Semina
Ciências Exatas e Tecnológicas

e-ISSN 1679-0375
ENGINEERINGS

ORIGINAL ARTICLE DOI 10.5433/1679-0375.2024.v45.50630

BERTugues: A Novel BERT Transformer Model Pre-trained for Brazilian Portuguese

BERTugues: Um Modelo Transformer BERT Inovador Pré-treinado para o Português Brasileiro

Ricardo Mazza Zago[1], Luciane Agnoletti dos Santos Pedotti[2]

Received: 17 May 2024 *Received in revised form: 30 November 2024* *Accepted: 2 December* *Available online: 20 December*

ABSTRACT

Large Language Models (LLMs) are trained for English or multilingual versions, with superior performance in English. This disparity occurs because, in the training of multilingual models, only a relatively small amount of data is added for each additional language. Consequently, while these models can function in Portuguese, their performance is suboptimal. The first BERT model (Bidirectional Encoder Representations from Transformers) specifically trained for Brazilian Portuguese was BERTimbau in 2020, which enhanced performance across various text-related tasks. We followed the training approach of BERT/BERTimbau for BERTugues, while implementing several improvements. These included removing rarely used characters in Portuguese from the tokenizer, such as oriental characters, resulting in the addition of over 7,000 new tokens. As a result, the average length of sentence representations was reduced from 3.8 words with more than one token to 3.0, which positively impacted embedding performance by improving metrics relevant to classification problems. Two additional enhancements involved embedding emojis as tokens – an essential step for capturing conversational nuances – and filtering low-quality texts from the training dataset. These modifications improved performance across various tasks, raising the average F1 score from 64.8 % in BERTimbau to 67.9 % in BERTugues.

keywords large language models, LLMs, BERT, NLP, foundation models

RESUMO

Grandes modelos de texto, ou LLMs, geralmente são treinados para o inglês ou versões *multilíngues*, cuja performance em inglês é superior. Isto ocorre, pois no treinamento das versões *multilíngues*, apenas uma quantidade relativamente pequena de dados de cada idioma é adicionada. Desta forma, mesmo que funcionem com o português, a eficiência é prejudicada. O primeiro modelo BERT (*Bidirectional Encoder Representations from Transformers*) treinado especialmente para o português brasileiro foi o BERTimbau em 2020, que elevou a performance em diversas tarefas de texto. Com o BERTugues, seguimos a abordagem de treinamento do BERT/Bertimbau e realizamos algumas melhorias. As alterações foram a remoção de caracteres poucos utilizados no português do tokenizador, como caracteres orientais, ganhando mais de 7.000 tokens, diminuindo o tamanho médio da representação de frases, de 3,8 palavras com mais de um token para 3,0, o que está relacionado ao desempenho dos *embeddings*, melhorando métricas relevantes para problemas de classificação. Duas melhorias adicionais envolveram a inclusão de *emojis* como tokens, o que é essencial para capturar as nuances das conversas, e a filtragem de textos de baixa qualidade dentro do conjunto de treinamento. Essas mudanças melhoraram o desempenho em várias tarefas, com uma média de F1 de 64,8% no BERTimbau para 67,9%.

palavras-chave modelos de linguagem de grande porte, LLMs, BERT, PNL, modelos fundamentais

[1]School of Electrical and Computer Engineering, Unicamp, Campinas, SP, Brazil. ricardoz@outlook.com
[2]Prof. Dr., Academic Department of Electronics, UFTPR, Curitiba, PR, Brazil. lucianesantos@utfpr.edu.br

Introduction

Artificial Intelligence (AI) is the word of the moment, with both corporate entities and governmental bodies increasingly channeling resources into the development of AI technologies. A notable illustration of this trend is GPT-4 (OpenAI, 2023), a prominent example of a Large Language Model (LLM). The burgeoning interest in AI has led to the introduction of several competing models, such as Google's Gemini and Anthropic's Claude. In both academic research and industrial applications, there is a substantial and growing demand for such models. Organizations and government agencies are particularly invested in models like ChatGPT due to their capabilities in automating customer service operations, conducting sophisticated data analysis, and enhancing operational efficiency. These advancements contribute to cost reduction and enable more rapid, personalized service delivery.

In 2017, Google introduced the transformer architecture (Vaswani et al., 2017), which contains the encoder and decoder components. Models like BERT (Bidirectional Encoder Representations from Transformers) use the encoder, GPT uses the decoder, and there are models like T5 that use both (Devlin et al., 2018). The function of a decoder is more intuitive to comprehend. ChatGPT employs this type of component, with the primary objective being the prediction of the next token, which could be a word or a subword fragment. These models are vast in scale. While earlier versions were relatively modest, containing merely millions of parameters, contemporary models have grown exponentially. The smallest models now consist of approximately 2 billion parameters (Google Gemma or Microsoft Phi3), and the most advanced ones range between 400 billion, like Meta Llama 3.1 (Llama Team, 2024), to over 1 trillion parameters (GPT 4). The training process for such models is exceedingly costly, with the required hardware infrastructure estimated to exceed 1 billion dollars. In Llama 3.1, Meta used 16k Nvidia H100 (Llama Team, 2024), which each costs about US$ 40,000.00, just in GPUs we estimate the data center costs to about US$ 640 millions.

The encoder's primary objective is to convert a word or phrase into an embedding vector that encapsulates the semantic essence of the given phrase. These vector representations enable the encoding of meaning, such that phrases with similar contexts are positioned in closer proximity within the vector space. These embeddings serve as foundational elements for various machine learning tasks, including, but not limited to, sentence similarity assessment, classification tasks, sentiment analysis, and content moderation. The number of parameters varies between 30 million to 1 billion.

As an example, let's consider the task of content moderation on a website. In this task, we must classify a user's post as toxic or not. Using a decoder-based model, we would need to write a series of instructions, known as a prompt, which the model would follow to analyze the text and perform the classification. In an encoder-based model, as proposed in this work, it is necessary to create a dataset with texts labeled as toxic or non-toxic. Each text is then transformed into a vector representation, and this representation is used in a classification model.

Both approaches have their advantages and disadvantages. The generalization capability of models like ChatGPT allows for the rapid development of solutions. However, this comes with the need for more computational power to process the model, which can involve over 1,000 times more parameters compared to BERT-based models.

Models can be trained in one or several languages. The original BERT was trained just in English (Devlin et al., 2018), and later, a multilingual version, mBERT, was released. However, training specialized versions in one language has superior performance, as demonstrated by BERTimbau (Souza et al., 2020).

In BERTimbau, we identified potential areas for improvement, such as excluding tokens that contain characters rarely used in Portuguese. Notably, a significant portion of the Bertimbau's vocabulary—approximately one-quarter of the tokens—are unlikely to be utilized in Portuguese texts. For instance, tokens like 国 are included in the vocabulary. Additionally, we propose incorporating emojis and implementing filters to remove insufficient quality texts from the BrWac corpus (Wagner et al., 2018).

The training and use of models like BERT involve two steps. The first is pre-training, which requires a large amount of text in the order of billions of words and takes days or months, depending on the capacity of GPUs/TPUs. In this stage, the model learns about the structure of the language. The second is fine-tuning, where the model, which already knows the language, is then trained for its final use, for example user's content moderation. The fine-tunning training can be done in a few minutes or hours.

In this study, we aim to pre-train a language model in Brazilian Portuguese that is applicable across a broad range of tasks, rather than being limited to specific applications. Our approach addresses the methodological limitations identified in the Bertimbau model. By implementing these enhancements, we achieved a notable improvement in performance, as evidenced by an increase in the average F1 score across five benchmarks. These benchmarks include movie review classification, three juridical similarity tasks, and an independent readability benchmark, with the mean score rising from 64.8% to 67.9%.

The article is divided into the following sections: in the Literature review, we provide a bibliographic review of previous works and techniques in the area of training and the use of text embedding models. In the Materials and methods section, the steps taken to train the BERTugues are explained. In the Results, we compare the performance of the Tokenizer and the model, comparing the embeddings in a series of tasks. Finally, in the Conclusions section, we summarize the findings and discuss future steps.

Literature review

In the development of models for use with texts in various classification tasks, such as sentiments of product reviews, it is necessary to transform the texts into numbers for then the use of models. Various techniques have been proposed over the years. A traditional one would be TF-IDF (Term Frequency-Inverse Document Frequency), which uses the relative word count.

More advanced techniques were created, such as Word2Vec, in 2013 by researchers at Google (Mikolov et al., 2013), which was the first large-dimension representation trained on billions of words. It became famous for enabling mathematical operations; for example, Queen - Woman + Man has a very close representation of King. An unresolved issue was the embeddings' lack of contextual differentiation. For instance, the word "well", which can signify either a positive condition or a water source, was represented by the same embedding across both interpretations.

In 2017, with the paper "Attention Is All You Need" (Vaswani et al., 2017), researchers from Google introduced the Transformer architecture. In 2018, they introduced the BERT model (Devlin et al., 2018), which enabled the generation of contextual embeddings for English. The following year, a version supporting 104 languages, including Portuguese, which we will call mBERT, was released.

BERTimbau was the first BERT-based model pre-trained for Portuguese (Souza et al., 2020, 2023); it followed the same steps as the original paper's training, using only texts in Portuguese with the BrWac dataset (Wagner et al., 2018). It demonstrated that models specialized in just one language can perform much better than those trained in several languages. De Souza (2020) observed that the quality metrics of BERTimbau improved more significantly compared to mBERT when the model's tokenization resulted in fewer tokens. Consequently, Some opportunities for improvement in the model were found. The main one is the presence of a large number of oriental characters in continuation tokens, such as 国, 因, 聽, and 徐 which are unlikely to be used in Portuguese.

To the best of our knowledge, Bertimbau is the only pre-trained BERT class, from 110 to 330 million parameters, model specifically developed for Brazilian Portuguese with general applicability. Other models, which will be discussed next, are specialized for specific domains and have been trained within limited contextual frameworks.

The JurisBERT model (Viegas, 2022), developed specifically for Semantic Textual Similarity (STS) tasks within the Brazilian legal domain, was trained from scratch using only legal documents and court rulings from Brazilian courts. This specialized training allowed JurisBERT to outperform BERTimbau when applied to legal texts.

The LegalBERT-pt model (Silveira et al., 2023), akin to JurisBERT, was specifically developed for the legal domain. It was trained on documents from ten Brazilian courts, encompassing four types of legal texts: initial petitions, petitions, decisions, and sentences. The model is designed to perform two key tasks: Named Entity Recognition (NER) and text classification.

The Sabiá model (Pires et al., 2023) is a decoder model, meaning it is an instruction-following model, developed by the Brazilian company Maritaca. It originated from earlier versions of Meta's Llama and continued training with Portuguese texts. To the best of our knowledge, it is the first GPT-like model specifically trained for Brazilian Portuguese, much like Bertimbau was for BERT. This additional training has improved its performance in Portuguese language tasks.

In general, datasets used for training models are often constructed through web scraping from internet sites. However, this process offers limited control over the nature of the extracted texts. For example, when performing web scraping on a news website, one might extract both full articles and lists of topics from the site, such as “Business”, “Arts”, “Lifestyle”, among others. Additionally, there is the possibility that the words in the extracted text may appear concatenated, without proper spaces between them (e.g., “wordsmergedtogether”).

In our research, we explored various techniques aimed at addressing issues related to the removal of insufficient quality texts. DeepMind, in its development of the Gopher model (Rae et al., 2021), identified that a significant portion of the textual data available on the internet is unsuitable for model training. To mitigate this, they proposed a series of filtering methods to clean the text corpus, including the removal of texts that exhibit the following characteristics:

- Has less than 50 words or more than 100,000 words;
- The average size of the words in text is less than 3 or more than 10 characters;
- The ratio between the number of symbols and words is greater than 10% for the hashtag and three dots;
- More than 90% of the lines start with bullet points or more than 30% end with three dots;
- Less than 80% of the words have alphanumeric characters;
- Do not have at least 2 common words in English (the, be, to, of, and, that, have, with).

Each criterion is designed to address specific issues within the text. For instance, restricting the average word length from 3 to 10 character range helps to prevent the problem of splited (s p l i t e d) or concatenated words. The presence of lines starting with bullet points is a typical characteristic of content found on web-scraped spam sites. Common word filters are employed to exclude texts in languages other than English.

Materials and methods

BERTugues

For the pre-training of BERTugues, we followed the methodology previously employed in BERTimbau, with several enhancements based on a thorough review of the literature. The most significant improvement was the removal of non-Latin characters, particularly those from oriental scripts, which are rarely or never used in Portuguese. Additionally, we incorporated emojis into the token vocabulary and performed a quality filtering based on Gopher paper (Rae et al., 2021).

As the training codes for BERTimbau are unavailable, we rewrote it following the description in de Souza (2020). We used the Huggingface Transformers (Wolf et al., 2020), for transformer implementation, Tokenizers (Zucker, 2024), to tokenize the text, and Datasets (Lhoest et al., 2021) libraries, to load and work with the dataset. The datasets used to pretrain were: Portuguese Wikipedia (“Wikimedia”, 2024) for the tokenizer and BrWac for pre-training the model (Wagner et al., 2018).

Tokenizer

The Tokenizers library creates the model’s vocabulary; common words usually appear whole in the vocabulary, while uncommon ones may be broken into more than one token. An example of this break could be the word “conversaram”, which could be broken into the tokens “conversa” and ‘##ram”, where the “##” denotes that it is the continuation of a word. 99 slots for new tokens, [unusedxx], where xx represents the token number, were added to be used in some task that requires new tokens. Five special tokens were also added, which denoted:

- [CLS]: the beginning of the sentence;
- [SEP]: the separation between sentences and the end of a sentence;
- [PAD]: space not used in the sentence size for the batch;
- [MASK]: token to be predicted by the model;
- [UNK]: unknown, a word that cannot be tokenized.

For the tokenizer's training, a copy of Wikipedia in Portuguese was used. Initially, we started using a sample of BrWac for training the tokenizer, which resulted in non-representative tokens of the language with many punctuation characters in sequence. Thus, we used Wikipedia and did a pre-treatment, removing oriental characters of little use in Portuguese. At the end of the process, our tokenizer yields a total of 30,522 tokens, which matches the number of tokens found in the original English BERT model.

BrWac Cleaning

The cleaning of BrWac consisted of two stages. As in BERTimbau, the first was to remove HTML tags and treat mojibakes with the ftfy (Speer, 2019) and BeautifulSoup libraries ("Beautiful", 2023).

Then, the Gopher (Rae et al., 2021) filters with small changes were applied. Allowing sentences with at least 3 (three) and a maximum of 200,000 words with at least one common word in Portuguese contrasts the two necessary in Gopher. Our goal was to allow shorter sentences, as we are also interested in this use. In total, 3% (110,051) of the texts from the dataset were removed.

Pre-training

For pre-training, we followed the two objectives of BERT: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). These objectives are defined in the last layers of the model, which are called heads, with the goal of the transformer model learning the structure of the language. They can be exchanged after pre-training for a task-specific head in a process known as transfer learning. The embeddings generated in the encoder of the transformer can be used directly for training classification models in a supervised process without further training of the language model, if necessary.

In MLM, the goal is to predict a masked word in the text, using the [MASK] token. The NSP is to predict whether the next paragraph following a [CLS] token is a continuation of the text or a paragraph from another text. An example of a sentence that could be used in training, assuming that each word is a model token, that is, that no word was broken into a subtoken with the beginning "##":

```
[CLS] Eduardo abriu os [MASK], mas não
quis se levantar. [SEP] Ficou deitado e viu
que [MASK] eram. [SEP]
```

The sentence always starts with [CLS], and separating the two sentences there is the token [SEP] at the end as well. Half of the time, the sentence after the first [SEP] is the continuation of the first, and the other half is a random sentence chosen from the other texts. The model learns to differentiate whether a sentence is a continuation or not of the previous one. Regarding the MLM task, 15% of normal word tokens undergo one of 3 options; in 80% of cases, they are replaced by the [MASK] token, in 10% by a random word, and the remainder of the word is kept. The model must then discover what word should be in place of the [MASK] token.

This pre-training was performed using an Nvidia GPU, with fp16 floating point format. Like BERTimbau and RoBERTa, each training epoch had different words masked, while in the original BERT, possibly due to the masking being done at the time of building the training dataset and not in the training process itself, the maskings were fixed.

The training parameters included an effective batch size of 256 sequences, with gradient accumulation employed to overcome memory limitations and achieve the desired batch size. The optimizer was Adam with a learning rate of 1e-4, β_1 = 0.9, β_1 = 0.999, L2 weight decay of 0.01, and learning warm up in the first 10,000 steps going from 0 to the learning rate.

As attention has quadratic complexity about the number of tokens in the sentence, it is usual to train the model with fewer tokens and increase that number at the end of training. The model was trained with a maximum length of 128 tokens per sentence until iteration 820,000, representing 5.7 epochs, and the remaining 180,000, with 512, representing 3.4 epochs.

Results

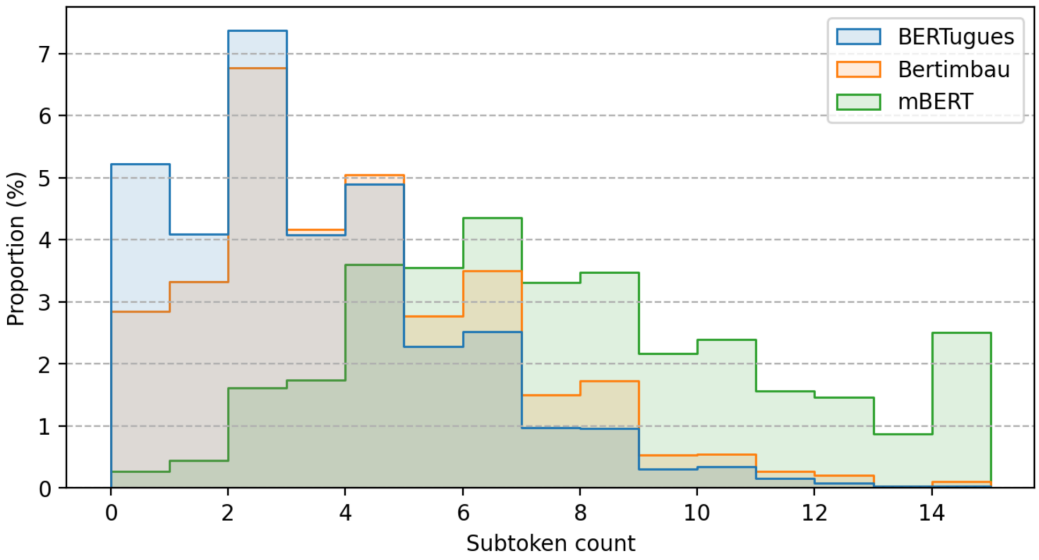
We divided the performance comparison of the models into two distinct parts. The first part focuses on analyzing the average length of tokenized sentences produced by the tokenizers of leading Portuguese and multilingual models. This analysis is rooted in the hypothesis proposed by de Souza (2020), which posits that a more efficient tokenizer, i.e., one that generates fewer tokens per sentence, tends to improve model performance across a variety of use cases. The second part entails evaluating the embeddings generated by different models on multiple tasks, some of which have already been explored in the literature, by measuring the performance metrics associated with each task.

Tokenizer Performance comparison

The comparison between the performance of the tokenizer was carried out using the ASSIN 2 dataset (Real et al., 2020), the same one used in the comparison carried out in the master's dissertation that gave rise to BERTimbau (de Souza, 2020). It provides sentence pairs consisting of a premise and a hypothesis, along with similarity scores ranging from 1 to 5. The sentence pairs were concatenated, and the number of subtokens, identified by the prefix "##", were counted.

In Figure 1, we demonstrate a comparison between the tokenizers of mBERT, BERTimbau, and BERTugues. We verified that the average number of words broken into more than one token of mBERT was 7.4, in BERTimbau 3.8 and BERTugues 3.0, a reduction of 21% in relation to BERTimbau, demonstrating the better efficiency of the tokenizer of BERTugues.

Figure 1 - Number of subtokens per sentence of ASSIN 2



Model Performance comparison

For the performance comparison between models, we selected four tasks, along with an additional benchmark test from the literature, incorporating BERTugues.

The first task involved a text classification problem using the IMDB movie reviews dataset (Fred, 2019), which contains movie reviews from the IMDB website (IMDb, 2024), classified as either positive or negative, representing a binary classification problem. In this task, we employed the BERTugues representation of the sentence and fed it into a Random Forest model from the scikit-learn library (Pedregosa et al., 2011) to perform the classification. A Random Forest is an ensemble algorithm that combines multiple decision trees to improve accuracy and reduce overfitting. It uses majority voting for classification and was chosen due to its generally good performance with default parameters.

For tasks 2 through 4, we utilized the benchmarks already published in the JurisBERT model paper (Viegas, 2022), which compares the performance of mBERT and BERTimbau with models fine-tuned or trained on legal texts for sentence similarity tasks using cosine similarity and the F1 metric.

We reused the code (Viegas & Alfaneo, 2023), incorporating BERTugues into the evaluation. This task employs Brazilian legal texts from the STJ, PJERJ, and TJMS to assess sentence similarity, where the dataset includes sentence pairs with a binary classification indicating whether they are similar or not.

Following the online release of BERTugues and prior to the publication of this article, Ribeiro (Ribeiro et al., 2024) applied the model to the task of evaluating text readability. In this study, BERTugues demonstrated superior performance compared to BERTimbau-large, achieving higher evaluation metrics and further showcasing the model's robustness. The specific task involved classifying the readability levels of texts sourced from Portuguese language exams conducted by Camões, I.P., the official institute responsible for promoting the Portuguese language. These texts were categorized across five distinct proficiency levels, and the evaluation was carried out using the Macro F1 metric.

The F1 score is a widely used performance metric in classification models that combines precision and recall into a single value, providing a balanced evaluation of a model's accuracy, particularly in situations where class imbalance is present. It is defined as the harmonic mean of precision, which measures the proportion of true positive predictions among all positive predictions made by the model, and recall, which evaluates the model's ability to correctly identify all actual positive instances. A generalized version of the F1 score, known as Macro F1, extends this concept to multi-class classification problems by calculating the F1 score independently for each class and then averaging the results, treating all classes equally regardless of their frequency.

In text-related tasks utilizing transformer models, fine-tuning the entire model is generally considered best practice, rather than solely training the classification layer on top of the embeddings. However, this process is computationally expensive and typically requires specialized hardware such as GPUs or TPUs. In contrast, using the model without fine-tuning, simply extracting the embeddings, offers a more efficient alternative. Although this extraction remains relatively time-consuming, it can be performed on a CPU, making it a more accessible approach for environments with limited hardware resources. Ensuring that our model performs well in this context is a key objective.

In Table 1, a comparison of the performance of the models is presented: mBERT, BERTimbau, and BERTugues. In 4 out of 5 tasks, BERTugues was superior to BERTimbau base, and in 3 out of 5, it was also superior to BERTimbau large, a much larger and computationally expensive model. The average performance was approximately 3.1 p.p. higher than the BERTimbau base. In all cases, it was superior to mBERT, as expected.

Table 1 - Performance of the models in various tasks.

Model	IMDB (F1)	STJ (F1)	PJERJ (F1)	TJMS (F1)	Readability (Macro F1)(Ribeiro et al., 2024)	Average (F1)
Multilingual BERT	72.0%	30.4%	63.8%	65.0%	N/A	57.8%
BERTimbau-Base	82.2%	35.6%	63.9%	71.2%	71.3%	64.8%
BERTimbau-Large	**85.3%**	43.0%	63.8%	**74.0%**	71.7%	67.6%
BERTugues-Base	84.0%	**45.2%**	**67.5%**	70.0%	**72.8%**	**67.9%**

Conclusion

In this work, we presented a new BERT class model pre-trained for Brazilian Portuguese and made the model publicly available (Zago, 2023). For the best of our knowledge, this is the best BERT class model available for Brazilian Portuguese.

We trained the model using a new methodology, a process that took several months. We compared the performance in tasks previously made available in the literature, raising mean performance about the state of the art of open text models with less than 120 million parameters by 3.1 p.p.. This enhancement is attributed primarily to the superior tokenizer employed in our approach.

We introduced improvements to the training methodology compared to the state of the art model, BERTimbau (Souza et al., 2020), such as the removal of characters that are infrequently used in Brazilian Portuguese. This modification enabled the tokenizer to perform more efficiently, as it required fewer tokens to represent sentences, resulting in an estimated 21 % reduction in token usage.

This reduction is crucial in enhancing the model's performance compared to Bertimbau by enabling the representation of the same sentence with fewer tokens.

In his master's thesis, which led to the development of Bertimbau, de Souza (2020) noted that Bertimbau exhibited performance comparable to that of Multilingual BERT when the number of tokens of the phrase was similar, and it outperformed Multilingual BERT when sentences were tokenized into fewer tokens. This suggests that by decreasing the average number of tokens per sentence, the model's performance can be improved. Furthermore, this reduction allows for the processing of longer sentences that would otherwise exceed the model's 512-token limit.

Additionally, we expanded the model's vocabulary to include emojis, which are particularly relevant in sentiment analysis tasks. Furthermore, in alignment with a methodology proposed by DeepMind (Rae et al., 2021), we applied a filtering process to the BrWac dataset, eliminating low-quality texts. This filtering resulted in a reduction of approximately 3 % of the total text corpus.

In the following steps, we plan to train the large version of BERT, which has approximately three times the number of parameters. For this, greater computational capacity will be necessary. Furthermore, train models of the BERT family with more recent architectures, with RoBERTa and DeBERTa for Portuguese, which will have superior performance to BERTugues, but again require greater hardware capacity for training or will have a longer training time than BERTugues.

Author contributions

R.M. Zago contributed to this work in the following capacities: conceptualization, data curation, formal analysis, investigation, methodology, original draft writing, validation, and visualization. L.A. dos S. Pedotti contributed to this work through critical revision and thorough editing.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Beautiful Soup Documentation. (2023). https://www.crummy.com/software/BeautifulSoup/bs4/doc/#module-bs4

de Souza, F. C. (2020). *BERTimbau: modelos BERT pré-treinados para português brasileiro*. [Master's Thesis, Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação]. Repositório. https://repositorio.unicamp.br/Busca/Download?codigoArquivo=466423&tipoMidia=0

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805

Fred, L. (2019). IMDB PT-BR. https://www.kaggle.com/datasets/luisfredgs/imdb-ptbr

IMDb. (2024). Imdb service. https://www.imdb.com/pt/

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., … Wolf, T. (2021). Datasets: Uma biblioteca comunitária para processamento de linguagem natural. In H. Adel & S. Shi (Eds.), *Anais da conferência de 2021 sobre métodos empíricos em processamento de linguagem natural: Demonstrações de sistemas* (pp. 175–184). Associação para Linguística Computacional. https://doi.org/10.18653/v1/2021.emnlp-demo.21

Llama Team. (2024). The Llama 3 Herd of Models. *ArXiv*, *3*, 1–92. https://doi.org/10.48550/arXiv.2407.21783

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Arxiv*, *3*, 1-12. https://doi.org/10.48550/arXiv.1301.3781

OpenAI. (2023). Gpt-4 technical report. https://doi.org/10.48550/ARXIV.2303.08774

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pires, R., Abonizio, H., Almeida, T. S., & Nogueira, R. (2023). Sabiá: Portuguese Large Language Models. In Sociedade Brasileira de Computação, *Anais da Brazilian Conference on Intelligent Systems* [Anais]. 13º Brazilian Conference on Intelligent Systems, Porto Alegre, Brasil. https://sol.sbc.org.br/index.php/bracis/article/view/28417

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G. v. d., Hendricks, L. A., Rauh, M., Huang, P.-S., … Irving, G. (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *Arxiv*, *2*, 1-120. https://doi.org/10.48550/ARXIV.2112.11446

Real, L., Fonseca, E., & Oliveira, H. G. (2020). The ASSIN 2 Shared Task: A Quick Overview. In P. Quaresma, R. Vieira, S. Aluísio, H. Moniz, F. Batista, & T. Gonçalves (Eds.), *Computational Processing of the Portuguese Language* (pp. 406–412). Springer International Publishing. https://doi.org/10.1007/978-3-030-41505-1_39

Ribeiro, E., Mamede, N., & Baptista, J. (2024, March). Automatic text readability assessment in European Portuguese. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, & R. Amaro (Eds.), *Proceedings of the 16th international conference on computational processing of portuguese* (pp. 97–107). Association for Computational Lingustics. https://aclanthology.org/2024.propor-1.10

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., & Furtado, V. (2023). LegalBert-pt: Um modelo de linguagem pré-treinado para o domínio jurídico do português brasileiro. In M. C. Naldi & R. A. C. Bianchi (Eds.), *Intelligent systems. bracis 2023. lecture notes in computer science* (Vol. 14197). Springer, Cham. https://doi.org/10.1007/978-3-031-45392-2_18

Souza, F. C., Nogueira, R. F., & Lotufo, R. A. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In R. Cerri, & R. C. Prati (Eds,), *Intelligent Systems* (pp. 403-417). Springer, Cham. https://doi.org/10.1007/978-3-030-61377-8_28

Souza, F. C., Nogueira, R. F., & Lotufo, R. A. (2023). Bert models for brazilian portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing*, *149*, 110901. https://doi.org/https://doi.org/10.1016/j.asoc.2023.110901

Speer, R. (2019). ftfy: Fixes Text for You, Version 5.5. https://doi.org/10.5281/zenodo.2591652

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Arxiv*, *5*, 1-15. https://doi.org/10.48550/ARXIV.1706.03762

Viegas, C. F. O. (2022). *JurisBERT: Transformer-based model for embedding legal texts.* [Master's Thesis, Universidade Federal de Mato Grosso do Sul]. Repositório. https://repositorio.ufms.br/handle/123456789/5119

Viegas, C. F. O., & Alfaneo. (2023). Brazilian-legal-text-benchmark. https://github.com/alfaneo-ai/brazilian-legal-text-benchmark

Wagner, J. A., Filho, Wilkens, R., Idiart, M., & Villavicencio, A. (2018). O Corpus brWaC: Um novo recurso aberto para o Português Brasileiro. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S.

Piperidis, & T. Tokunaga (Eds.), *Anais da décima primeira conferência internacional sobre recursos linguísticos e avaliação (LREC 2018)*. Associação Europeia de Recursos Linguísticos (ELRA). https://aclanthology.org/L18-1686

Wikimedia Downloads. (2024). https://dumps.wikimedia.org/backup-index.html

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Q. Liu, & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zago, R. (2023). Bertugues-base-portuguese-cased. https://huggingface.co/ricardoz/BERTugues-base-portuguese-cased

Zucker, A. (2024). Huggingface tokenizers. https://github.com/huggingface/tokenizers