



Open Sentence Embeddings for Portuguese with the Serafim PT* Encoders Family

Luís Gomes^(✉), António Branco, João Silva, João Rodrigues,
and Rodrigo Santos

NLX—Natural Language and Speech Group, Department of Informatics Faculdade de Ciências, Campo Grande, University of Lisbon, 1749-016 Lisboa, Portugal
{luis.gomes, antonio.branco, jrsilva, jarodrigues, rdsantos}@fc.ul.pt

Abstract. Sentence encoder encode the semantics of their input, enabling key downstream applications such as classification, clustering, or retrieval. In this paper, we present Serafim, a family of open-source sentence encoders for Portuguese with various sizes, suited to different hardware/compute budgets. Each model exhibits state-of-the-art performance and is made openly available under a permissive license, allowing its use for both commercial and research purposes. Besides the sentence encoders, this paper contributes a systematic study and lessons learned concerning the selection criteria of learning objectives and parameters that support top-performing encoders.

Keywords: Sentence encoder · Large Language Model · Portuguese

1 Introduction

Under the mainstream paradigm for Natural Language Processing (NLP), meaning is ultimately represented under the form of a vector, or embedding, for linguistic forms of any length be it a word, a multi-word sentence, a multi-sentence paragraph or a multi-paragraph text, etc. And under the mainstream neural architecture for NLP, the Transformer [27], a representation of the meaning of a linguistic input can be obtained out of the latent representations on its layers for that input, for instance by picking the representation at a top layer.

This architecture has unfolded into different families of encoder-decoders, decoders and encoders. Despite the outstanding visibility that the decoder models have deservedly garnered, especially with the availability of ChatGPT for the general public, the models of the encoder family have not lost their traction as they have maintained a competitive performance in non-generative tasks, especially in those tasks primarily related to classification [30].¹

¹ As a way of confirmation of this remark, the top performing model in the SuperGLUE benchmark (<https://super.gluebenchmark.com/leaderboard>) is an encoder, namely the Vega v2 model [30] at the time of this writing.

Accordingly, encoders have been the type of Transformer resorted to when the objective is to obtain a most performant representation of the meaning, aimed at being embedded in some further downstream tasks. Though this may concern any type of linguistic input—words, sentences, paragraphs, etc.—, the literature has referred to this as sentence encoding, or sometimes sentence embedding, a terminology that we will also abide by here.

Getting at the meaning representation of linguistic expressions per se is a necessary step when it is crucial to determine the degree of semantic synonymy, similarity or relatedness among them. A prototypical use case is found at the downstream task of semantic-driven information retrieval, where at some point the input query and the relevant snippets to be retrieved are eventually screened and selected on the basis of the respective meaning representations.

With a regular encoder, the typical approach to compare sentences is to concatenate them, using a separator token, and feed that to the model. The resulting embeddings are then provided to a separate layer that learns to classify those sentences with respect to their semantic relation. For some tasks, however, this approach has the drawback of requiring a large number of comparisons, which easily becomes intractable. For instance, in the prototypical information retrieval scenario, any new query would be paired with every candidate snippet to be retrieved—which in real usage scenarios have a very large volume—so that the embeddings of their concatenation can be calculated and classified.

To deal with this, standalone sentence embeddings are calculated by relying on a siamese encoder, an approach introduced by sBERT [16] that became known in the literature as sentence encoder. Given two sentences, these are fed, one at a time, to the same encoder, and the resulting embeddings are pooled into two embeddings and passed through a loss function designed to bring closer together in vector space the embeddings of semantically similar sentences.

An encoder trained in this way learns to map input into embeddings that live in a latent space where distance reflects semantic similarity. This allows to semantically compare sentences, through their embeddings, using metrics such as cosine distance, thus leveraging a wide range of applications such as semantic classification, clustering, retrieval, etc. Returning to the prototypical retrieval scenario, instead of calculating the embedding of every query-candidate pair for each new query, a sentence encoder can pre-calculate standalone embeddings for every candidate in the data set, decoupled from any particular query.

Adding to the difficulty of accessing sufficient computational resources, most NLP research focuses on English, which is just one of the around 7,000 idioms on the planet. Consequently, there is a lack of competitive and open sentence encoders specifically developed for the vast majority of languages other than English, which happens to be also the case for Portuguese. To the best of our knowledge, there are a couple of publicly published encoders that were developed specifically for Portuguese. However, as we discuss in this paper, they present drawbacks, including in what concerns their sub-optimal performance level.

Prompted by this motivation, we report here on a systematic study concerning sentence encoding for Portuguese. As a result, we present Serafim, a family of

sentence encoder models, available in different sizes, that establish a new state of the art for the respective core benchmarks of semantic text similarity (STS) and information retrieval (IR) in Portuguese. With three versions, for 100 Million, 335M and 900M parameters, these encoders are an open code, open access and open license collection of top performing sentence encoders in their class specifically developed for Portuguese, covering both the European variant, spoken in Portugal (PTPT) and the American variant, spoken in Brazil (PTBR).²

The remainder of this paper is structured as follows: Sect. 2 discusses related work; Sect. 3 describes the method proposed in this study; Sect. 4 introduces the data sets and base models used; Sect. 5 reports on the experiments and discusses the results obtained; and finally, Sect. 6 closes this paper with concluding remarks and considerations on future work.

2 Related Work

This section presents work related to the goal of this paper, namely sentence encoding for Portuguese.

Sentence Encoders not Specifically for Portuguese. There exist embedders able to handle Portuguese as their underlying encoder is multilingual and had some exposure to Portuguese during its pre-training. Besides these, it is also interesting to consider encoders for the best resourced language, that is English.

To guide the choice of which of these models to compare with our own, we rely on the information provided on the SBERT site, with exhaustive information on sentence encoding,³ and on the results reported in the Massive Text Embedding Benchmark (MTEB) Leaderboard [13].⁴ We selected a few that are best either of their class or under some interesting criterion. These models and associated criterion are listed in Table 1, and their full names are in Table 5.

Table 1. Sentence encoders not specifically for Portuguese. Second column reports sizes of models (parameters) and embeddings (bits)

Model	Size	Reason for choice
distiluse 135 multilingual	135M 512	best multilingual model [†]
mpnet 109 en	109M 768	best English model [†]
gte 137 en	137M 768	best under 250M [‡]
gist embedding 109 en	109M 768	2nd-best under 250M [‡]
gte 434 en	434M 1024	best under 1B [‡]
mini 23 en	23M 384	fast, close to best [†]

[†]According to SBERT site. [‡]According to MTEB benchmark.

² Serafim models are available at <https://huggingface.co/PORTULAN>.

³ https://www.sbert.net/docs/pretrained_models.html.

⁴ <https://huggingface.co/spaces/mteb/leaderboard>.

The STJ-IRIS Family of Encoders for Portuguese. Aimed at supporting a semantic search system for Supremo Tribunal de Justiça (Supreme Court of Justice), the STJ-IRIS encoders [12] are built on top of the Legal BERTimbau encoder, a BERTimbau Large [22] model adapted to the legal domain.

The encoders in this STJ-IRIS family vary in terms of the methods and data used for their creation. Two unsupervised methods are used in the adaptation to the legal domain, namely the standard masked language modeling (MLM) and TSDAE [29]. The sentence embedding training that follows also varies in terms of the task and training objective, resulting in over 20 models in this family.

For the sake of reference and comparison, we select the STJ-IRIS models in Table 2, whose scores on their respective HuggingFace model card indicates them as having the best performance on the various datasets used in their study.

Table 2. Chosen STJ-IRIS models

Model	Best on (Pearson score)	Aliased here as
mlm-mkd-nli-sts-v1	IRIS-STs (0.850)	stjiris 335 top irtis-sts
tsdae-mkd-nli-sts-v1	STs-B (0.855)	stjiris 335 top sts-b
mlm-gpl-nli-sts-MetaKD-v0	ASSIN (0.811)	stjiris 335 top assin
tsdae-mkd-nli-sts-v0	ASSIN2 (0.841)	stjiris 335 top assin2

It is worth noting that these encoders were trained with data from a specific domain, and thus not all available PT data was used. Also, recent methods for sentence encoding that are known to support better performance—viz. CoSENT [24], AnglE [10], GISTEmbed [21]—were not used. All this suggests that there is likely room for improvement.

Other Models for Portuguese. On HuggingFace, one can find a few other models that are presented as having been specifically trained for Portuguese, though they are not documented in some peer reviewed publication.

One finds, for instance, `jmbrito/ptbr-similarity-e5-small`,⁵ which results from fine-tuning of `intfloat/multilingual-e5-small` on ASSIN2. It is reported to achieve a score of 0.79934, though unclear under what metric. We also find a set of 22 sentence embedding models under the name `MTEB-PT`.⁶ Instead of being built upon BERT-like contextual embeddings, like the sentence encoders mentioned above, they all appear to use simple, non-contextual word embeddings (Word2Vec, Glove, etc.). No performance scores are reported.

As pointed out, these models lack a peer-reviewed publication or do not report performance scores and as such we will not consider them for our study.

⁵ <https://huggingface.co/jmbrito/ptbr-similarity-e5-small>.

⁶ <https://huggingface.co/mteb-pt>.

3 Methodology

The SBERT architecture is the mainstream approach to sentence encoders. It is rather straightforward and has not changed since its introduction in [16]. Much of the research has focused instead on the development of new training methods, in particular developing loss functions that can be used with different types of data, as each has specific use-cases, strengths and weaknesses.

Supervised Methods. We performed two types of supervised training. The first takes as input data pairs of sentences annotated with a similarity score, and uses a loss function based on the distance between the outcome of the similarity among sentence embeddings (measured with a vector similarity function such as cosine) and the label (scaled to the range of the selected similarity function). This is known as the Semantic Text Similarity (STS) task.

We experimented with three selected loss functions. The first is the canonical cosine similarity loss, proposed in the original SBERT paper, which we used as starting point and baseline for our exploratory experiments. The other two loss functions are the recently proposed CoSENT [24] loss and the AngleE [10] loss, which are tweaked versions of the same working principle as the first one. Both CoSENT and AngleE perform significantly better than the original cosine similarity loss, but very close to each other in all trial experiments.

Another type of supervised training makes use of either pairs of semantically close sentences or triples containing one anchor sentence, one positive sentence (semantically close to the anchor) and one negative sentence (either unrelated or semantically opposite to the anchor). Unlike the previous method, here no scores are given to quantify the relatedness of input sentences. This is the type of training supported by the Multiple Negatives Ranking (MNR) loss [8], which, more formally, takes as input a set of positive pairs (a_i, b_i) which is augmented with negative pairs (a_i, b_j) , obtained by pairing each a_i with all b_j sentences, assuming that any two sentences not paired at the input are unlikely to be related. Without a score for quantifying the semantic closeness of positive and negative pairs, this loss function maximizes the difference between the cosine similarities of the positive and negative pair embeddings, thus bringing the embeddings of positive pairs closer together than the embeddings of negative ones.

Building on the MNR loss, GISTEmbed [21] brings in a *guide model*, which is another sentence encoder used to generate embeddings for every sentence and filter out negative pairs that have a similarity too close (or higher) than positive pairs. By filtering out these negative pairs, GISTEmbed improves the signal-to-noise ratio of the training data, and thus the performance of the resulting model, at the expense of additional memory and compute time for the guide model.

In our exploratory experiments, we compared the MNR and GISTEmbed losses and opted for the latter because, despite requiring more time to train on the same data amount, it has shown significantly better performance on fixed-data trial experiments. To train the final model, we used a small interim model trained with a combination of the other methods as guide. The interim model performed comparably to the STJ-IRIS models, despite being 3 times smaller.

Unsupervised Methods. To mitigate the scarcity of data annotated in a way that is suitable to the training of sentence encoders, we also experimented with methods that do not require labeled data. Typically, these do not deliver results as good as supervised methods and as such we use unsupervised methods not as an alternative to supervised ones but as a supplement to them, allowing us to get some extra (pre-)training from unlabeled data. We experimented with Transformer-based Sequential Denoising Auto-Encoder (TSDAE) [29] and Contrastive Tension Loss with In-Batch Negatives (CT) [3].

TSDAE relies on the concept of a denoising auto-encoder. Given a sentence, the approach introduces some noise into it, through the deletion of random tokens, before calculating its sentence embedding. It then tries to reconstruct the original, denoised sentence from the embedding.

CT uses the same technique as MNR to automatically create negative pairs from positive ones. However, instead of using positive pairs of closely related sentences, this method only requires individual sentences, which are paired with themselves. Two slightly different instances of the model are trained simultaneously and, for each input pair, one instance is used to encode the first sentence and the other to encode the second, producing similar but not identical embeddings for identical sentences. Both models are trained with a binary cross-entropy objective, whereby positive pairs should score 1 and negative pairs 0. In the end, one of the models is discarded and the other is kept as the resulting model.

Sequential Transfer Learning. Each of the final Serafim models result from a combination of three training methods, applied sequentially in four stages with different types of data. Based on our experiments, the following order produces the best performing models:

First, the pre-trained base encoder models are trained for one epoch with Contrastive Tension Loss with In-Batch Negatives, using a set of selected high-quality Portuguese corpora, already segmented at sentence level.

Then, the models are trained for 10 epochs using GISTEmbed on data obtained by converting available NLI datasets to (anchor, positive, negative) triples –this conversion is described in the next section along with the NLI datasets. The validation splits of all the STS datasets are combined and used as validation set to select the best model checkpoint.

The third training stage employs the training splits of all available STS datasets combined, for 20 epochs, using both the CoSENT and Angle loss functions separately. Like in the previous stage, the STS validation splits are used to choose the best performing checkpoint. The performance on the validation set is also used to choose the best of the two models, one trained with CoSENT and the other with Angle. The models produced at this stage are the Serafim models for Semantic Textual Similarity tasks.

In the fourth and final stage, the models are trained for one epoch on a large information retrieval dataset, consisting of (*query, relevant passage, irrelevant passage*) triples that are fed to the GISTEmbed training method. The resulting models are the Serafim models for Information Retrieval.

4 Data and Base Models

Raw Text. For the unsupervised training of our models, we made use of several of the Portuguese corpora from the OPUS Collection [26]. While these are parallel corpora, we take only the Portuguese part. We made this choice as the corpora from this collection are generally well-curated and have better quality than some monolingual web crawl. We selected the following corpora:

EUbookshop v2 [5]: Corpus of documents from the Publications Office of the European Union with 4,172K sentences and 102.6M tokens, PTPT variant.

Europarl v8 [9]: Proceedings of the European Parliament with 2,000K sentences and 51.0M tokens, PTPT variant.

TED 2020 v1 [17]: a crawl of nearly 4000 TED and TED-X transcripts from July 2020, translated by a community of volunteers, with 329K sentences and 5.2M tokens, mostly PTBR variant.

Tatoeba v2023-04-12 [25]: database of sentences and translations, from voluntary contributions, released under CC-BY 2.0 FR, with 228K sentences and 1.6M tokens, mostly PTBR variant.

Labeled Data. We train and evaluate Serafim on data sets, available from HuggingFace, for Semantic Text Similarity (STS), Natural Language Inference (NLI) and Information Retrieval (IR) tasks, namely:

ASSIN [6] is a data set that consists of sentence pairs manually annotated for STS (a 1–5 score). Both PTBR and PTPT are provided, with each variant having 2.5k/0.5k/2.5k sentences for training/development/testing.

ASSIN2 [15] is similar to ASSIN, but it is only for PTBR, though a larger size (6.5k/0.5k/3k) for that variant.

STsb Multi MT [23] results from the automatic translation of the STS benchmark used in the SemEval shared task [4]. Its size is 5.7k/1.5k/1.3k.

IRIS-STs [12], unlike the data sets above, is a very small STS data set from a specific domain (legal), that resulted from the same project that created the STJ-IRIS family of models mentioned in Sect. 2. This data is of the PTPT variant and its size is 1.7k/0.5k/0.5k.

PLUE [7] was obtained through automatic translation of the GLUE [28] benchmark (English) to PTBR. It comprises several tasks, from which we use only MNLI (393k/19.6k/19.6k) and SNLI (511k/9.8k/9.8k), which are both data sets for the NLI task where pairs of sentences are labeled as *entailment*, *neutral* or *contradiction*. We adapted these pairs into (anchor, positive, negative) sentence triples, for use with the GISTEmbed loss, as follows: pairs labeled as entailment (resp. contradiction) are considered as positive pairs (resp. negative), and sentences that are paired with both positive and negative sentences become anchors, resulting in a total of 534k triples.

mMARCO [2] is a multilingual version of the MS MARCO [1] data set for Information Retrieval. The translation was done automatically, using Google

Translate, and includes a translation to PTBR. The data set contains a collection of 8.8 million passages and 100k queries, and for each query a list of most relevant passages (qrels file). The mMARCO training set for passage retrieval contains about 40 million of (*query, relevant passage, irrelevant passage*) triples that were produced from the queries, passage collection and qrels. mMARCO does not comprise the MS MARCO test set,⁷ so we performed this translation ourselves, also using Google Translate.

Foundation LLMs Underlying Serafim PT*. A sentence encoder model is built upon an existing, pre-trained encoder. The current state-of-the-art encoders for Portuguese, which form the “encoder ecosystem” described in [19], are the Albertina [18] and the BERTimbau [22] models. These cover a range of sizes, from 100 million to 1.5 billion parameters and, apart from BERTimbau which was pre-trained on Brazilian Portuguese (PTBR) only, support both the PTBR and European Portuguese (PTPT) variants.

5 Experiments and Discussion

A selection of the main results obtained with our experiments for **Semantic textual Similarity (STS)** are in Table 3. The test data are represented in columns, with the three leftmost concerning PTPT and the remainder ones PTBR. The encoders are represented in rows, clustered into four groups.⁸

The two groups at the top include encoders for Portuguese fine-tuned for STS: the first group with our encoders; the second group with the best representatives of previous encoders in the literature.⁹

The two groups at the bottom include the baseline encoders. The penultimate group includes the foundational encoders for Portuguese, Albertina and BERTimbau, that are the base code for our sentence encoders and that were not fine-tuned for sentence encoding. The last group includes state-of-the-art sentence encoders that are multilingual or specific for English.

Portuguese Datasets are Crucial for Having Top Performance. Sentence encoders specifically trained for Portuguese (top two groups) outperform sentence encoders that were not trained for Portuguese (bottom groups) but were nevertheless used to produce sentence embeddings for Portuguese. This happens by a very large margin of over some 10 to 20 points on average.

Fine-Tuning is Needed for Top Performing Sentence Encoders. When fine-tuning foundational models for sentence encoding, the resulting encoders

⁷ The test set is found in the msmarco-test2019-queries.tsv file, downloaded from the webpage of the 2019 edition of the TREC Deep Learning Track (<https://microsoft.github.io/msmarco/TREC-Deep-Learning-2019.html>).

⁸ Their fully fledged identification is in Table 5.

⁹ From the stjiris encoder collection [12], a particular one was selected when it was top performing among them all in at least one of the test datasets according to their Hugging Face models cards. It is worth noting that Pearson correlation scores are reported there, while here, Table 3 reports Spearman correlation scores.

Table 3. Results on STS, with Spearman correlation coefficient

Model	IRIS-STS	STSB	ASSIN	ASSIN	ASSIN2
	ptpt	ptpt	ptpt	ptbr	ptbr
<i>Our sentence encoders</i>					
serafim 900	0.8231	0.8570	0.8307	0.7883	0.8267
serafim 335	0.8226	0.8486	0.8210	0.7978	0.8323
serafim 100	0.8290	0.8350	0.8119	0.7871	0.8214
<i>Previous sentence encoders for Portuguese</i>					
stjiris 335 top iris-sts	0.7772	0.8486	0.7985	0.7734	0.8127
stjiris 335 top sts-b	0.7873	0.8550	0.8001	0.7713	0.8160
stjiris 335 top assin	0.7724	0.8364	0.7829	0.7829	0.8112
stjiris 335 top assin2	0.7526	0.8539	0.7724	0.7724	0.8105
<i>Base encoders</i>					
albertina 900 ptpt	0.6560	0.5815	0.6183	0.6183	0.5754
albertina 900 ptbr	0.6392	0.5781	0.6239	0.6239	0.5648
bertimbau 335 ptbr	0.6641	0.5801	0.5445	0.6608	0.6192
albertina 100 ptpt	0.6295	0.5604	0.6099	0.6099	0.6034
albertina 100 ptbr	0.6279	0.5177	0.6126	0.6126	0.5875
bertimbau 100 ptbr	0.6869	0.5388	0.5968	0.6442	0.6145
<i>Previous top-ranking multilingual and English sentence encoders</i>					
gte 434 en	0.6705	0.6906	0.5147	0.5170	0.5971
gte 137 en	0.6977	0.6590	0.5802	0.5626	0.5503
distiluse 135 multilingual	0.7648	0.7687	0.7374	0.6982	0.7170
mpnet 109 en	0.6484	0.6212	0.5437	0.5062	0.5631
gist embedding 109 en	0.6736	0.6679	0.5214	0.5387	0.5450
mini 23 en	0.6386	0.6156	0.5664	0.5300	0.5644

(top group) outperforms the base models (third group) by a large margin, of well over 20 points on average.

Our Sentence Encoders are the State of the Art for Portuguese. For every test dataset, our family of encoders (top group) outperform previous encoders for Portuguese (second group) by 0.02 (average weighted by dataset size) in terms of Spearman correlation coefficient, with the deltas: 0.0417 (IRIS-STS), 0.0020 (STSB), 0.0306 (ASSIN), 0.0149 (ASSIN), 0.0163 (ASSIN2). As expected, they also outperform both all baselines (third group) and all multilingual and English-specific alternatives (bottom group).

Separate Training for Portuguese Variants Could Lead to Improved Performance. In our exploratory experiments, we found that with the amount of data currently available for Portuguese, we got better performance by training with all data, including both language variants, than training separate models

for each variant. However, despite all Serafim models being trained with the same data, we observe that Serafim 900, based on Albertina PTPT, is better on the PTPT datasets while Serafim 335, based on BERTimbau (PTBR), takes the lead on PTBR ones. Thus, if more data becomes available in the future, training separately for each variant could lead to an increase in performance.

Larger Encoders Tend to Outperform Smaller Ones as expected. Our largest encoder, with 900M parameters, has the top scores in two datasets (STSb and ASSIN) and the second best scores in the others. The second largest, 335M, always scores better than the smaller one, except for IRIS-STS, where the smaller one, 100M, actually is the top one, outperforming even the largest 900M encoder.

Datasets Offer Different Levels of Difficulty. As expected, some datasets are easier to resolve, with 0.8570 for STSb, than others, with 0.7978 for ASSIN. In this connection, it is worth noting that the IRIS-STS dataset is not only the smallest one but also the only one of a specific domain (legal). All the others are of general domain. The higher performance of Serafim 100, compared to the larger Serafim encoders, seems to indicate that the smaller model has adapted better to the specific domain, with fewer data available, than the larger ones.

Table 4. Results on IR, with MRR@10

Model	mMARCO ptbr
<i>Our sentence encoders</i>	
serafim 900 ir	0.8539
serafim 335 ir	0.8278
serafim 100 ir	0.8118
<i>Previous sentence encoders for Portuguese</i>	
stjiris 335 top assin	0.6197
stjiris 335 top assin2	0.5564
stjiris 335 top sts-b	0.5250
stjiris 335 top iris-sts	0.4998
<i>Previous top-ranking multilingual and English sentence encoders</i>	
gte 434 en	0.7414
gist embedding 109 en	0.7227
gte 137 en	0.7104
mini 23 en	0.6875
mpnet 109 en	0.6685
distiluse 135 multilingual	0.6494

Information Retrieval (IR). A selection of the main results obtained with our experiments for IR are in Table 4. The Serafim IR models result from the fourth training stage, described in Sect. 3, using the mMARCO data (see Sect. 4). For

evaluation, we use the MS MARCO test set of the 2019 edition of the TREC Deep Learning Track. With scores clearly over 0.80, the Serafim IR encoders outperform both the previous encoders for Portuguese and the multilingual and English encoders by a large margin, where the best one among them is the GTE encoder for English with 109M parameters, scoring 0.74. The Serafim encoders set thus the state of the art for Portuguese concerning IR under this dataset and, as expected, larger models have better performance.

6 Conclusion

This paper presents a systematic study concerning sentence encoding for Portuguese. Its central result is the contribution of a family of new state-of-the-art sentence encoders for this language, with 100M, 335M and 900M parameters, that are openly distributed as open source under an open license, and that serve both its European variant, spoken in Portugal, and its American variant, spoken in Brazil. These join a wider family, which includes decoders [20], and data sets [14]. As future work, to further enhance encoders performance, we will seek to acquire larger volumes of data by exploring data augmentation techniques.

Table 5. Names of the sentence encoders as they appear on HuggingFace

Model	Name on HuggingFace	Ref.
stjiris 335 top iris-sts	stjiris/bert-large-portuguese-cased-legal-mlm-mkd-nli-sts-v1	[12]
stjiris 335 top sts-b	stjiris/bert-large-portuguese-cased-legal-tdsae-mkd-nli-sts-v1	[12]
stjiris 335 top assin	stjiris/bert-large-portuguese-cased-legal-mlm-gpl-nli-sts-MetaKD-v0	[12]
stjiris 335 top assin2	stjiris/bert-large-portuguese-cased-legal-tdsae-mkd-nli-sts-v0	[12]
albertina 100 ptpt	PORTULAN/albertina-100 m-portuguese-ptpt-encoder	[18]
albertina 100 ptbr	PORTULAN/albertina-100 m-portuguese-ptbr-encoder	[18]
bertimbau 100 ptbr	neuralmind/bert-base-portuguese-cased	[22]
bertimbau 335 ptbr	neuralmind/bert-large-portuguese-cased	[22]
albertina 900 ptpt	PORTULAN/albertina-900 m-portuguese-ptpt-encoder	[18]
albertina 900 ptbr	PORTULAN/albertina-900 m-portuguese-ptbr-encoder	[18]
distiluse 135 multilingual	distiluse-base-multilingual-cased-v1	[16]
mpnet 109 en	all-mpnet-base-v2	n/a
gist embedding 109 en	avsolatorio/GIST-Embedding-v0	[21]
gte 137 en	Alibaba-NLP/gte-base-en-v1.5	[11]
gte 434 en	Alibaba-NLP/gte-large-en-v1.5	[11]
mini 23 en	all-MiniLM-L6-v2	n/a

Acknowledgement. This work 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); and IMPROMPT—Image Alteration with Language Prompts, funded by FCT (CPCA-IAC/AV/590897/2023).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bajaj, P., Campos, D., Craswell, N., et al.: MS MARCO: a human generated machine reading comprehension dataset. [arXiv:1611.09268](https://arxiv.org/abs/1611.09268) (2018)
2. Bonifacio, L., Jeronymo, V., Queiroz Abonizio, H., et al.: mMARCO: a multilingual version of the MS MARCO passage ranking dataset. [arXiv:2108.13897](https://arxiv.org/abs/2108.13897) (2022)
3. Carlsson, F., Gogoulou, E., Yliää, E., Cuba Gyllensten, A., Sahlgren, M.: Semantic re-tuning with contrastive tension. In: ICLR (2021)
4. Cer, D., Diab, M., et al.: SemEval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th SemEval (2017)
5. EUbookshop. <https://bookshop.europa.eu/>
6. Fonseca, E., Santos, L., Criscuolo, M., Aluísio, S.: ASSIN: avaliação de similaridade semântica e inferência textual. In: 12th PROPOR, pp. 13–15 (2016)
7. Gomes, J.R.S.: PLUE: Portuguese language understanding evaluation (2020). <https://github.com/ju-resplande/PLUE>
8. Henderson, M., et al.: Efficient natural language response suggestion for smart reply (2017)
9. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Proceedings of Machine Translation Summit X: Papers, pp. 79–86 (2005)
10. Li, X., Li, J.: Angle-optimized text embeddings. [arXiv:2309.12871](https://arxiv.org/abs/2309.12871) (2023)
11. Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards general text embeddings with multi-stage contrastive learning. [arXiv:2308.03281](https://arxiv.org/abs/2308.03281) (2023)
12. Melo, R., Santos, P.A., Dias, J.: A semantic search system for the Supremo Tribunal de Justiça. In: Progress in Artificial Intelligence, pp. 142–154 (2023)
13. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: MTEB: massive text embedding benchmark. [arXiv:2210.07316](https://arxiv.org/abs/2210.07316) (2022)
14. Osório, T., et al.: PORTULAN ExtraGLUE datasets and models: kick-starting a benchmark for the neural processing of Portuguese. In: BUCC Workshop (2024)
15. Real, L., Fonseca, E., Gonçalo Oliveira, H.: The ASSIN 2 shared task: a quick overview. In: Proceedings of the 14th PROPOR, pp. 406–412 (2020)
16. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of EMNLP-IJCNLP, pp. 3982–3992 (2019)
17. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of EMNLP, pp. 4512–4525 (2020)
18. Rodrigues, J., Gomes, L., Silva, J., et al.: Advancing neural encoding of Portuguese with Transformer Albertina PT-*. In: Proceedings of EPIA (2023)
19. Santos, R., Rodrigues, J., Gomes, L., et al.: Fostering the ecosystem of open neural encoders for Portuguese with Albertina PT* family. [arXiv:2403.01897](https://arxiv.org/abs/2403.01897) (2024)
20. Santos, R., Silva, J., et al.: Advancing generative AI for Portuguese with open decoder Gervásio PT*. In: SIGUL workshop (2024)
21. Solatorio, A.V.: GISTEmbed: guided in-sample selection of training negatives for text embedding fine-tuning. [arXiv preprint arXiv:2402.16829](https://arxiv.org/abs/2402.16829) (2024)
22. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Intelligent Systems, pp. 403–417 (2020)
23. STSB Multi MT. https://huggingface.co/datasets/PhilipMay/stsb_multi_mt
24. Su, J.: CoSENT: a more effective sentence vector scheme than Sentence BERT. <https://kexue.fm/archives/8847>
25. Tatoeba. <https://tatoeba.org/>

26. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of LREC, pp. 2214–2218 (2012)
27. Vaswani, A., et al.: Attention is all you need. *NeurIPS* **30** (2017)
28. 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 (2018)
29. Wang, K., Reimers, N., Gurevych, I.: TSDAE: using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. In: Findings of the Association for Computational Linguistics: EMNLP, pp. 671–688 (2021)
30. Zhong, Q., et al.: Toward efficient language model pretraining and downstream adaptation via self-evolution: a case study on SuperGLUE. [arXiv:2212.01853](https://arxiv.org/abs/2212.01853) (2022)