Aroeira: A Curated Corpus for the Portuguese Language with a Large Number of Tokens

Thiago Lira, Flávio Cação, Cinthia Souza, João Valentini, Edson Bollis, Otavio Oliveira, Renato Almeida, Marcio Magalhães, Katia Poloni, Andre Oliveira, and Lucas Pellicer

Instituto de Ciência e Tecnologia Itaú-Unibanco (ICTi), São Paulo, Brazil
{thlira15, flavio.nakasato}@gmail.com, {cinthia.mikaela-souza, joao.valentini22, edson.bollis, otavio.rodrigues-oliveira, renato-augusto.almeida, marcio.chiara-magalhaes, katia.poloni, andre.seidel-oliveira, lucas.pellicer}@itau-unibanco.com.br

Abstract. The emphasis on constructing extensive datasets for training large language models (LLM) has recently increased, and current literature predominantly features datasets for high-resource languages such as English and Chinese. However, there is a notable scarcity of high-quality corpora for the Portuguese language. To address this limitation, we propose Aroeira, a curated corpus explicitly designed for training large language models in the Portuguese language, with a focus on the Brazilian Portuguese one. The Aroeira Corpus consists of 100 GB of texts from various internet platforms, processed through a comprehensive pipeline to ensure superior quality. The pipeline handles downloading, text extraction, language identification, application of quality and bias filters, and storage, all tailored for the Portuguese language. The resulting corpus contains 35.3 million documents and over 15.1 billion tokens, surpassing the largest previously available corpus in this domain.

1 Introduction

Most modern natural language processing (NLP) methodologies, including Large Language Models (LLMs), rely on extensive text corpora for precise training and weight adaptation [18]. Large-scale training corpora, or pre-training corpora, are fundamental for developing foundational models, which serve as the basis for numerous task-specific adaptations [7]. State-of-the-art LLM training pipelines utilize various types of datasets: (i) pre-training corpora to acquire language structure, syntax, and semantics; (ii) instruction fine-tuning datasets to enhance the model's capability to follow instructions; (iii) preference datasets to rank responses; and (iv) evaluation datasets to measure model performance [23].

Recent research shows that the size and diversity of pre-training corpora significantly impact LLM performance [14, 18]. Most pre-training datasets are available in English and Chinese, which are high-resource languages, while other languages have significantly fewer tokens [23]. Although multilingual corpora can help mitigate data scarcity for low-resource languages, these datasets are

often unbalanced, favoring high-resource languages [17]. This imbalance affects the performance of multilingual models for less-represented languages and models trained on multilingual corpora do not perform as well as those trained on monolingual corpora [34]. Therefore, it is essential to train or fine-tune models in the target languages to capture linguistic nuances, structures, and domain-specific or cultural knowledge [28].

A direct implication of this scenario is the necessity of making high-quality plain-text corpora available to encourage research on specific model languages and the development of better-performing approaches. Therefore, we introduce Aroeira: a curated Portuguese language-specific corpus composed of approximately 100 GB of text. The content was extracted from recent Common Crawl[1] (CC) web pages (up until 2023) and fully curated to remove web tags, ensuring quality and bias filtering. To the best of our knowledge, Aroeira is the largest highly-curated Portuguese corpus available to date. It has the potential to influence new instruction fine-tuning and evaluation dataset studies while guiding the development of preference datasets and large models.

Aroeira was created based on a double-pipeline inspired by [30]. The pipeline comprises two key steps: data quality management and content safety assurance. These steps ensure the size and quality necessary for a Portuguese corpus to train safe LLMs effectively. As part of the content safety step, we investigated techniques for filtering hazardous content and mitigating biases in our corpora [16]. This effort resulted in a custom Portuguese word dictionary, which encompasses offensive words, as well as terms, expressions, and phrases that include sexism, homophobia, ableism, racism, hate speech, and political, religious, and regional prejudice [13, 24, 26].

We highlight our main contributions:

- Introduction of Aroeira, a 100 GB Portuguese corpus from diverse internet sources. Our dataset surpasses the largest currently available corpus for training language models in Portuguese in terms of size, quality, and representativeness.
- Development of a parameterizable double-pipeline, which includes: downloading, extracting, language identification, quality filtering, and text storage in the data step; filtering sexual content, toxic data, and bias in the content safety step.
- Creation of a dictionary to filter biased terms and mitigate social bias in the Portuguese language.

This paper is organized as follows. In Section 2, we present related work in corpus extraction. Sections 3 and 4, we describe the methodology for generating the corpus and the configuration of hyperparameters used in the quality filters, respectively. In Section 5, we analyze the volumetry of Aroeira in terms of year distribution, knowledge domains, document length, and other relevant results. Finally, Section 6 presents conclusions and future works.

[1] Available at: https://commoncrawl.org/

2 Related Work

The largest Portuguese language corpus is BrWac [35] which has approximately 25 GB of textual data distributed in 3.53 Mi documents totaling 2.68 Bi tokens. Another large corpus is the Carolina 1.2 Ada [11], which contains approximately 2.11 Mi documents and a total of 11 GB of textual data. When we compare this corpus with the corpora of other languages, the gaps become evident. Gao et al [14], for example, propose The Pile, a corpus with 825 GB of texts in English. The corpus is derived from various data sources, including scientific articles, patent documents, and forums.

An inspiring work for Aroeira is Colossal Clean Crawled Corpus (C4) [30], a curated English-only corpus. C4 was created using Common Crawl (CC) data extracted in April 2019 and comprises approximately 750 GB of clean English text. Similar to our approach, they apply filters to the raw data. CLUECorpus2020 [36] was constructed using cleaned data from CC, resulting in a high-quality Chinese pre-training corpus of 100 GB and 36 Bi tokens. MassiveText [29] is a collection of large English datasets created with data from different sources. MassiveText contains 2.35 Bi documents, equivalent to 10.5 TB of text. More recently, Sabiá [28] applied a similar filtering methodology of MassiveText to the Portuguese section of ClueWeb dataset [27] and managed to retrieve a curated dataset. WuDaoCorpora [38] is a 3 TB Chinese corpus with 1.08 Tri of Hanzi characters collected from 822 Mi web pages.

It is also worth mentioning that a current trend is the proposal of multilingual corpora. The C4Corpus authors [17] present the construction of a 12 Mi web page corpus containing more than 50 languages, including Portuguese. English has a volume of 7.7 Mi (64.2%) documents while Portuguese has only 0.3 Mi (2.5%). RedPajama [10] is a large multilingual corpus, containing 100 Bi text documents extracted from 84 CC snapshots. Quality signals were applied to 30 billion documents, and deduplication was performed on 20 billion documents. It claims to have English (69.8%), Deutch (9.2%), Spanish (8.8%), French (7.8%), and Italian (4.4%).

As we can see, the corpora available in English and Chinese have a massive data amount, easily surpassing corpora in Portuguese and other languages. However, we know that the amount of internet information available in English and Chinese is greater than in Portuguese.

Another important aspect is the biases present in corpora and texts. Language is a highly relevant avenue for manifesting social hierarchies, pre-established concepts, and standard forms of treatment [6]. Various efforts are being made to evaluate data biases and how they impact the behavior of language models. The paper [24] analyzed 93 social groups that receive stigmatized treatment by NLP models. Work [25] created StereoSet to measure stereotypical treatment in certain ethnic groups, and paper [26] developed a benchmark dataset for measuring biases related to gender, race, age, sexual orientation, and others.

These aspects are relevant in a context with strong normative motivations and the need to create responsible AI. Many ways exist to mitigate text biases, such as data augmentation, content filtering, rebalancing, masking, and many

others [13]. Our work uses the concept studied by [16], where filtering sensitive content can result in models with more equitable treatment of different ethnic groups. The work specifically uses word co-occurrence in the filtering process.

Based on these past works, we can see that, in general, they focus on creating corpora for training language models for high-resource language tasks. Thus, there is a necessity for creating a Portuguese corpus since the amount of large Brazilian Portuguese models has drastically increased recently. We can cite Bertimbau [33], PTT5 [9], Bertaú [12], Sabiá [28, 4], Cabrita [22], and Bode [15].

3 Aroeira

In this section, we detail the steps of the corpus creation (double-pipeline) which is divided into two objectives, (i) collect (Data Pipeline) and (ii) ensure content safety (Content Safety Pipeline). Our whole pipeline contains nine steps: data collection and sampling, text extraction, language identification, deduplication, and quality filters in Data Pipeline, and sexual content filter, toxic data filter, bias filter, and categorization in Content Safety Pipeline. Figure 1 presents the entire workflow.

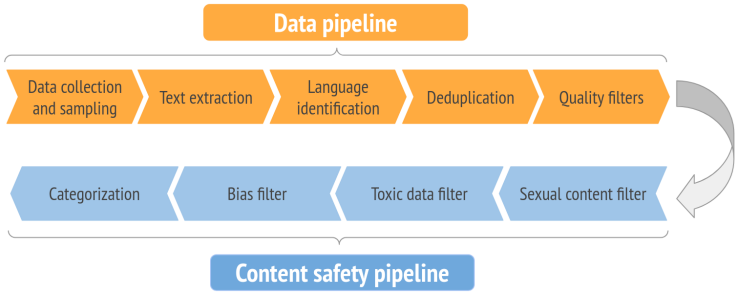

Fig. 1. Double pipeline: Data Pipeline contains collection and sampling, text extraction, language identification, deduplication, and quality filters; and content safety pipeline encompasses sexual content filter, toxic data filter, bias filter, and categorization.

3.1 Data collection and sampling

The data collection and sampling step involves downloading and extracting Portuguese text from raw Web ARChive (WARC) in files. All data is sourced from Common Crawl (CC), which contains petabytes of scraped internet content from millions of web pages. We use the raw HTTP files as the initial material, from which we extract and filter Portuguese text as detailed in Subsections 3.2 and 3.3. CC organizes its datasets by date, each comprising thousands of individual shards of scraped content. We sampled shards from datasets ranging from 2015 to 2023, prioritizing more recent data.

3.2 Text extraction

This computational step involves using multiple cloud machines in parallel. These instances download and process raw files by extracting text from the HTML and filtering for Portuguese text. The resulting data is then processed further using a single machine containing a key-value database for deduplication, as explained in Subsection 3.4. We opted to work with WARC files to ensure better quality text, which includes handling raw HTML files and extracting texts ourselves. A Python library called Trafilatura [5] extracted only natural language text from the HTML files. Metadata from webpages were saved for later use in the pipeline.

3.3 Language identification

Roughly 0.2% of the pages in each shard are in Portuguese. To filter these pages, it is necessary to detect the language they were written in automatically. As utilized by [14] and [1], we employ Meta AI's pre-trained fastText model, which can detect 176 languages. For each downloaded page, the text is extracted using the Trafilatura library [5] and the fastText[2] [20] models used to determine the language. Pages identified as Portuguese with the highest probability by fastText were selected.

3.4 Deduplication

The purpose of this step is to remove duplicated data from the corpus. To achieve this goal, we use two deduplication approaches. The first is a page-level approach, which identifies and removes pages with duplicate URLs. The second is the document-level approach, which aims to remove significant overlapping documents. We employ the MinHashLSH algorithm to calculate the Jaccard similarity between documents, considering whether two document similarity exceeds 0.7 [29].

3.5 Quality filters

A significant amount of data available on the internet may be insufficient in terms of quality for linguistic model formation. Some examples include automatically generated text and text not written for human consumption [29]. This step aims to retain only pages written by humans for humans. To achieve this, we applied a series of ten quality filters:

- **Number of tokens**: Removes pages with fewer than a minimum number of tokens (in this work, we used the same tokenizer employed by GPT-2), as texts with low token counts are generally not informative;
- **Number of words**: Removes pages that do not attend specified upper and lower word limits, excluding punctuation and special characters;

[2] Available at: https://fasttext.cc/

- **Type Token Ratio (TTR)**: The ratio of unique words (types) to total words (tokens) [31]. TTR [37] serves as an indicator of text quality;
- **Symbols-word ratio**: Removes pages whose symbol word percentages exceed limits. Any special character is considered a symbol;
- **Symbols at the beginning of the text**: Removes pages with an excessive number of symbols at the beginning of the text;
- **Stopwords**: The presence of stopwords may indicate text coherence [29];
- **N-gram repetition**: Excessive repetition of sentences, paragraphs, or n-grams indicates low informational content [29];
- **Number of sentences**: Removes pages with fewer than a specified number of sentences;
- **Lorem ipsum**: Removes pages containing the term "Lorem ipsum" [30];
- **Valid words**: Removes pages whose percentage of words found in a language dictionary is below a specified threshold.

The thresholds for each filter are detailed in Section 4.

3.6 Sexual content filter

To maintain the integrity of the corpus, a filter was applied to remove sexual content from the data. We verified whether a URL was present in the Université Toulouse 1[3] (UT1) blocklist for each page collected. As noted by [2], the UT1 blocklist is an extensive compilation of block lists frequently used for internet access control at schools. It was developed with the help of automated systems and human contributors and currently includes 3.7 million entries. For this work, we utilized a filtered version of this blocklist tailored for Brazilian websites. It should be noted that this filter only excludes websites marked as adult content. For the remaining content, we randomly selected 25,000 examples and used a Mistral 7B [19] model to extract pejorative sexual terms. These terms were then reviewed by humans and used as the final sexual content filter.

3.7 Toxic data filter

In this step, our objective is to identify and remove potential toxic content. Toxicity definition is a rude, disrespectful, or unreasonable comment likely to incite an argument [13]. Our filter comprises a dictionary of insults and pejorative terms. We evaluated exact matches of dictionary words with document terms and removed documents that exceeded a specified threshold percentage of words in the dictionary. The dictionary used for this filter was created by merging two lists of words[4,5]. We reinforce this filter does not aim to eliminate all the data containing toxic words, but rather to remove content with a significant toxic content proportion.

[3] https://dsi.ut-capitole.fr/blacklists/index_en.php

[4] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/pt

[5] https://github.com/dunossauro/chat-detox/blob/main/palavras.txt

3.8 Bias filter

Our work involves a step to identify and eliminate potential biases in the text based on contextual cues. We construct a dictionary of Portuguese expressions used in biased contexts.

When compiling this dictionary, it is crucial to consider the society dynamic and its relationship with linguistics [6]. We filtered the corpus by checking for exact matches between the dictionary words and those in the text [16]. Various types of social biases were mapped, including gender, religion, race, sexist expressions, xenophobia, homophobia, ableism, fatphobia, and politics [24, 26].

3.9 Categorization

This phase aims to categorize each page into a specific knowledge domain. We identified 27 knowledge domains, covering several categories and subjects, such as: blog posts, news articles, marketing, movies, social media, health, culinary recipes, books, scientific articles, politics, etc. The information regarding each domain can be leveraged to balance the data for specific tasks or augment datasets where knowledge is lacking. Essentially, we map the URLs of different pages to each knowledge scope and assign a topic to each URL. A pt-pt text category has been introduced to differentiate between Brazilian Portuguese and European Portuguese, as Brazilian Portuguese predominates in the dataset.

The identified knowledge domains and their distribution are discussed in the subsection 5.2.

4 Qualitative Configuration Test

We generated a 1 GB sample of texts and analyzed the distribution of metrics such as the number of tokens, word count, and TTR. We use this sample to find the best configuration for our double-pipeline to produce the resultant datasets. Different value sets were empirically tested, and for each one, we checked the correctness of page removals and recorded the number of pages excluded after the filters were applied.

Analysis was carried out on this sample to find the optimal configuration. Moreover, we perform qualitative analyses to verify the removal appropriateness. This qualitative test helped us define satisfactory values that we should filter to obtain content with good textual quality, i.e., a text that is diverse in words, fluid, with few repetitions, and with semantically relevant content. It is worth noting that we also evaluated the number of potentially toxic words and possible biases contained in the texts.

Table 1 shows the optimal configuration to obtain texts that exceed our minimum quality requirements.

5 Results

The created corpus was evaluated concerning five groups of requirements. The first requirement is the created corpus must be larger than the existing corpus

Table 1. Double-pipeline final configuration.

Parameter	Description	Value
min_tokens	Minimum number of tokens	30
min_words	Minimum number of words	20
max_words	Maximum number of words	10000
TTR	Type Token Ratio	0.2
max_symbols	Maximum percentage of symbols-words	0.70
fs_symbols	Maximum number of symbols at the beginning of the sentence	6
min_stopwords	Minimum percentage of stopwords	0.02
occurrence_ngram	3-gram repeat percentage	0.3
num_sentences	Minimum number of sentences	2
valid_words	Minimum percentage of valid words	0.2
toxic_content	Maximum percentage of toxic content	0.2
max_word_bias	Maximum number of biased words	10

for the Portuguese language (see Subsection 5.1). The second requirement is the corpus must be diverse, i.e., containing data from different sources (see Subsection 5.2). The third requirement is the corpus covers the most recent to the least recent information (see Subsection 5.3). The fourth requirement is that the corpus presents high-quality text indicators (see Subsection 5.4). Finally, the fifth requirement is that the corpus avoids introducing or increasing bias.

Due to the large corpus size, Subsections 5.4 and 5.5 utilize a 10% randomly generated sample to present the results. Consequently, Figure 4 and Table 3 were created based on this sample size.

5.1 Corpus size

The first requirement evaluated was the corpus size. We collected terabytes of data from different CC dumps. Each dump has approximately 0.2% of texts in Portuguese. It is worth noting that the documents may be of poor quality, contain inappropriate or biased content, and be duplicated due to the CC not filtering the data. Therefore, a corpus cleaning step was necessary to ensure that the final corpus was composed only of non-duplicated documents to respect quality criteria. At the end of this process, we obtained a corpus of 100 GB.

Table 2 presents the created corpus statistics alongside other Portuguese corpora. Aroeira surpasses brWac [35] and Carolina 1.2 Ada [11] in size, document quantity, and token number. Thus, our corpus is potentially a more diverse resource regarding texts and tokens than the available resources.

5.2 Knowledge domains

The second requirement evaluated was the distribution of knowledge domains within the created corpus. A mapping of different URLs to their respective knowledge domains was conducted. Each base URL was verified against a dictionary. When there were no matches, keywords were used to determine the

Table 2. Corpora size comparison.

Language	Corpus	Size	#Documents	#Tokens
Portuguese	Aroeira	**100 GB**	**35.3 Mi**	**15.1 Bi**
	brWac	25 GB	3.53 Mi	2.68 Bi
	Carolina 1.2 Ada	11 GB	2.11 Mi	0.82 Bi
English	MassiveText	**10.5 TB**	**2.35 Bi**	**2.3 Tri***
	The Pile	825 GB	-	-
	C4	750 GB	-	-
Chinease	WuDaoCorpora	**3 TB**	**822 Bi**	-
	CLUEcorpus2020	100 GB	2.35 Bi	36 Bi
Multilingual	RedPajama	**260 TB***	**100 Bi**	**30.4 Tri**
	C4Corpus	29 GB^c	12 Mi	10.8 Bi

Note: Bold letters are best values, "*" point calculated or no paper, and "^{c}" show compressed values.

document's domain (Subsection 3.9). Figure 2 illustrates the document distribution across these domains.

The complexity of the corpus strongly correlates with downstream data performance [3]. Therefore, an extensive representation of knowledge domains can contribute to the generation of more robust models, potentially improving in-context few-shot learning performance [32].

Most documents could not be assigned to a specific domain and are marked as NR (Not Recognized). Among those that were categorized, blog posts and news articles were the most frequent, although other categories such as institutional texts, e-commerce, and internet forums were also found. Knowledge domains are essential for evaluating the quality of the data in the corpus and for filtering data used in the pre-training phase of domain-specific language models.

5.3 Distribution of documents over time

Our third analysis is the distribution of the corpus documents over time. This temporal analysis is important to identify possible temporal biases such as outdated texts. Our corpus presents a recent data distribution, which indicates more up-to-date texts.

Figure 3 illustrates that our data set comprises documents spanning up to 7 years, beginning in 2017. The bulk of the data is from 2017 to 2019, but a notable portion of recent data is from 2021, 2022, and 2023. This distribution meets the need for both recent and extensive data. As a result, models can train on up-to-date information and include recent and common terms used in Portuguese.

5.4 Quality Indicators

We used TTR value, symbol word percentage, stopword percentage, valid words, and toxic content as quality indicators. Figure 4 shows the results obtained.

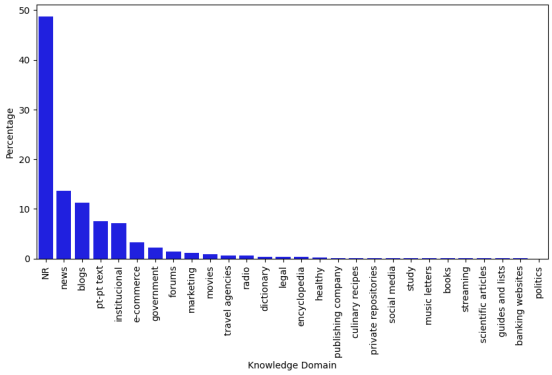

Fig. 2. Distribution of knowledge domains.

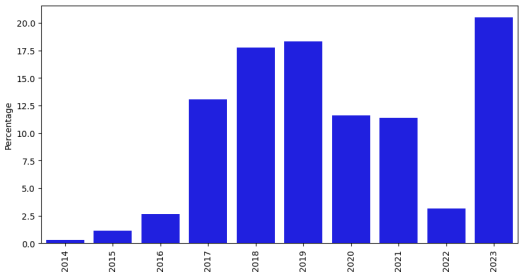

Fig. 3. Distribution of documents over time.

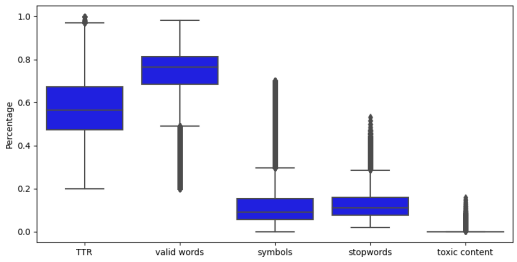

Fig. 4. Quality indicators. Higher TTR and percentage of valid words indicate better corpus quality, while lower values for other indicators also signify better quality.

We have two metrics that indicate the diversity of the content present in our corpus. The TTR indicates the variability of tokens in a sentence, and we aim for this value to be as high as possible, as it is a sign of texts composed of varied tokens with low word repetition. A TTR threshold of 0.5 is a quality parameter for the text. We obtained a distribution curve with the first quartile close to 0.5 TTR, a median of 0.57, and the third quartile above 0.65. Thus, most of the data in our corpus reaches a satisfactory TTR value, which is a strong indicator of non-repetitive texts.

Another indicator of variability is the percentage of valid tokens or keywords. This indicator measures the frequency of contextually relevant words within the text, and we also want higher values for more diverse and fluid texts. We achieved an excellent distribution in this indicator, with most data above 0.7. This distribution value is another strong sign of lexically diverse texts.

In contrast, the metrics for the percentage of symbols and the percentage of stopwords are indicators of less fluid texts, with many symbols interrupting the text or commonly used words that do not add semantic value to the documents (e.g., “to”, “for”, or “the”). Our goal is to minimize these metrics as much as

possible. We achieved our goal of reducing these values, obtaining distributions with low values, with the third quartile below 0.2 in both metrics. This result is a strong indication that the texts in the corpus are fluid.

Finally, we aim to minimize the percentage of potentially toxic words, decreasing to close to 0 (no potentially toxic words). The results show that we achieved this goal in almost all the texts in the corpus, except for some outliers that do not exceed 0.2 percent of toxic content.

5.5 Bias

We performed a word co-occurrence analysis to identify biases in our corpus. This technique has proven effective in demonstrating stereotypical treatment of a particular social group [13], which we do not desire. It is worth noting that this method is one of several mitigation approaches, and the corpus may still exhibit bias. We analyzed three groups of biases as shown in Table 3, they are: (i) Gender, (ii) Religion, and (iii) Race. We selected different words for each group and analyzed the context in which these words were inserted.

We have chosen representative terms indicative of social groups and conducted an analysis focusing on those with the highest co-occurrence frequencies. We isolate the terms "Man" and "Woman" to evaluate gender bias. We selected "Atheist", "Christian", "Buddhist", "Evangelical", "Jewish", "Muslim" and "Umbandist" for religious bias. Finally, we highlighted the words "White", "Black", "Asian" and "Hispanic" for racial bias. The Tables 3 respectively represent the results for gender, religious, and racial bias.

No stereotypical treatment or hazardous behaviors described in the literature exist in the analyzed groups, such as associations with crime, income, and others [16]. Furthermore, the words selected among the various social groups are very similar, which suggests a more equanimous treatment in our proposed corpus.

6 Conclusion

Recent studies have shown significant improvements in the performance of language models trained on large corpora [8, 14]. Consequently, the interest in creating large datasets has grown. Most existing research focuses on high-resource languages like English and Chinese, with considerable efforts made to develop multilingual corpora. However, there is a pressing need to develop large datasets for lower-resource languages.

This work aims to address this gap by developing language models for lower-resource languages, specifically Portuguese. We created the largest curated corpora for training or pre-training Portuguese language models. To achieve this, we implemented a double-pipeline process to extract data while ensuring content safety. The process includes downloading, text extraction, language identification, deduplication, quality filtering, filtering for sexual content and toxicity, bias filtering, categorization, and storage. This effort involved collecting terabytes of

Table 3. Related word co-occurrence.

Topic	Word	1	2	3	4	5	6	7	8	9	10
Gender	Woman (Mulher)	man (homem)	day (dia)	year (ano)	life (vida)	son (filho)	women (mulheres)	mother (mãe)	house (casa)	against (contra)	husband (marido)
	Man (Homem)	woman (mulher)	gave (deu)	year (ano)	life (vida)	world (mundo)	son (filho)	spider (aranha)	well (bem)	day (dia)	because (porque)
Religion	Atheist (Ateu)	gave (deu)	Christian (cristão)	faith (fé)	all (todo)	religion (religião)	person (pessoa)	life (vida)	because (porque)	state (estado)	religious (religioso)
	Christian (Cristão)	gave (deu)	life (vida)	all (todo)	world (mundo)	Christ (cristo)	church (igreja)	must (deve)	Jesus (Jesus)	can (pode)	love (amor)
	Buddhist (Budista)	temple (templo)	monk (monge)	meditation (meditação)	zen (zen)	practice (prática)	Buddhism (budismo)	year (ano)	tradition (tradição)	all (todo)	about (sobre)
	Evangelical (Evangélico)	pastor (pastor)	church (igreja)	means (meio)	day (dia)	Christian (cristão)	gave the (deu o)	year (ano)	hospital (hospital)	Catholic (católico)	Brazil (brasil)
	Jewish (Judeu)	people (povo)	state (estado)	Jesus (jesus)	all (todo)	gave (deu)	Israel (israel)	day (dia)	history (história)	Christian (cristão)	other (outro)
	Muslim (Muçulmano)	world (mundo)	Christian (cristão)	all (todo)	country (país)	Arab (árabe)	Jewish (judeu)	about (sobre)	state (estado)	can (pode)	other (outro)
	Umbanda (Umbandista)	Umbanda (umbanda)	religion (religião)	great (grande)	day (dia)	good (bom)	all (todo)	catholic (católico)	about (sobre)	true (verdadeiro)	end (fim)
Race	White (Branco)	black (preto)	river (rio)	castle (castelo)	color (cor)	blue (azul)	core (core)	wine (vinho)	red (vermelho)	green (verde)	first (primeiro)
	Black (Negro)	red (rubro)	river (rio)	hole (buraco)	white (branco)	movement (movimento)	side (lado)	about (sobre)	year (ano)	Brazil (Brasil)	humor (humor)
	Hispanic (Hispânico)	world (mundo)	Spanish (espanhol)	qualified (qualificado)	black (negro)	work (trabalho)	about (sobre)	great (grande)	soccer (futebol)	public (público)	Mexico (México)
	Asian (Asiático)	southeast (sudeste)	country (país)	market (mercado)	continent (continente)	countries (países)	China (China)	south (sul)	east (leste)	year (ano)	Africa (África)

data, resulting in a curated dataset of approximately 100 GB for Aroeira's construction.

Our results demonstrate that our corpus fulfills the requirements of corpus size, knowledge domains, document distribution over time, quality indicators, and bias mitigation. We conducted statistical analyses of the corpus to better comprehend the size of the collected documents in terms of the number of tokens, words, and sentences. Additionally, we analyzed bias to recognize potential harms in the created corpus, a distinguishing factor in building models free from social biases. Our findings conclude that the corpus created is of high quality and diversity, with minimal bias.

We are eager to advance our work by training models with encoder architectures, such as BERT models [21]. Furthermore, we plan to pre-train language models with Aroeira to obtain higher-quality models in Portuguese. We intend to research existent instruction and evaluation datasets. Finally, conducting comparative bias analyses on models trained with Aroeira relative to other available corpora [13] will be highly valuable.

Acknowledgments. We thank Instituto de Ciência e Tecnologia Itaú-Unibanco (ICTi) and Itaú-Unibanco SA for the technical support, resources, and financial aid in the development of the Aroeira corpus. It's also noteworthy the fact that ChatGPT (OpenAI) was employed in the writing process, contributing to thorough grammatical and semantic reviews.

Data Availability Our Aroeira corpus is available for download in the Hugging Face repository: `https://huggingface.co/datasets/Itau-Unibanco/aroeira` and is under the CC-BY-NC 4.0 license.

Bibliography

[1] Abadji, J., Ortiz Suárez, P.J., Romary, L., Sagot, B.: Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus (2021). https://doi.org/10.14618/IDS-PUB-10468

[2] Abadji, J., Suarez, P.O., Romary, L., Sagot, B.: Towards a cleaner document-oriented multilingual crawled corpus (Jan 2022)

[3] Agrawal, A., Singh, S.: Corpus complexity matters in pretraining language models. pp. 257–263 (01 2023). https://doi.org/10.18653/v1/2023.sustainlp-1.20

[4] Almeida, T.S., Abonizio, H., Nogueira, R., Pires, R.: Sabi\'a-2: A new generation of portuguese large language models. arXiv preprint arXiv:2403.09887 (2024)

[5] Barbaresi, A.: Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In: Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 122–131. Association for Computational Linguistics (2021), https://aclanthology.org/2021.acl-demo.15

[6] Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in nlp. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.485

[7] 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. arXiv preprint arXiv:2108.07258 (2021)

[8] Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[9] Carmo, D., Piau, M., Campiotti, I., Nogueira, R., Lotufo, R.: Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. arXiv preprint arXiv:2008.09144 (2020)

[10] Computer, T.: Redpajama: An open source recipe to reproduce llama training dataset (2023), https://github.com/togethercomputer/RedPajama-Data

[11] Crespo, M.C.R.M., et al.: Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information (Mar 2023)

[12] Finardi, P., Viegas, J.D., Ferreira, G.T., Mansano, A.F., Caridá, V.F.: Berta\'u: Ita\'u bert for digital customer service. arXiv preprint arXiv:2101.12015 (2021)

[13] Gallegos, I.O., et al.: Bias and fairness in large language models: A survey. Computational Linguistics pp. 1–79 (Jun 2024). https://doi.org/10.1162/coli_a_00524
[14] Gao, L., et al.: The pile: An 800gb dataset of diverse text for language modeling (Dec 2020)
[15] Garcia, G.L., et al.: Introducing bode: A fine-tuned large language model for portuguese prompt-based task. arXiv preprint arXiv:2401.02909 (2024)
[16] Garimella, A., Mihalcea, R., Amarnath, A.: Demographic-aware language model fine-tuning as a bias mitigation technique. In: He, Y., et al. (eds.) Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 311–319. Association for Computational Linguistics, Online only (Nov 2022), https://aclanthology.org/2022.aacl-short.38
[17] Habernal, I., Zayed, O., Gurevych, I.: C4Corpus: Multilingual web-size corpus with free license. In: Calzolari, N., et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 914–922. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1146
[18] Hoffmann, J., et al.: An empirical analysis of compute-optimal large language model training. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 30016–30030. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf
[19] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
[20] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics (April 2017)
[21] Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, p. 2 (2019)
[22] Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., Caridá, V.: Cabrita: closing the gap for foreign languages. arXiv preprint arXiv:2308.11878 (2023)
[23] Liu, Y., Cao, J., Liu, C., Ding, K., Jin, L.: Datasets for large language models: A comprehensive survey. arXiv preprint arXiv:2402.18041 (2024)
[24] Mei, K., Fereidooni, S., Caliskan, A.: Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In: 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23, ACM (Jun 2023). https://doi.org/10.1145/3593013.3594109

[25] Nadeem, M., Bethke, A., Reddy, S.: Stereoset: Measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.416
[26] Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R.: Crows-pairs: A challenge dataset for measuring social biases in masked language models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.154
[27] Overwijk, A., Xiong, C., Callan, J.: Clueweb22: 10 billion web documents with rich information. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3360–3362 (2022)
[28] Pires, R., Abonizio, H., Almeida, T.S., Nogueira, R.: Sabiá: Portuguese large language models. In: Brazilian Conference on Intelligent Systems. pp. 226–240 (2023)
[29] Rae, J.W., et al.: Scaling language models: Methods, analysis & insights from training gopher (Dec 2021)
[30] Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), \url{http://jmlr.org/papers/v21/20-074.html}
[31] Richards, B.: Type/token ratios: what do they really tell us? Journal of Child Language **14**(2), 201–209 (Jun 1987). https://doi.org/10.1017/s0305000900012885
[32] Shin, S., Lee, S.W., Ahn, H., Kim, S., Kim, H., Kim, B., Cho, K., Lee, G., Park, W., Ha, J.W., Sung, N.: On the effect of pretraining corpora on in-context learning by a large-scale language model (2022)
[33] Souza, F., Nogueira, R., Lotufo, R.: Bertimbau: pretrained bert models for brazilian portuguese. In: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9. pp. 403–417. Springer (2020)
[34] Virtanen, A., et al.: Multilingual is not enough: Bert for finnish (Dec 2019)
[35] Wagner Filho, J.A., et al.: The brWaC corpus: A new open resource for Brazilian Portuguese. In: Calzolari, N., et al. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1686
[36] Xu, L., Zhang, X., Dong, Q.: Cluecorpus2020: A large-scale chinese corpus for pre-training language model. arXiv preprint arXiv:2003.01355 (2020)
[37] Youmans, G.: Measuring lexical style and competence: The type-token vocabulary curve. Style **24**(4), 584–599 (1990), http://www.jstor.org/stable/42946163
[38] Yuan, S., Zhao, H., Du, Z., Ding, M., Liu, X., Cen, Y., Zou, X., Yang, Z., Tang, J.: Wudaocorpora: A super large-scale chinese corpora for pre-training language models. AI Open **2**, 65–68 (2021)