starting point, DeBERTa, and its pre-training was done over data sets of Portuguese, namely over a data set we gathered for PT-PT and over the brWaC corpus for PT-BR. The performance of Albertina and competing models was assessed by evaluating them on prominent downstream language processing tasks adapted for Portuguese. Both Albertina versions are distributed free of charge and under a most permissive license possible and can be run on consumer-grade hardware, thus seeking to contribute to the advancement of research and innovation in language technology for Portuguese.

Keywords: Portuguese · Large language model · Foundation model · Encoder · Albertina · DeBERTa · BERT · Transformer · Deep learning

1 Introduction

In recent years, the field of Artificial Intelligence has come to successfully exploit the paradigm of deep learning, a machine learning approach based on large artificial neural networks [17]. Applied to Natural Language Processing (NLP), deep learning gained outstanding traction with notable breakthroughs under the distributional semantics approach, namely with word embedding techniques [19] and the Transformer neural architecture [28]. These neural models acquire

N. Moniz et al. (Eds.): EPIA 2023, LNAI 14115, pp. 441–453, 2023.
https://doi.org/10.1007/978-3-031-49008-8_35

semantic representations from massive amounts of data in a self-supervised learning process that ultimately results in the so-called *Foundation Models* [4].

Self-supervision is accomplished in NLP through language modeling [3] and was initially adopted in shallow neural models such as Word2Vec [19] for word embeddings. Over time, this approach was scaled beyond the single-token level to sequence transduction with encoding-decoding models based on recurrent [27] or convolution networks and occasionally supported by attention mechanisms [2].

A particular neural network architecture, the Transformer, has stood out among all others, showing superior performance by a large margin, sometimes even surpassing human-level performance [30,31], and became mainstream in virtually every NLP task and application [4]. Several variants have spun out from the base Transformer architecture (encoder-decoder), including the landmark encoder BERT [7] and the outstanding decoder GPT [5], which have been most successfully adapted to downstream tasks, complemented with techniques such as transfer learning [21], fine-tuning [22] or few-shot prompting [5].

The large scale of foundation models is crucial to their strength and successful deployment. Adding to the difficulty of accessing sufficiently large computational resources, most NLP research is focused on the English language, which is just one of the around 7,000 idioms on the planet. Consequently, there is a lack of competitive and openly available foundation models specifically developed for the vast majority of languages other than English, which happens to be also the case for Portuguese. This restrains the scientific progress and the innovative pace related to those languages, as well as curtailing other societal benefits, further enlarging the digital divide between English and other languages.

To the best of our knowledge, there are a couple of publicly published models that were developed specifically for Portuguese, namely for its European variant from Portugal (PT-PT) and for its American variant from Brazil (PT-BR). However, they present considerable drawbacks, namely in what concerns their sub-optimal performance level and the non-existent public distribution of encoders for the PT-PT variant.

Accordingly, there is important motivation and considerable room for improvement in creating new and better encoders for Portuguese, which we developed and present in this paper—and named as Albertina PT-*.[1] On a par with an encoder for PT-BR that sets a new state of the art for this language variant, its twin PT-PT version is an original contribution to the state-of-the-art concerning Portuguese: a freely available neural encoder specifically developed for its European variant with highly competitive performance, whose reporting publication is publicly available and which is openly distributed.

The remainder of this paper is organized as follows. Section 2 provides an overview of existing models with support for Portuguese, with a particular focus on the pre-existing BERTimbau, for PT-BR. The data sets used in pre-training and evaluating our model are presented in Sect. 3. Section 4 describes Albertina

[1] The models can be obtained here: The Albertina-PT-PT model can be obtained at https://huggingface.co/PORTULAN/albertina-ptpt and the Albertina-PT-BR model can be obtained at https://huggingface.co/PORTULAN/albertina-ptbr.

PT-* and its pre-training and fine-tuning procedures. The evaluation results of its versions on downstream tasks are discussed in Sect. 5. Section 6 closes the paper with concluding remarks.

2 Related Work

Regarding related work, we consider Transformer-based encoder models that, to the best of our knowledge, are concerned with the Portuguese language. Accordingly, besides searching the literature, we also screened the Hugging Face [13] model repository, as it has become the main source of NLP models.

Multiple studies [6,7,18,25] have shown that language-specific foundation models perform better than multilingual ones. This realization has thus led to a few initiatives that created language-specific encoders, trained from scratch for a single language, such as ERNIE for Chinese [26], BERTje for Dutch [6], CamemBERT for French [18], and MarIA for Spanish [10], among others.

Nevertheless, given it is not always viable to create a model specifically for a given language due to a lack of available data or computing resources, multilingual models have been resorted to as a temporary yet common mitigation for this problem for many languages. These are models that are pre-trained on data that include a mix of languages—albeit English is typically present in a greater amount—and are thus capable of modeling multiple languages.

2.1 Encoders Whose Multilingual Data Set Included Portuguese

Taking the number of Hugging Face downloads as a proxy for popularity and user base size, the stand-out models that support Portuguese are multilingual, namely XML-Roberta, available in Base and Large sizes, Multilingual BERT (mBERT) Base Cased, and DistilBERT Base.

Several task-specific or domain-specific models have been built upon these multilingual foundations. For instance, BioBERTpt (Portuguese Clinical and Biomedical BERT) [24] was created by fine-tuning mBERT on clinical notes and biomedical literature in Portuguese.

2.2 Encoders Specifically Concerned with Portuguese

To the best of our knowledge, for PT-PT there is the encoder presented in [20], but it is not possible to find therein clear evaluation results against downstream tasks and, most importantly, the distribution of that model is not announced.

As for PT-BR, there are a couple of encoders publicly distributed. That is the case of BERTabaporu,[2] which is of limited interest though, given its quite narrow domain, as it is a BERT-based encoder trained on Twitter data. The most popular of these two encoder models for PT-BR, by far, is BERTimbau [25].

BERTimbau is available in two model sizes, Base, with 110 million parameters, and Large, with 330 million parameters. In both cases, the authors took

[2] https://huggingface.co/pablocosta/bertabaporu-base-uncased.

an existing BERT-based model as starting point and, after discarding the word embeddings and the masked language modeling head layers, performed a hefty 1 million steps of additional pre-training on the brWaC corpus (see Sect. 3.1).

- BERTimbau Base took multilingual mBERT Base [7] as its starting point. It was pre-trained with a batch size of 128 and sequences of 512 tokens during 4 days on a TPU v3-8, performing about 8 epochs on the corpus [25, §5.1].
- BERTimbau Large took the monolingual English BERT Large [7] as the starting point, given there was no multilingual mBERT available in Large size. It was pre-trained with sequences of 128 tokens in batches of size 256 for the first 900,000 steps and sequences of 512 tokens in batches of size 128 for the final 100,000 steps. Its pre-training took 7 days on a TPU v3-8 instance and performed about 6 epochs on the corpus [25, §5.1].

Both the Base and Large variants of BERTimbau outperform mBERT in a couple of downstream tasks in Portuguese, with the Large variant being better [25]. Given this was an inaugural general-domain encoder for Portuguese, it set the state of the art for those tasks in Portuguese.[3]

Since the creation of BERTimbau, improved Transformer-based architectures have been developed that, together with more efficient training techniques, should allow better-performing models to be developed. This strengthens the motivation to develop alternative, state-of-the-art encoders also for PT-BR.

3 Data Sets

We proceed now with presenting the data sets used to pre-train Albertina PT-* and the data sets used to fine-tune it for the downstream tasks where it was extrinsically evaluated, for both PT-PT and PT-BR variants.

3.1 Data Sets for the Pre-training Stage

To secure conditions for comparability with BERTimbau, for the pre-training of Albertina PT-BR we resorted to the same data set, the brWaC corpus (Brazilian Portuguese Web as Corpus) [29]. It contains 2.7 billion tokens in 3.5 million documents and was obtained from crawling many different sites to ensure diversity. The authors report that some effort was made to remove duplicated content.

As for the pre-training of the Albertina PT-PT, we resorted to a data set that resulted from gathering some openly available corpora of European Portuguese from the following sources:

[3] As such, BERTimbau has come to serve as the basis for several other task-specific models available in Hugging Face. These task-specific models, however, appear to be unpublished, unnamed, or provide no information on their Hugging Face page; as such, they will not be covered in the present paper.

- OSCAR [1]: the OSCAR data set includes documents in more than one hundred languages, including Portuguese, and it is widely used in the literature. It is the result of a selection performed over the Common Crawl[4] data set, crawled from the Web, that retains only pages whose metadata indicates permission to be crawled, that performs deduplication, and that removes some boilerplate, among other filters. Given that it does not discriminate between the Portuguese variants, we performed extra filtering by retaining only documents whose meta-data indicate the Internet country code top-level domain of Portugal. We used the January 2023 version of OSCAR, which is based on the November/December 2022 version of Common Crawl.
- DCEP [11]: the Digital Corpus of the European Parliament is a multilingual corpus including documents in all official EU languages published on the European Parliament's website. We retained its Portuguese portion.
- Europarl [14]: the European Parliament Proceedings Parallel Corpus is extracted from the proceedings of the European Parliament from 1996 to 2011. We retained its Portuguese portion.
- ParlamentoPT: the ParlamentoPT is a data set we obtained by gathering the publicly available documents with the transcription of the debates in the Portuguese Parliament.

We filtered these data using best practice in the literature, resorting to BLOOM [16] pre-processing pipeline,[5] resulting in a data set of 8 million documents, containing around 2.2 billion tokens.

The number of documents from each source—Europarl, DCEP, ParlamentoPT, and OSCAR data—corresponds approximately to 15%, 20%, 31%, and 34% of the entire data set for PT-PT, respectively. All these data sets are publicly available, including ParlamentoPT.[6]

3.2 Data Sets for the Fine-tuning Concerning Downstream Tasks

We organized the data sets used for downstream tasks into two groups. In one group, we have the two data sets from the ASSIN 2 benchmark, namely STS and RTE, that were used to evaluate BERTimbau [25].

In the other group of data sets, we have the translations into PT-BR and PT-PT of the English data sets used for a few of the tasks in the widely-used GLUE benchmark [31], which allowed to test both Albertina variants on a wider variety of downstream tasks.

[4] https://commoncrawl.org/.

[5] We skipped the default filtering of stopwords since it would disrupt the syntactic structure, and also the filtering for language identification given the corpus was pre-selected as Portuguese.

[6] ParlamentoPT was collected from the Portuguese Parliament portal in accordance with its open data policy (https://www.parlamento.pt/Cidadania/Paginas/DadosAbertos.aspx, and can be obtained here: https://huggingface.co/datasets/PORTULAN/parlamento-pt.

ASSIN 2

ASSIN 2 [23] is a PT-BR data set of approximately 10,000 sentence pairs, split into 6,500 for training, 500 for validation, and 2,448 for testing, annotated with semantic relatedness scores (range 1 to 5) and with binary entailment judgments. This data set supports the task of semantic text similarity (STS), which consists of assigning a score of how semantically related two sentences are, and the task of recognizing textual entailment (RTE), which given a pair of sentences, consists of determining whether the first entails the second.

We did not create a PT-PT version of ASSIN 2. That would require transposing the data set, which is PT-BR, into PT-PT; however, to the best of our knowledge, there is no automatic translation system for direct translation between those variants. One solution would be to translate through an intermediate language, say English or Spanish, and then translate the result into PT-PT, but doing this would likely highly degrade the quality of the resulting benchmark by a factor that would not be possible to determine.

GLUE Tasks Translated

GLUE [31] has become a standard benchmark for model evaluation on downstream tasks. As the original GLUE is in English, we resort to PLUE [8] (Portuguese Language Understanding Evaluation), a data set that was obtained by automatically translating GLUE [31] into PT-BR. We address four tasks from those in PLUE, namely:

- two similarity tasks: MRPC, for detecting whether two sentences are paraphrases of each other, and STS-B, for semantic textual similarity;
- and two inference tasks: RTE, for recognizing textual entailment,[7] and WNLI, for coreference and natural language inference.

To obtain the PT-PT version of this benchmark, we automatically translated the same four tasks from GLUE using DeepL Translate,[8] which specifically provides translation from English to PT-PT as an option.[9]

4 Albertina PT-* Model

We describe the pre-training of the Albertina language model for Portuguese, in its two PT-PT and PT-BR versions, as a continuation of the pre-training of DeBERTa with our data sets. We also address its fine-tuning for the downstream tasks considered for its extrinsic evaluation.

[7] This is the same task as the ASSIN 2 RTE, but on different source data.
[8] https://www.deepl.com/.
[9] This is distributed at https://huggingface.co/datasets/PORTULAN/glue-ptpt.

4.1 The Starting Encoder

We take DeBERTa [12] as our starting encoder since it is reported to improve on multiple strong encoders and surpass human performance on the SuperGLUE benchmark. The main novelty in DeBERTa comes from two techniques, namely *disentangled attention* and *enhanced mask decoder*, which are related to how information about the relative and the absolute positions of tokens is encoded and handled by the model.

In other BERT-like encoders and Transformers in general, information about the position of tokens is represented as a vector, such as, for instance, a sinusoidal embedding, that is added to the content embedding of the token. The disentangled attention mechanism in DeBERTa uses separate content (H) and relative position (P) embeddings, and the attention mechanism attends separately to these embeddings. So, when calculating the cross attention $A_{i,j}$ between tokens i and j, the disentangled attention mechanism incorporates not only the usual content-to-content attention $H_i H_j^T$ but also content-to-position $H_i P_{j|i}^T$ attention and position-to-content $P_{i|j} H_j^T$ attention.

The second specific mechanism in DeBERTa, the enhanced mask decoder, incorporates information about the absolute position of tokens right before the softmax layer to predict the masked tokens. Usually, all three inputs (Query, Key, and Value) to the self-attention calculation come from the hidden states in the preceding layer, but in the enhanced mask decoder of DeBERTa the Query input is based on the absolute position of the token.

As codebase, we resorted to the DeBERTa V2 XLarge, for English, that is available from Hugging Face.[10] We use the Transformers [32] library with Accelerate [9]. The model has 24 layers with a hidden size of 1536, and a total of 900 million parameters. This version brings some changes to the original DeBERTa paper [12]. In particular: (i) it uses a vocabulary size of 128,000 and the *sentencepiece* tokenizer [15], (ii) it adds an additional convolution layer to the first Transformer layer, and (iii) it shares the position projection and content projection matrices in the attention layer.

4.2 Pre-training Albertina PT-BR

For the training of Albertina PT-BR, the brWaC data set was tokenized with the original DeBERTa tokenizer with a 128-token sequence truncation and dynamic padding. The model was trained using the maximum available memory capacity[11] resulting in a batch size of 896 samples (56 samples per GPU without gradient accumulation steps). We chose a learning rate of 1e-5 with linear decay and 10k warm-up steps based on the exploratory experiments. In total, around 200k training steps were taken across 50 epochs. Additionally, we used the standard BERT masking procedure with a 15% masking probability.

[10] https://huggingface.co/microsoft/deberta-v2-xlarge.

[11] The PT-BR model was trained for 1 day and 11 hours on a2-megagpu-16gb Google Cloud A2 VMs with 16 GPUs, 96 vCPUs and 1.360 GB of RAM.

4.3 Pre-training Albertina PT-PT

To train Albertina PT-PT, the data set was tokenized with the original DeBERTa tokenizer. The sequences were truncated to 128 tokens and dynamic padding was used during the training. The model was trained using the maximum available memory capacity[12] resulting in a batch size of 832 samples (52 samples per GPU and applying gradient accumulation in order to approximate the batch size of the PT-BR model). Similarly to the PT-BR variant above, we opted for a learning rate of 1e-5 with linear decay and 10k warm-up steps. However, since the number of training examples is approximately twice of that in the PT-BR variant, we reduced the number of training epochs to half and completed only 25 epochs, which resulted in approximately 245k steps.

4.4 Fine-tuning Albertina and BERTimbau

Albertina PT-BR and BERTimbau Large were fine-tuned for each of the 6 tasks described above (4 from GLUE and 2 from ASSIN 2), while Albertina PT-PT was fine-tuned on the 4 GLUE tasks only (as ASSIN-2 is for PT-BR). Each of these model-task combinations was fine-tuned for a range of sets of hyperparameter values, with the purpose of selecting the best-performing set of hyperparameters for each combination. Specifically, we experimented with dropout 0 and 0.1, learning rate 1e-6, 5e-6 and 1e-5, 32bit and 16bit floating point precision, and random seeds 41, 42, and 43. When combined, these ranges resulted in a considerable experimental space, with 36 experiments for each model-task pair. In every such experiment, the whole model was fine-tuned (not just its output head), for 5 epochs with batches of 16 examples.

5 Experimental Results

The experimental results obtained are reported in this section. Every score reported is the average of three runs with different seeds. The set of hyperparameters that produced the highest score on the development data for a given model/task was selected to subsequently evaluate it. It is the corresponding score over the test data that is reported.

Table 1. Performance on the ASSIN 2 tasks RTE (Accuracy) and STS (Pearson). Higher values indicate better performance, with the best results in bold.

	RTE	STS
Albertina PT-BR	**0.9130**	**0.8676**
BERTimbau Large	0.8913	0.8531

[12] The PT-PT model was trained for 3 days on a2-highgpu-8gb Google Cloud A2 VMs with 8 GPUs, 96 vCPUs and 680 GB of RAM.

5.1 Improving the State of the Art on ASSIN 2 Tasks

The performance scores of Albertina PT-BR and BERTimbau Large on the RTE task and STS task of ASSIN 2 are displayed in Table 1. Our model improves the state of the art for PT-BR on these two tasks by a quite competitive margin.

5.2 Setting the State of the Art on Portuguese GLUE Tasks

The performance of Albertina and BERTimbau Large are compared again, this time on the four tasks from PLUE, in PT-BR. As displayed in Table 2, our model continues to show superior performance, in three of these four tasks.

Table 2. Performance on the PLUE tasks, for PT-BR, namely RTE and WNLI (Accuracy), MRPC (F1) and STS-B (Pearson)

	RTE	WNLI	MRPC	STS-B
Albertina PT-BR	0.7545	0.4601	0.9071	**0.8910**
BERTimbau Large	0.6546	**0.5634**	0.8873	0.8842
Albertina PT-PT	**0.7960**	0.4507	**0.9151**	0.8799

Table 3 shows the performance of Albertina on the same four tasks from GLUE as before, but now automatically translated to PT-PT.

Table 3. Performance on the GLUE tasks translated into PT-PT, namely RTE and WNLI (Accuracy), MRPC (F1) and STS-B (Pearson)

	RTE	WNLI	MRPC	STS-B
Albertina PT-PT	**0.8339**	**0.4225**	**0.9171**	0.8801
Albertina PT-BR	0.7942	0.4085	0.9048	**0.8847**

5.3 Discussion

In this study, we present a Transformer-based foundation model that establishes a new state-of-the-art performance for multiple benchmark data sets in Portuguese. It is worth noting that the better efficacy of our model, compared to the pre-existing BERTimbau, goes on par with its better efficiency, as efficacy is achieved with significantly reduced computational requirements compared to pre-existing models. In particular, while the BERTimbau model was trained over one million steps, our model required less than a quarter of a million steps. Our model's ability to achieve superior performance with less training

time/computation likely results from resorting to all pre-trained layers, including the first layer, concerning word embeddings, and the last layer, concerning masked token prediction (the masked language modeling head), in contrast to the common practice in the literature of resetting these two layers to random weights to continue the pre-training.

With the cross-evaluation, the motivation for the creation of separated versions for the two language variants PT-PT and PT-BR is somewhat empirically justified: when evaluated on PT-PT tasks, Albertina PT-PT outperforms Albertina PT-BR in all tasks except one, where it is only marginally inferior, cf. Table 3; conversely, when evaluated on PT-BR data, Albertina PT-BR outperforms Albertina PT-PT in half of the tasks, cf. Table 2.

Although not directly comparable, the state-of-the-art English models using the original GLUE data sets[13] show performance results that are slightly superior to the results with Albertina. We hypothesized that this is due mainly to the fact that the English models were evaluated on the respective GLUE test sets (by being submitted to the automatic GLUE benchmark online), while Albertina was not. The reason was that the GLUE online service for testing was not available when we needed it and provided no notice about whether it would reopen. We had thus to evaluate our model offline, and thus on a different split of the data. We used the original development set for evaluation, and from the original training set, we used 10% for development and the rest for actual training. Moreover, we consider that the WNLI task was particularly affected by this difference in data partition given its limited sample size.

6 Concluding Remarks

In this paper, we presented Albertina PT-*, a state-of-the-art foundation model for Portuguese with 900 million parameters, of the encoder class, available in two versions, one for the European Portuguese variant from Portugal (PT-PT), and one for the American Portuguese variant from Brazil (PT-BR). To the best of our knowledge, there is no pre-existing encoder specifically developed for PT-PT that has been made publicly available and distributed for reuse. Hence, our Albertina PT-PT is a contribution in that direction and thus sets the state of the art for this variant of Portuguese. As for PT-BR, our Albertina encoder improves the state of the art, taking into account the previous level that was set by the pre-existing encoder BERTimbau, with 330 million parameters, showing superior performance in five out of six downstream tasks used for extrinsic evaluation.

As future work, we will be seeking to progress along a number of directions that may help to secure improvements in the performance of Albertina PT-*. We will experiment with training our encoder versions from scratch on Portuguese data only. It will be important to keep searching for and using better data in terms of quality (boilerplate cleaning, etc.), coverage of different genres, domains and registers, and coverage of additional Portuguese variants. And last but not least, we will be trying to obtain better encoders for Portuguese by virtue of

[13] https://gluebenchmark.com/leaderboard.

improved design, increasing their size, experimenting with more architectures, or finding better hyper-parameters.

Acknowledgments. This research was partially supported by: PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT (PINFRA/22117/2016); ACCELERAT.AI—Multilingual Intelligent Contact Centers, funded by IAPMEI (C625734525-00462629); ALBERTINA—Foundation Encoder Model for Portuguese and AI, funded by FCT (CPCA-IAC/AV/478394/2022); and LIACC—Artificial Intelligence and Computer Science Laboratory (FCT/UID/CEC/0027/2020).

References

1. Abadji, J., Ortiz Suarez, P., Romary, L., Sagot, B.: Towards a cleaner document-oriented multilingual crawled corpus. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), pp. 4344–4355 (2022)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations (ICLR) (2015)
3. Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. Adv. Neural Inf. Process. Syst. **13** (2000)
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models (2021). arXiv:2108.07258
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
6. De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M.: BERTje: A Dutch BERT model (2019). arXiv:1912.09582
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186 (2019)
8. Gomes, J.R.S.: PLUE: Portuguese language understanding evaluation. github.com/ju-resplande/PLUE (2020)
9. Gugger, S., Debut, L., Wolf, T., Schmid, P., Mueller, Z., Mangrulkar, S.: Accelerate: Training and inference at scale made simple, efficient and adaptable (2022). github.com/huggingface/accelerate
10. Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pámies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C.P., Armentano-Oller, C., Rodriguez-Penagos, C., Gonzalez-Agirre, A., Villegas, M.: MarIA: Spanish language models. Procesamiento del Lenguaje Natural, pp. 39–60 (2022)
11. Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., Varga, D.: DCEP -digital corpus of the European parliament. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC) (2014)
12. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations (2021)
13. Hugging Face. huggingface.co/Accessed Apr. 2023

14. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of Machine Translation Summit X: Papers, pp. 79–86 (2005)
15. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66–71 (2018)
16. Laurençon, H., Saulnier, L., Wang, T., Akiki, C., del Moral, A.V., Scao, T.L., Werra, L.V., Mou, C., Ponferrada, E.G., Nguyen, H., et al.: The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
18. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219 (2020)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). arXiv:1301.3781
20. Miquelina, N., Quaresma, P., Nogueira, V.B.: Generating a European Portuguese BERT based model using content from Arquivo.pt archive. In: Proceedings of the Intelligent Data Engineering and Automated Learning 23rd International Conference (IDEAL), pp. 280–288 (2022)
21. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
22. Peters, M.E., Ruder, S., Smith, N.A.: To tune or not to tune? adapting pretrained representations to diverse tasks. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP), pp. 7–14 (2019)
23. Real, L., Fonseca, E., Gonçalo Oliveira, H.: The ASSIN 2 shared task: a quick overview. In: 14th International Conference on the Computational Processing of the Portuguese Language (PROPOR), pp. 406–412. Springer (2020)
24. Schneider, E.T.R., de Souza, J.V.A., Knafou, J., Oliveira, L.E.S., et al.: BioBERTpt–a Portuguese neural language model for clinical named entity recognition. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, pp. 65–72. Association for Computational Linguistics (2020)
25. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Intelligent Systems: 9th Brazilian Conference (BRACIS), pp. 403–417. Springer (2020)
26. Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al.: Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation (2021). arXiv:2107.02137
27. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Adv. Neural Inf. Process. Syst. **27** (2014)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
29. Wagner Filho, J.A., Wilkens, R., Idiart, M., Villavicencio, A.: The brWaC corpus: a new open resource for Brazilian Portuguese. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC) (2018)
30. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., et al.: Superglue: A stickier benchmark for general-purpose language understanding systems. Adv. Neural Inf. Process. Syst. **32** (2019)

31. Wang, A., Singh, A., Michael, J., et al.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the EMNLP Workshop BlackboxNLP, pp. 353–355 (2018)
32. Wolf, T. et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)