The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al.

The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset

Hugo Laurençon[1*] **Lucile Saulnier**[1*] **Thomas Wang**[1*] **Christopher Akiki**[2*]
Albert Villanova del Moral[1*] **Teven Le Scao**[1*]

Leandro von Werra[1] **Chenghao Mou**[3] **Eduardo González Ponferrada**[4] **Huu Nguyen**[5]
Jörg Frohberg [32] **Mario Šaško** [1] **Quentin Lhoest** [1]

Angelina McMillan-Major[1,6] **Gérard Dupont**[7] **Stella Biderman**[8,9] **Anna Rogers**[10]
Loubna Ben allal[1] **Francesco De Toni**[11] **Giada Pistilli**[1] **Olivier Nguyen** [28]
Somaieh Nikpoor[12] **Maraim Masoud**[13] **Pierre Colombo**[14] **Javier de la Rosa**[15]
Paulo Villegas[16] **Tristan Thrush**[1] **Shayne Longpre**[17] **Sebastian Nagel**[19] **Leon Weber**
[20] **Manuel Romero Muñoz** [21] **Jian Zhu** [22] **Daniel van Strien** [23] **Zaid Alyafeai** [24]
Khalid Almubarak [25] **Vu Minh Chien** [26] **Itziar Gonzalez-Dios** [27] **Aitor Soroa** [27]
Kyle Lo [29] **Manan Dey** [30] **Pedro Ortiz Suarez** [31] **Aaron Gokaslan** [18] **Shamik Bose**[3]
David Ifeoluwa Adelani[33] **Long Phan**[34] **Hieu Tran**[34] **Ian Yu**[35] **Suhas Pai**[36]
Jenny Chim[37]

Violette Lepercq[1] **Suzana Ilić**[1] **Margaret Mitchell**[1] **Sasha Luccioni**[1] **Yacine Jernite**[1]

[1]Hugging Face [2]Leipzig University [3]Independent Researcher [4]Ferrum Health
[5]Ontocord.ai [6]University of Washington [7]Mavenoid [8]EleutherAI [9]Booz Allen Hamilton
[10]University of Copenhagen [11]University of Western Australia [12]CAIDP
[13]Independent Researcher [14]CentraleSupélec [15]National Library of Norway
[16]Telefonica I+D [17]MIT [18]Cornell University [19]Common Crawl
[20]Humboldt-Universität zu Berlin and Max Delbrück Center for Molecular Medicine [21]Narrativa
[22]University of Michigan, Ann Arbor [23]British Library
[24]King Fahd University of Petroleum and Minerals
[25]Prince Sattam bin Abdulaziz University (PSAU) [26]DETOMO Inc.
[27]HiTZ Center, University of the Basque Country (UPV/EHU) [28]ServiceNow
[29]Allen Institute for AI [30]SAP [31]Mannheim University [32]Apergo.ai [33]Saarland University
[34]VietAI Research [35]Aggregate Intellect [36]Bedrock AI [37]Queen Mary University of London

[*] Equal contributions

Abstract

As language models grow ever larger, the need for large-scale high-quality text datasets has never been more pressing, especially in multilingual settings. The BigScience workshop, a 1-year international and multidisciplinary initiative, was formed with the goal of researching and training large language models as a values-driven undertaking, putting issues of ethics, harm, and governance in the foreground. This paper documents the data creation and curation efforts undertaken by BigScience to assemble the Responsible Open-science Open-collaboration Text Sources (**ROOTS**) corpus, a 1.6TB dataset spanning 59 languages that was used to train the 176-billion-parameter BigScience Large Open-science Open-access Multilingual (**BLOOM**)(BigScience Workshop, 2022) language model. We further release a large initial subset of the corpus and analyses thereof, and hope to empower large-scale monolingual and multilingual modeling projects with both the data and the processing tools, as well as stimulate research around this large multilingual corpus.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

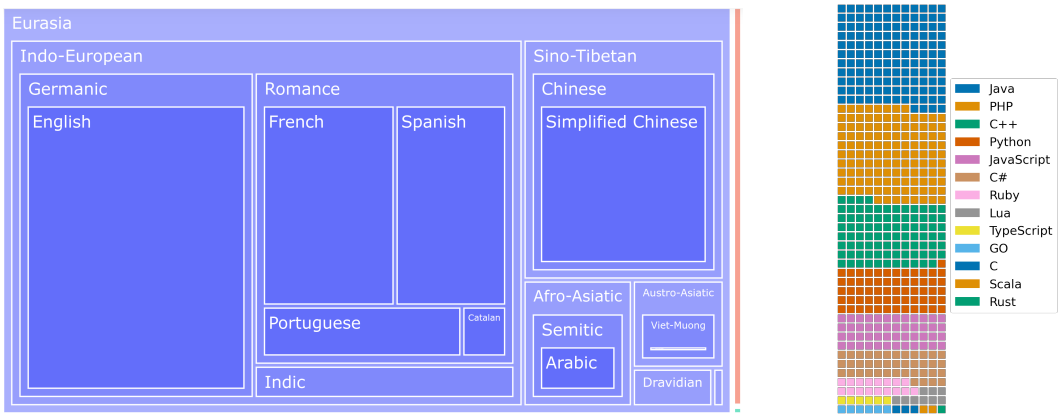


Figure 1: Overview of ROOTS. Left: A treemap of natural language representation in number of bytes by language family. The bulk of the graph is overwhelmed by the 1321.89 GB allotted to Eurasia. The orange rectangle corresponds to the 18GB of Indonesian, the sole representative of the Papunesia macroarea, and the green rectangle to the 0.4GB of the Africa linguistic macroarea. Right: A waffle plot of the distribution of programming languages by number of files. One square corresponds approximately to 30,000 files.

1 Introduction

BigScience[1] started in May 2021 as a one-year long open collaborative research initiative that gathered over a thousand participants around the world to study large language models (LLM). One of the founding goals of BigScience was to train an open-access, massively multilingual LLM, comparable in scale to GPT-3 (Brown et al., 2020) yet trained on a better documented and more representative multilingual dataset. The overall BigScience workshop was designed as a collaborative (Caselli et al., 2021; Bondi et al., 2021) and value-driven (Birhane et al., 2021) endeavor. Throughout the process of building this corpus we engaged in simultaneous investigation of ethical (Talat et al., 2022), sociopolitical (McMillan-Major et al., 2022), and data governance issues (Jernite et al., 2022) with the explicit goal of doing good for and by the people whose data we collected.

Sourcing and building the dataset was organized around four working groups: **Data Governance** which helped define the project's values and design our approach to data usage and release in an international context, **Data Sourcing and Preparation** which was tasked with overseeing data collection, curation efforts, and **Privacy** for privacy risks and sanitizing the dataset, **Legal Scholarship** which helped define the multi-jurisdiction legal context in which the entire workshop was to operate, and we discuss practical implications throughout the paper where appropriate. An overview of the BigScience Corpus is provided in figure 1.

The goal of the current paper is twofold: (1) we present a preliminary gated, subject to committing to the BigScience ethical charter[2], release of a large subset of ROOTS[3] (2) we release the numerous data tools[4] that were developed along the way and enabled us to curate, source, clean and inspect all 498 constituent datasets that come together to constitute ROOTS. This includes a preliminary results of the analyses that are currently being developed to study the corpus.

1.1 Outline of the Paper

The remainder of this paper details our approach to curating a web-scale dataset covering 59 languages, 46 natural languages and 13 programming languages — the language choice was chiefly driven by the communities who participated in the effort given the importance we placed on language expertise. Our final corpus is made up of two main components: 62% of the text comes from a community-selected and documented list of language data sources and its collection process is described in section 2, and

[1] https://bigscience.huggingface.co/
[2] https://hf.co/spaces/bigscience/ethical-charter
[3] https://hf.co/bigscience-data
[4] https://github.com/bigscience-workshop/data-preparation

38% consists of text extracted from a pre-processed web crawl, OSCAR (Ortiz Suárez et al. (2020)), filtered with the help of native speakers, which is described in section 3.

1.2 Related Work

Large Language Models and Large Text Corpora The current dominant paradigm in natural language processing relies heavily on pre-trained models: large language models that can then be fine-tuned on a downstream task (Howard and Ruder, 2018; Devlin et al., 2018) or even used as-is without additional data (Radford et al., 2019; Brown et al., 2020). In this paradigm, performance is directly correlated on both the model size and the dataset size and quality (Kaplan et al., 2020), with recent models trained on up to 1.4 trillion tokens (Hoffmann et al., 2022) and dataset creation pipelines representing a significant part of large language model projects. Most such datasets, however, are not released, hindering further research. Exceptions include the Pile (Gao et al., 2020), a curated corpus of datasets for language modeling that has become widely used for training state-of-the-art English-language models (Lieber et al., 2021; Smith et al., 2022; Black et al., 2022; Zhang et al., 2022), and C4 and mC4 (Raffel et al., 2020; Xue et al., 2020), which have powered the T5 family of models; CC100 (Conneau et al., 2020) which has seen heavy use for multilingual modeling; and OSCAR (Ortiz Suárez et al., 2019), which has enabled monolingual non-English models.

Tooling, Visualization, and Replication Upstream from the finalized training datasets is the issue of processing methods and pipelines: both the operations that the datasets go through and the engineering effort required to apply them at terabyte scales. Existing work tends to fall on a spectrum from no details at all (Brown et al., 2020) to detailed filtering instructions, with (Raffel et al., 2020) or without the dataset release (Rae et al., 2021) to detailed filtering instructions with the accompanying code (Gao et al., 2020; Conneau et al., 2020; Ortiz Suárez et al., 2019). Even when the code is released, it tends to be built and tailored for the project's purpose. Consequently, large projects that do not re-use an existing dataset outright usually build their own pipeline rather than re-use an existing one on new data. However, data tools that were built and packaged in order to be used for other projects exist, such as OSCAR's Ungoliant and Goclassy (Abadji et al., 2021; Ortiz Suárez et al., 2019), which provides a distributed Common Crawl processing pipeline; CCNet (Wenzek et al., 2020), built for quality filtering of multilingual Common Crawl dumps; and OpenWebText (Gokaslan and Cohen, 2019), enabling Reddit dump processing.

Documenting Textual Corpora in NLP An inspiration for our work is a recent emphasis on a more in-depth documentation of what is included and what is not in the corpora used for training NLP models . The most notable example of this is the Pile, for which the authors themselves analyze and document a variety of syntactic and semantic properties of the dataset including structural statistics (n-gram counts, language, document sizes), topical distributions across its components, social bias and sentiment co-occurrence, pejorative content, and information about licensing and authorial consent, in addition to releasing a datasheet (Biderman et al., 2022). Other LM pre-training datasets that have been documented and analyzed include C4 (Dodge et al., 2021; Luccioni and Viviano, 2021; Kreutzer et al., 2022), OSCAR (Kreutzer et al., 2022) and BookCorpus (Bandy and Vincent, 2021) . While this kind of documentation is far from standard practice, it is becoming increasingly common given recent calls for better documentation (Rogers, 2021; Bender et al., 2021) as well as empirical studies on data memorization in language models (Carlini et al., 2019, 2022).

2 (Crowd) Sourcing a Language Resource Catalogue

The first part of our corpus, accounting for 62% of the final dataset size (in bytes), was made up of a collection of monolingual and multilingual language resources that were selected and documented collaboratively through various efforts of the BigScience Data Sourcing working group. The first such effort consisted in creating a tool to support metadata collection through open submissions, called the BigScience Catalogue and running a series of hackathons in collaboration with locally-focused ML and NLP communities such as Masakhane, Machine Learning Tokyo and LatinX in AI where participants could add and document entries for their languages to the catalogue (McMillan-Major et al., 2022). This yielded a set of 252 sources, including at least 21 per considered language category. We focused on metadata collection as a way to support selection of the sources for the final dataset and documentation of the final dataset. In parallel, working group participants gathered additional

Arabic language resources in the Masader repository (Alyafeai et al., 2021), and proposed a list of websites of interest to increase the geographical diversity of our English, Spanish, and Chinese language data. Finally, in order to explicitly test large language models' ability to handle computer code along with natural language, we selected code data available on GitHub and StackExchange.

2.1 Obtaining Data from the Identified Resources

Gathering Identified Datasets and Collections. First, we leveraged the BigScience Catalogue and the Masader repository to start obtaining text from identified sources, which included both existing NLP datasets and collections of documents of various compositions. Given the diversity of sources, hosting methods, data custodians, and formats, collecting this text required a collaborative effort. To that end, we established a 2-phase approach: first, collect as many data sources as possible in an easily accessible location; second, map all of them to a common format to ease further processing.

In the first phase, we organized an open hackathon to start gathering identified sources on the Hugging Face Datasets hub (Lhoest et al., 2021), in a dedicated organization[5] (in order to manage access controls). In the second phase, the collected datasets were furthered processed via (1) *Language segmentation*, whereby data sources were split using metadata for each covered language in order to obtain monolingual datasets, and the use of (2) *Uniform interface* whereby a document consisted of two fields: "text" for the actual text content, and "meta" with a JSON representation of metadata for a given document, containing sufficient information to trace documents back to their original sources.

Pseudo-Crawled Data. Of the various categories of language resources identified through the data sourcing effort, websites stood out as one that required a particular effort and dedicated pipeline. We decided to design such a pipeline based on "pseudo-crawling": that is, rather than crawling the websites ourselves, we retrieved pages corresponding to the target domain names from 18 snapshots archived by Common Crawl in 2020 and 2021 in Web ARChive (WARC) format (Mohr et al., 2008). These domain names came from two main sources: the homepage field in the metadata of the 252 above-mentioned catalogue entries when available (192 in total), and the 456 websites proposed by participants asynchronously to improve the geographical diversity of our language sources; which yielded a total of 614 unique domain names after deduplication.

We collected URLs contained within those domains using the Common Crawl index. The index provides metadata for every document including the page URL, WARC filename and record offsets, fetch status, content MIME type, etc. We ran a query matching all documents that share the domain name with a seed using Amazon Athena on Common Crawl's columnar index[6]. 48 of the 614 initial seed domain names had no matches in the index and were therefore left out. Once we obtained the document metadata, we fetched the WARC records using HTTP range requests with the start and end byte offsets. Since HTML web pages constitute the largest portion of pages contained in the Common Crawl dumps, we decided to only extract text from HTML pages. Documents in other formats were filtered out, ie XML, PDF, etc. 27 domain names were additionally removed from the list at this stage as we had not retrieved any HTML pages for them.

To extract the text from the HTML pages, we first minified the HTML code. Minification is the removal of unnecessary characters from the source code of a website. Inspired by Aghajanyan et al. (2022), we removed from the DOM-HTML all the sub-trees contained in a <*script*>, <*style*>, <*header*>, <*iframe*>, <*footer*> and <*form*> tag as well as all the sub-trees associated with a <*body*>, <*div*>, <*p*>, <*section*>, <*table*>, <*ul*>, <*ol*> or <*dl*> tag whose textual content was less than 64 characters long. The text was then extracted from the nodes of this new DOM-HTML. While concatenating the text extracted, we applied a set of rules to reconstruct the structure of the text without its HTML code, inspired by what Common Crawl does to extract its WET files (Appendix A.1). The overall procedure enabled us to obtain text datasets for 539 domain names.

GitHub Code. We collected a code dataset from BigQuery[7] using the same language selection as AlphaCode (Li et al., 2022). The dataset was then deduplicated of exact matches and filtered for source files with between 100 and 200,000 characters, between 15-65% alphabetic characters, a max

[5]https://hf.co/bigscience-catalogue-data

[6]https://commoncrawl.org/2018/03/index-to-warc-files-and-urls-in-columnar-format/

[7]"GitHub on BigQuery: Analyze all the open source code"

line length of 20-1000 characters, and a token length standard deviation of more than 3. Due to a bug in the pre-processing pipeline the dataset was also filtered for GPL licenses only.

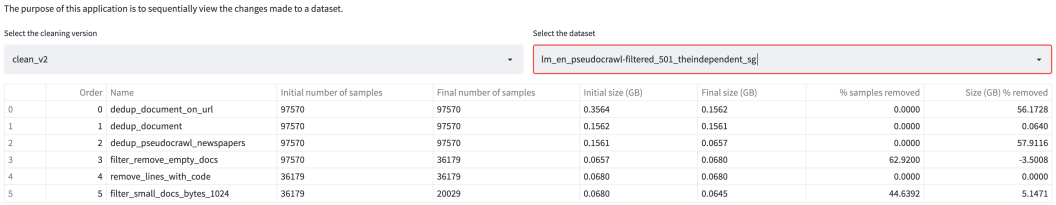
Merging and Deduplicating Sources. After gathering and processing language data via the three pipelines outlined above, we took a final step to manually inspect, deduplicate, and make a further selection of the sources. First, we addressed dataset overlap we found by looking through our sources. For example: *OpenITI* was present in both its raw form as well as a processed version. Consensus was reached to choose the latter version. Non-trivial datasets overlap included *s2orc* (Lo et al., 2020), *Arxiv* (Clement et al., 2019) and the *PubMed Central* subset of the Pile (Gao et al., 2020). We also performed cross-pipeline dataset deduplication, removing the pseudo-crawled Wikipedia and GitHub in favor of their other versions. We also removed datasets that we found had a high incidence of documents that were not fully in natural language (e.g. unexpected instances of SEO, HTML tags etc...), as well as very small datasets in the higher-resourced languages. Finally, pseudo-crawled sources were further processed to remove menus (with a heuristic consisting of removing lines that occurred in more than 1% of pages in a given domain) and pages that had a high incidence of character ngram repetition, low language identification confidence, or low proportion of closed class words (see Section 3). We then removed entire domains whose size was less than 2MB after this step, yielding 147 pseudo-crawl-based datasets, and a total of 517 datasets including all three pipelines.

2.2 Processing Pipeline for Quality Improvement on Crowdsourced Datasets

Once a text field was obtained, we attempted to improve the quality of that text. In the specific case of text extraction from HTML, we observe that not all text are relevant (menus, advertisements, repeated text on each page etc ...). In order to remove noisy data from our dataset, we applied a processing pipeline for each dataset consisting of a sequence of functions.

Functions were categorised as *document-scoped* or *dataset-scoped* functions. *Document-scoped* functions are operations that modify a document independently of other documents and *dataset-scoped* functions are operations that take into account the whole dataset. Orthogonal to this scope, functions were also separated into *cleaning* and *filtering* functions. *Cleaning functions* aim to remove text considered not part of the main document. Document-scoped cleaning functions can for example target leftover HTML tags. On the other end, dataset-scoped cleaning functions need the whole dataset to calculate a heuristic to determine how to modify each document. For instance, advertisements vary across datasets, making it harder to define a dataset-agnostic classifier for advertisement. Instead, we can index all the lines in a dataset and identify repeated lines on multiple pages as likely advertisements. An example is displayed in Appendix A.2. *Filtering functions* aim at removing an entire document from the corpus. The reasons for choosing to remove a document completely are diverse: it may be because the document is considered to be of too poor quality, to be too complex to automatically fix or too similar to other examples already present in the corpus. In the latter case, we speak of deduplication. Deduplication of a document is dependent on whether an equivalent document already exists somewhere else in the dataset and is thus necessarily a dataset-scope function. The notion of equivalent documents has been explored by Lee et al. (2022). In this case we provide deduplication via metadata (urls, normalised urls) and via text (exact string matching). An exhaustive list of functions is available in A.3.

As datasets came from heterogeneous sources with different properties, each needs its own set of processing functions to correspond to our definition of natural language documents. In order to support participants in deciding what functions to apply to which, we built and released a *streamlit*-based visualization tool (figure 2 helps understand the impact of each function, displaying how a document was altered/removed as well as estimated dataset level metrics (quantity of data removed in bytes or samples)). This rapid feedback loop enabled us to update the pipeline consequently in an iterative process to finetune each processing pipelines across datasets and languages with the input of native speakers. A specific example is shared in Appendix A.2. This resulted in 485 non-empty datasets.



	Order	Name	Initial number of samples	Final number of samples	Initial size (GB)	Final size (GB)	% samples removed	Size (GB) % removed
0	0	dedup_document_on_url	97570	97570	0.3564	0.1562	0.0000	56.1728
1	1	dedup_document	97570	97570	0.1562	0.1561	0.0000	0.0640
2	2	dedup_pseudocrawl_newspapers	97570	97570	0.1561	0.0657	0.0000	57.9116
3	3	filter_remove_empty_docs	97570	36179	0.0657	0.0680	62.9200	-3.5008
4	4	remove_lines_with_code	36179	36179	0.0680	0.0680	0.0000	0.0000
5	5	filter_small_docs_bytes_1024	36179	20029	0.0680	0.0645	44.6392	5.1471

Figure 2: Partial screenshot of the visualization tool. Users can look at how each function in the processing pipeline influenced high-level statistics. Influence on specific samples can be monitored via the same tool, see Appendix A.2

3 Processing OSCAR

We chose to complement the data obtained at the end of the process described in the previous section with additional Common Crawl-based[8] data motivated by two main reasons. First, given the project's overall goal of providing a trained LLM as a research artifact comparable to previously released ones that have relied extensively on this source, we assessed that not including it would constitute too much of a departure and risk invalidating comparisons. Relatedly, recent work has put a strong emphasis on the quantity of data being a strong factor in a trained model's performance on evaluation tasks (Kaplan et al., 2020; Hoffmann et al., 2022), and we were missing about one third of data in order to optimize our compute budget in this direction. With that in mind, we chose OSCAR version 21.09 (Ortiz Suárez et al., 2020), based on the Common Crawl snapshot of February 2021, to make up the remaining 38% of our final dataset.

However, crawled data suffers from several known issues. First, we wanted to only select documents written by humans for humans, and exclude machine-generated content e.g. search engine optimization (SEO). Crawled content also over-represents pornographic text across languages (Kreutzer et al., 2022), especially in the form of spam ads. Finally, it contains personal information that may constitute a privacy risk. The present section outlines our approach to mitigating those issues.

3.1 Data cleaning and filtering

Our first approach to addressing the above consists in defining quality indicators for web content. These can then be used to filter out specific pages by defining cutoff thresholds. Extensive descriptions for reproduction are available in appendix B. We filtered out documents with:

- Too high **character repetition** or **word repetition** as a measure of repetitive content.
- Too high ratios of **special characters** to remove page code or crawling artifacts.
- Insufficient ratios of **closed class words** to filter out SEO pages.
- Too high ratios of **flagged words** to filter out pornographic spam. We asked contributors to tailor the word list in their language to this criterion (as opposed to generic terms related to sexuality) and to err on the side of high precision.
- Too high **perplexity** values to filter out non-natural language.
- Insufficient **number of words**, as LLM training requires extensive context sizes.

The languages that we eventually considered in OSCAR were the languages for which we were able to obtain hyperparameters and the cutoff values for each of these indicators by native speakers. Specifically, we considered Arabic, Basque, Bengali, Catalan, Chinese, English, French, Hindi, Indonesian, Portuguese, Spanish, Urdu, and Vietnamese. The code used for filtering OSCAR, along with the language-specific parameters and cutoff values, are publicly available. We then asked native speakers of each language to use our visualization tool[9] to establish the thresholds for each filter. The percentage of documents removed after applying all these filters is given in Table 1, and the percentage of documents discarded by each filter independently is given in 3.

[8] https://commoncrawl.org/

[9] https://hf.co/spaces/huggingface/text-data-filtering

AR	EU	BN	CA	ZH	EN	FR	HI	ID	PT	UR	VI	ES
20.3	5.2	48.8	21.1	23.1	17.2	17.0	25.7	10.4	12.6	15.8	21.3	16.9

Table 1: Percentage of documents removed by the filtering per language (ISO 639-1 code).

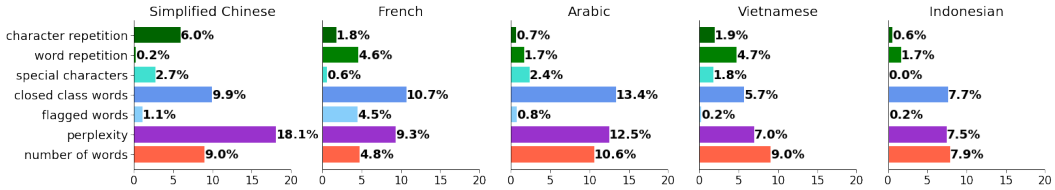


Figure 3: Percentage of documents discarded by each filter independently for 5 languages

3.2 Deduplication

Data deduplication has become a key tool for language model projects following research showing that it both improves performance on downstream tasks (Lee et al., 2022; Zhang et al., 2021) and decreases memorization of training data (Kandpal et al., 2022). To remove near duplicate documents in OSCAR (which is already exact-deduplicated) we initially used SimHash (Charikar, 2002; Manku et al., 2007), a hashing function that associates to two similar texts hashes with a low Hamming distance, with 6-grams and a Hamming distance threshold of 4. About 0.7% of the documents on average (0.07% $\sim$ 2.7%) were identified as near duplicates. However, because SimHash is essentially a bag-of-words algorithm, long documents are more likely to end up being similar to each other. In practice, we found false positives among long documents and decided not to discard documents in a same cluster of near-duplicates when they were longer than 6000 characters. Instead, we applied substring deduplication (Lee et al., 2022) based on Suffix Array (Manber and Myers, 1993) as a complementary method that clusters documents sharing a long substring, for documents with more than 6000 characters. We found on average 21.67% (10.61% $\sim$ 32.30%) of the data (in bytes) being duplicated.

3.3 Personally identifiable information

We used a rule-based approach leveraging regular expressions (Appendix B). The elements redacted were instances of *KEY* (numeric & alphanumeric identifiers such as phone numbers, credit card numbers, hexadecimal hashes and the like, while skipping instances of years and simple numbers), *EMAIL* (email addresses), *USER* (a social media handle) and *IP_ADDRESS* (an IPv4 or IPv6 address).

4 A First look at ROOTS

The efforts described in the previous sections come together in an assemblage of 1.6 Terabytes of multilingual text. Figure 4 puts that number into context by comparing the sizes of corpora typically used to train large language models. Documentation of the individual components of the corpus can be found in an interactive dataset card deck. In this section, we take initial steps towards further understanding of the corpus through statistical analyses of the aggregated data.

4.1 Natural Languages

The constitution of the corpus reflects the crowdsourcing efforts that enabled its creation. It comprises of 46 natural languages spanning 3 macroareas and 9 language families: Afro-Asiatic, Austro-Asiatic, Austronesian, Basque, Dravidian, Indo-European, Mande, Niger-Congo, Sino-Tibetan. At 30.03%, English constitutes the largest part of the corpus, followed by Simplified Chinese (16.16%), French (12.9%), Spanish (10.85%), Portuguese (4.91%) and Arabic (4.6%). A more detailed breakdown of the corpus can be found in the appendix and in an online interactive exploration tool[10],

[10]https://hf.co/spaces/bigscience-data/corpus-map

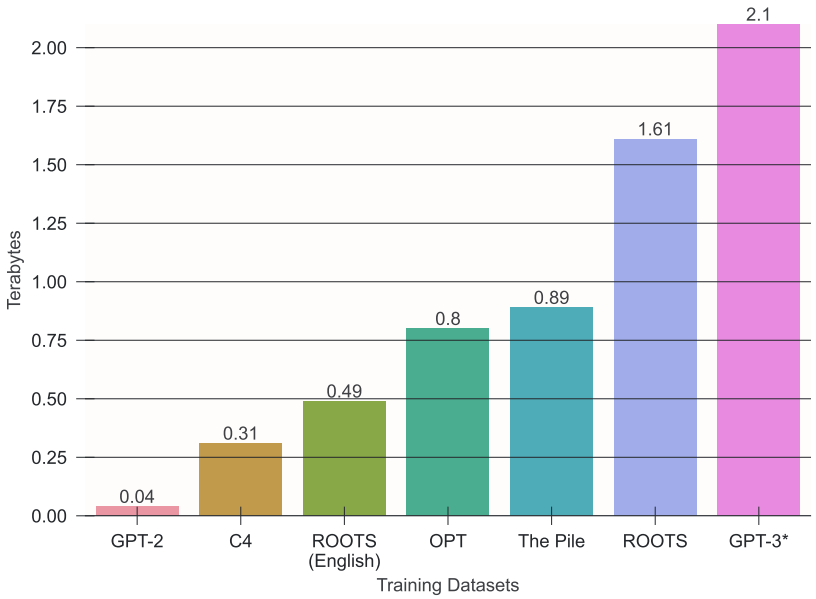


Figure 4: A raw size comparison to other corpora used to train large language models. The asterisk next to GPT-3 indicates the fact that the value in question is an estimate computed using the reported number of tokens and the average number of tokens per byte of text that the GPT-2 tokenizer produces on the `Pile-CC`, `Books3`, `OWT2`, and `Wiki-en` subsets of the Pile (Gao et al., 2020)

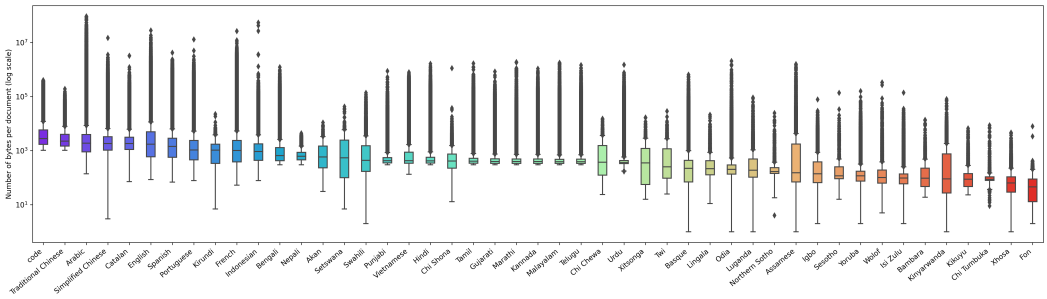


Figure 5: Size in bytes of every document in the corpus per language. The y-axis is in logarithmic scale. Box-and-whisker diagrams illustrate median, the first and third quartiles, whiskers drawn within the 1.5 IQR value and outliers

a screenshot of which is included in figure 1 to depict the byte-distribution of linguistic genera of the Eurasian macroarea subset of the corpus.

In order for the trained model to have an opportunity to learn long dependencies, the training corpus needs to contain long sequences of coherent text. At the same time, the previous post-processing steps only reduced the size of the documents. The median size of a document in our corpus is 1,129 bytes. Figure 5 shows the distribution of document sizes by language. A more detailed breakdown of the size of corpus on an online interactive tool.[11].

The distributions of the filter values for the different filters introduced in Section 3.1 and languages, for the Catalogue, Pseudo-Crawl and OSCAR (filtered) data are available in an online demo[12]. Examples for English are shown in figure 6. The different distributions reflect the diversity of sourcing and filtering of our main components. A notable example is the flagged word filter, for which the distribution for OSCAR is skewed right compared to the catalogue even after filtering.

[11]https://hf.co/spaces/bigscience-data/document-sizes

[12]https://hf.co/spaces/bigscience-catalogue-lm-data/filter_values_distributions

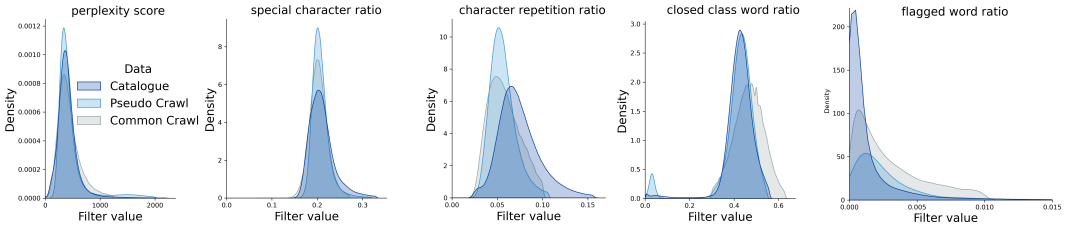


Figure 6: Some distributions of filter values for English. A filter value is the value that the filter gives to a document. These values are generally used to filter out documents that are too low or too high rated and also inform about the composition of the datasets.

4.2 Programming Languages

As depicted in the waffle plot in figure 1, the code subset of the corpus spans 13 programming languages, with Java, PHP, and C++ accounting for more than half of all documents.

Configuration and test files are abundant in most GitHub repositories but not as interesting for code modeling. To that end, we use a heuristic whose first step examines the first 5 lines of a file for the presence of keywords such as "configuration file" or "test file". Failing that, the second step is to see whether the occurrence of the literals `config` and `test` in a given file exceeds 5% of the total number of lines of that file. We find that 5.23% of the data consists of configuration files and 7.88% of test files.

Allamanis (2019) and Lopes et al. (2017) highlight the large fraction of near-duplicates present in code datasets and how they can inflate performance metrics. Exact match deduplication alone can miss a fair amount of near-duplicates. To detect them, we first compute the MinHash of all documents, then create a Locality Sensitive Hashing (LSH) index between files to find the duplicate clusters in linear time. We additionally evaluate the Jaccard similarities within duplicate clusters to remove some false positives. We find 10.9M duplicate files in the clusters and 4.1M unique files: almost 32% of the data consists of near-duplicates. Syntax checkers[13] are used to validate 500K samples of Python and PHP code. We find that only 1% of the Python data and 2% of the PHP files do not pass the syntax check.

4.3 Tokenizer analysis of the component datasets

A tokenizer trained on a dataset can be used as a proxy for its content (Gao et al., 2020). The relevant metric is the number of tokens produced for a byte of natural language. The more different the training corpus from the tokenized corpus, the more tokens will be produced as the tokenizer is forced to divide natural text in more numerous, more general, smaller tokens. This property has allowed us to spot errors associated with outlier values, such as incorrectly classified languages, or crawling error. In the following analysis, we use it in two ways: first, we can use tokenizers trained on different corpora to see how ours differs from them; and second, we can use a tokenizer trained on this corpus to assess which components are outliers. We exclude outliers smaller than 5 documents.

Figure 7 shows the tokens-per-byte measurement on English component datasets for the BLOOM tokenizer, trained on this corpus, the GPT-NeoX 20B tokenizer (Black et al., 2022), trained on the Pile, and the T5 tokenizer (Raffel et al., 2020), trained on C4. Those tokenizers may differ in algorithms and/or vocabulary size, but we won't be directly comparing them to each other.

The figure is ordered by BLOOM tokenizer token-per-byte values, which shows that the ordering is very similar for BLOOM and GPT-NeoX. However, it shows several bumps for T5: component datasets that are out of domain in C4 but not our corpus, for example technical and academic datasets such as `s2orc` or `royal_society_corpus`, domains absent from C4's Common Crawl-sourced data. Other such datasets include `global_voices`, which contains news about non-English-speaking regions including quotes in the original languages and `no_code_stackexchange`, which contains forums which, although in English, may be dedicated to technical matters, foreign languages, or very specific domains. Both are similar to our corpus but not to the Pile or C4.

[13] `py_compile` for Python and the `-l` flag for PHP

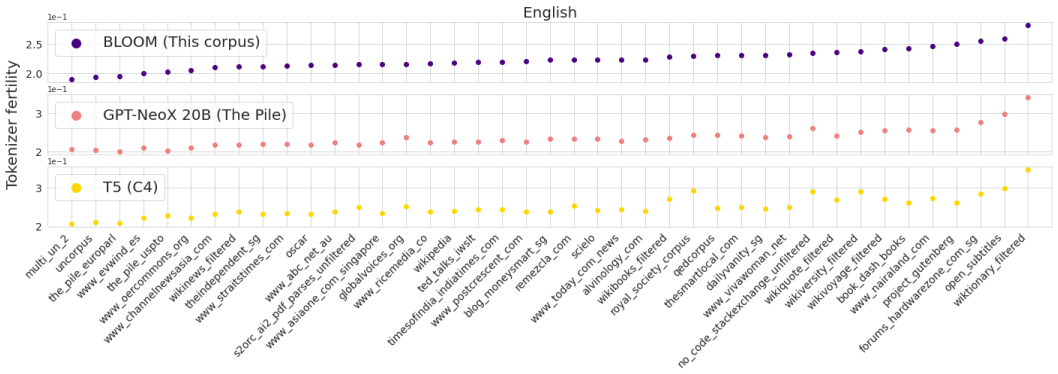


Figure 7: Tokens per byte for each English-language component for tokenizers trained on this corpus (BLOOM), the Pile (GPT-NeoX 20B) and C4 (T5). Lower values mean the component (X axis) is more similar in aggregate to the compared training corpus.

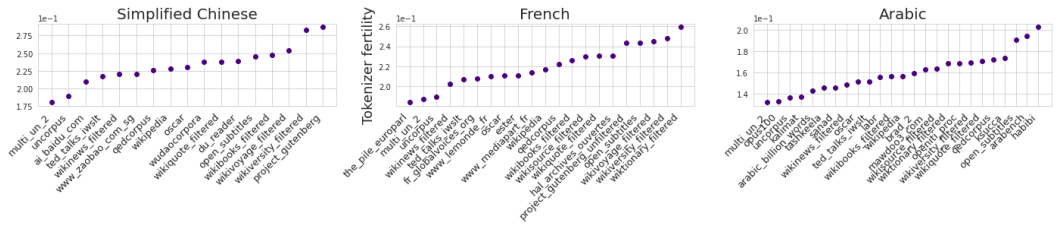


Figure 8: Tokens per byte for each French, Simplified Chinese, and Arabic component for tokenizers trained on this corpus. Lower values mean the component (X axis) is more similar in aggregate to the rest of the corpus.

Figure 8 additionally shows BLOOM fertilities for Simplified Chinese, French and Arabic components. Outlier, high-fertility components, e.g. datasets that differ from the rest of our corpus, tend to be the same for all languages. `project_gutenberg` contains old books with their original formatting (for example, "**********" to denote page ends). `wiktionary` contains definitions of words in foreign languages. `wikiversity` contains technical terms and LaTeX. `wikivoyage` contains tables formatted as text. Forums may contain the user and date information of the message, as well as internet slang or emoji. `arabench` is spoken Arabic, and `habibi` is classical Arabic with more diacritics than modern. We deem most of those deviations acceptable to represent the diversity of uses of text, which tokenizer analysis is able to surface from the rest of the dataset.

5 Conclusion

We have presented ROOTS, a massive multilingual corpus that was the result of an international collaboration between multidisciplinary researchers studying large language models. The efforts to put the corpus together were value-driven and prompted by a data-first approach to training the BLOOM model. We further release the tooling developed throughout the project, and are currently implementing a release strategy that is informed by both the licensing and governance needs of every data source for the corpus itself. We hope this paves the way toward a more reflected use of the data that makes its way into large language models.

Ethical Considerations and Broader Impacts Statement

As discussed in Section 1, the BigScience Research Workshop was conceived as a collaborative and value-driven endeavor from the start. All the ethical efforts were concentrated on implementing the values chosen first on the ethical charter and then on how to articulate those core values into specific ethical sensitive issues, such as data governance. This mechanism also allows ethical thinking to guide governance regarding technical matters. The articulation between the BigScience core values

and those chosen by the collaborators contributing to data efforts was central. The importance of this collective exercise is due to the social impact that technologies such as LLMs have on the people impacted, directly and indirectly, positively and negatively. Moral exercises based on consensus, discussion around values, and how to link technical actions to ethical reflections is a strength that we believe is important within ML research. A critical analysis from an ethical perspective is fundamental to making different disciplines coexist in thinking around the social impact of these technologies and well define the object of analysis, as in this case, a multilingual dataset.

BigScience Values

Motivated by recent work on the values encoded in current approaches to research in NLP and ML more broadly (Leahy and Biderman, 2021; Birhane et al., 2021), which finds that narrow definitions of performance and efficiency were often prioritized over considerations of social impact in research and development. Even more relevant to the corpus creation aspect of our project, Scheuerman et al. (2021) outline how data efforts in computer vision tend to prioritize "*efficiency [over] care; universality [over] contextuality; impartiality [over] positionality*...". These ML research programs and systems in turn support the development of new technologies that carry these same values when deploying these technologies in production (Winner, 2017). This limits the potential positive societal benefits of the rapid advances of NLP research while increasing risks considerably.

Aware of these challenges, participants in BigScience collaboratively drafted an ethical charter[2] formalizing our core values and how they are articulated. It establishes the core values in order to allow its contributors to commit to them, both individually and collectively, and to ground discussions and choices made throughout the project in a common document. These values include notably **openness** and **reproducibility** as a scientific endeavor aimed at advancing the state of the art in a way that can be understood, interrogated, and re-used; **responsibility** of the participants to consider the social and legal context, and the social and environmental consequences of their work; and **diversity** and **inclusivity**. These last two are especially relevant to our data efforts, which aim to include text representative of diverse languages, varieties, and uses through a participatory approach to curation.

Putting Our Values into Practice

Centering Participation in Data Curation Participatory approaches play a vital role in bridging the gaps between model development and deployment and in promoting fairness in ML applications (Rajkomar et al., 2018). They have received increased attention in recent years, with newer work calling to involve participants as full stake-holders of the entire research life-cycle rather to catering their role to *post hoc* model evaluation (Sloane et al., 2020; Caselli et al., 2021; Bondi et al., 2021), as exemplified by an organization like Maskhane (Nekoto et al., 2020) that brings together African researchers to collaboratively build NLP for African languages.

With regard to developing LLMs, BigScience stands in contrast to previous work on models of similar size (Brown et al., 2020; Zhang et al., 2022) — where the majority of the development occurs in-house — by promoting engagement with other communities at every stage of the project from its design to the data curation to the eventual model training and release. Specifically, on the data curation aspect which is the focus of this paper, the involvement of a wide range of participants from various linguistic communities aims to help with the following aspects. First, Kreutzer et al. (2022) have shown in recent work that multilingual text data curation done without involving language-specific expertise leads to resources that are very different from the intentions of their creators, and these limitations carry on to the models trained on these datasets. Second, resources that are developed in collaboration with other communities are more likely to be more directly relevant to them, and thus to avoid reduce replication of model development by making the artifacts and tools we develop useful to more people and for more languages. Third, intentional curation and proper documentation of web-scale corpora takes a significant amount of human work and expertise, which can be distributed between a large number of participants in community efforts. Finally, community involvement can help foster trust and collective ownership of the artifacts we create.

Addressing the Legal Landscape The legal status of webscraped datasets is extremely unclear in many jurisdictions, putting a substantial burden on both data creators and data users who wish to be involved with this process. While the principle of fair use generally protects academic researchers, it is not recognized in all jurisdictions and may not cover research carried out in an industry context. In

consultation with our **Legal Scholarship** and **Data Governance** working groups, we developed a framework (Jernite et al., 2022) to uphold the rights and responsibilities of the many stakeholders in NLP data generation and collection, and provide assurances to downstream users as to how they are and are not authorized to use the dataset (Contractor et al., 2020).

Limitations of the Approach.

While we believe that an approach grounded in community participation and prioritizing language expertise constitutes a promising step toward more responsible data curation and documentation, it still has important limitations. Among those, we primarily identify the use of data from the Common Crawl which represents a point of tension between our drive to present a research artifact that is comparable to previous work and values of consent and privacy (see Section 3). Our pre-processing removes some categories of PII but is still far from exhaustive, and the nature of crawled datasets makes it next to impossible to identify individual contributors and ask for their consent. Similar concerns apply to other existing NLP datasets we identified in the catalogue, including notably the WuDao web-based corpus (Yuan et al., 2021) which makes up a significant part of the Chinese language data. Additionally, while we hope that our intentional approach to selecting diverse data sources (mostly along axes of geographical diversity and domains) will lead to a more representative language dataset overall, our reliance on medium to large sources of digitized content still over-represents privileged voices and language varieties.

Acknowledgements

BigScience. This work was pursued as part of the BigScience research workshop, an effort to collaboratively build a very large multilingual neural network language model and a very large multilingual text dataset. This effort gathered 1000+ reasearchers from 60 countries and from more than 250 institutions.

Compute. The BigScience Workshop was granted access to the HPC resources of the Institut du développement et des ressources en informatique scientifique (IDRIS) du Centre national de la recherche scientifique (CNRS) under the allocation 2021-A0101012475 made by Grand équipement national de calcul intensif (GENCI). Model training ran on the Jean-Zay cluster of IDRIS, and we thank the IDRIS team for their responsive support throughout the project, in particular Rémi Lacroix.

References

Abadji, J., P. J. Ortiz Suárez, L. Romary, and B. Sagot (2021, July). Ungoliant: An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus. In *CMLC 2021 - 9th Workshop on Challenges in the Management of Large Corpora*, Limerick / Virtual, Ireland.

Abdelali, A., F. Guzman, H. Sajjad, and S. Vogel (2014, may). The amara corpus: Building parallel language resources for the educational domain. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Adelani, D. I., J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, S. Mayhew, I. A. Azime, S. H. Muhammad, C. C. Emezue, J. Nakatumba-Nabende, P. Ogayo, A. Anuoluwapo, C. Gitau, D. Mbaye, J. Alabi, S. M. Yimam, T. R. Gwadabe, I. Ezeani, R. A. Niyongabo, J. Mukiibi, V. Otiende, I. Orife, D. David, S. Ngom, T. Adewumi, P. Rayson, M. Adeyemi, G. Muriuki, E. Anebi, C. Chukwuneke, N. Odu, E. P. Wairagala, S. Oyerinde, C. Siro, T. S. Bateesa, T. Oloyede, Y. Wambui, V. Akinode, D. Nabagereka, M. Katusiime, A. Awokoya, M. MBOUP, D. Gebreyohannes, H. Tilaye, K. Nwaike, D. Wolde, A. Faye, B. Sibanda, O. Ahia, B. F. P. Dossou, K. Ogueji, T. I. DIOP, A. Diallo, A. Akinfaderin, T. Marengereke, and S. Osei (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics 9*, 1116–1131.

Aghajanyan, A., D. Okhonko, M. Lewis, M. Joshi, H. Xu, G. Ghosh, and L. Zettlemoyer (2022). HTLM: Hyper-text pre-training and prompting of language models. In *International Conference on Learning Representations*.

Agić, Ž. and I. Vulić (2019, July). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 3204–3210. Association for Computational Linguistics.

Allamanis, M. (2019). The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, pp. 143–153.

Alrabiah, M., A. Alsalman, and E. Atwell (2013, 01). The design and construction of the 50 million words ksucca king saud university corpus of classical arabic.

Aly, M. and A. Atiya (2013, August). LABR: A large scale Arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, pp. 494–498. Association for Computational Linguistics.

Alyafeai, Z., M. Masoud, M. Ghaleb, and M. S. AlShaibani (2021). Masader: Metadata sourcing for arabic text and speech data resources. *CoRR abs/2110.06744*.

Armengol-Estapé, J., C. P. Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, and M. Villegas (2021, August). Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, pp. 4933–4946. Association for Computational Linguistics.

Artetxe, M., I. Aldabe, R. Agerri, O. Perez-de Viñaspre, and A. Soroa (2022). Does corpus quality really matter for low-resource languages?

Ashari, A. (2018). Indonesian news articles published at 2017.

Bandy, J. and N. Vincent (2021). Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*.

Belinkov, Y., A. Magidow, A. Barrón-Cedeño, A. Shmidman, and M. Romanov (2019). Studying the history of the arabic language: language technology and a large-scale historical corpus. *Language Resources and Evaluation 53*(4), 771–805.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, pp. 610–623. Association for Computing Machinery.

Biderman, S., K. Bicheno, and L. Gao (2022). Datasheet for the pile. *arXiv preprint arXiv:2201.07311*.

BigScience Workshop (2022). Bloom (revision 4ab0472).

Bird, S., E. Klein, and E. Loper (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Beijing: O'Reilly.

Birhane, A., P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao (2021). The values encoded in machine learning research. *ArXiv abs/2106.15590*.

Black, S., S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, et al. (2022). Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode\# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136.

Bondi, E., L. Xu, D. Acosta-Navas, and J. A. Killian (2021). Envisioning communities: A participatory approach towards AI for social good. In M. Fourcade, B. Kuipers, S. Lazar, and D. K. Mulligan (Eds.), *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pp. 425–436. ACM.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Budiono, H. Riza, and C. Hakim (2009, August). Resource report: Building parallel text corpora for multi-domain translation system. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, Suntec, Singapore, pp. 92–95. Association for Computational Linguistics.

Bình, V. Q. (2021). Binhvq news corpus. https://github.com/binhvq/news-corpus.

Carlini, N., D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang (2022). Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Carlini, N., C. Liu, Ú. Erlingsson, J. Kos, and D. Song (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284.

Carrino, C. P., C. G. Rodriguez-Penagos, and C. Armentano-Oller (2021, March). Tecla: Text classification catalan dataset.

Caselli, T., R. Cibin, C. Conforti, E. Encinas, and M. Teli (2021). Guiding principles for participatory design-inspired natural language processing. *Proceedings of the 1st Workshop on NLP for Positive Impact*.

Cettolo, M., C. Girardi, and M. Federico (2012, May 28–30). WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, Trento, Italy, pp. 261–268. European Association for Machine Translation.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, New York, NY, USA, pp. 380–388. Association for Computing Machinery.

Chen, Y. and A. Eisele (2012, May). MultiUN v2: UN documents with multilingual alignments. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pp. 2500–2504. European Language Resources Association (ELRA).

Clement, C. B., M. Bierbaum, K. P. O'Keeffe, and A. A. Alemi (2019). On the use of arxiv as a dataset. *CoRR abs/1905.00075*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 8440–8451. Association for Computational Linguistics.

Contractor, D., D. McDuff, J. Haines, J. Lee, C. Hines, and B. Hecht (2020). Behavioral use licensing for responsible ai. *arXiv preprint arXiv:2011.03116*.

David, D. (2020, December). Swahili: News classification dataset.

De la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, and M. Grandury (2022). BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural 68*, 13–23.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Dodge, J., M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. pp. 1286–1305.

Einea, O., A. Elnagar, and R. Al Debsi (2019). Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in Brief 25*, 104076.

El-Haj, M. (2020, May). Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 1318–1326. European Language Resources Association.

El-Haj, M. and R. Koulali (2013). Kalimat a multipurpose arabic corpus. In *Second workshop on Arabic corpus linguistics (WACL-2)*, pp. 22–25.

El-Khair, I. A. (2016). 1.5 billion words arabic corpus. *arXiv preprint arXiv:1611.04033*.

Elnagar, A., L. Lulu, and O. Einea (2018). An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia Computer Science 142*, 182–189. Arabic Computational Linguistics.

Fan, A., S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al. (2021). Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res. 22*(107), 1–48.

Galliano, S., E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri (2006, May). Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Gao, L., S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gokaslan, A. and V. Cohen (2019). Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Goldhahn, D., T. Eckart, and U. Quasthoff (2012, May). Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pp. 759–765. European Language Resources Association (ELRA).

He, W., K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Liu, Y. Wang, H. Wu, Q. She, X. Liu, T. Wu, and H. Wang (2018). Dureader: a chinese machine reading comprehension dataset from real-world applications.

Heafield, K. (2011, July). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, pp. 187–197. Association for Computational Linguistics.

Ho, V. A., D. H.-C. Nguyen, D. H. Nguyen, L. T.-V. Pham, D.-V. Nguyen, K. V. Nguyen, and N. L.-T. Nguyen (2020). Emotion recognition for vietnamese social media text.

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. v. d. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre (2022). Training compute-optimal large language models.

Howard, J. and S. Ruder (2018). Fine-tuned language models for text classification. *CoRR abs/1801.06146*.

Jawaid, B., A. Kamran, and O. Bojar (2014). Urdu monolingual corpus.

Jernite, Y., H. Nguyen, S. Biderman, A. Rogers, M. Masoud, V. Danchev, S. Tan, A. S. Luccioni, N. Subramani, G. Dupont, J. Dodge, K. Lo, Z. Talat, D. Radev, A. Gokaslan, S. Nikpoor, P. Henderson, R. Bommasani, and M. Mitchell (2022). Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, New York, NY, USA. Association for Computing Machinery.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov (2017, April). 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.

Kandpal, N., E. Wallace, and C. Raffel (2022). Deduplicating training data mitigates privacy risks in language models. *arXiv preprint arXiv:2202.06539*.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling laws for neural language models. *CoRR abs/2001.08361*.

Kermes, H., S. Degaetano-Ortlieb, A. Khamis, J. Knappen, and E. Teich (2016, May). The royal society corpus: From uncharted data to corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, pp. 1928–1931. European Language Resources Association (ELRA).

Kowsher, M., M. Uddin, A. Tahabilder, M. Ruhul Amin, M. F. Shahriar, and M. S. I. Sobuj (2021, September). Banglalm: Bangla corpus for language model research. Online. IEEE.

Kreutzer, J., I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. O. Suarez, I. Orife, K. Ogueji, R. Niyongabo, T. Nguyen, M. Müller, A. Müller, S. Muhammad, N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. Dossou, S. Dlamini, N. de Silva, S. Çabuk Ballı, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics 10*(0), 50–72.

Kudo, T. (2018, July). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp. 66–75. Association for Computational Linguistics.

Kunchukuttan, A. (2020). The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Kunchukuttan, A., D. Kakwani, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar (2020). Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Kunchukuttan, A., P. Mehta, and P. Bhattacharyya (2018, May). The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kurniawan, K. and S. Louvan (2018). Indosum: A new benchmark dataset for indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)*, pp. 215–220. IEEE.

Külebi, B. (2021, October). ParlamentParla - Speech corpus of Catalan Parliamentary sessions.

Leahy, C. and S. Biderman (2021). The hard problem of aligning AI to human values. In *The State of AI Ethics Report*, Volume 4, pp. 180–183. The Montreal AI Ethics Institute.

Lee, K., D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini (2022). Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lhoest, Q., A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush,

and T. Wolf (2021, November). Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online and Punta Cana, Dominican Republic, pp. 175–184. Association for Computational Linguistics.

Li, Y., D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. D. Lago, T. Hubert, P. Choy, C. d. M. d'Autume, I. Babuschkin, X. Chen, P.-S. Huang, J. Welbl, S. Gowal, A. Cherepanov, J. Molloy, D. J. Mankowitz, E. S. Robson, P. Kohli, N. de Freitas, K. Kavukcuoglu, and O. Vinyals (2022). Competition-level code generation with alphacode.

Lieber, O., O. Sharir, B. Lenz, and Y. Shoham (2021). Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*.

Lison, P. and J. Tiedemann (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Lo, K., L. L. Wang, M. Neumann, R. Kinney, and D. Weld (2020, July). S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 4969–4983. Association for Computational Linguistics.

Lopes, C. V., P. Maj, P. Martins, V. Saini, D. Yang, J. Zitny, H. Sajnani, and J. Vitek (2017). Déjàvu: a map of code duplicates on github. *Proceedings of the ACM on Programming Languages 1*(OOPSLA), 1–28.

Luccioni, A. S. and J. D. Viviano (2021). What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *Published in the Proceedings of ACL 2021*.

Mahendra, R., A. F. Aji, S. Louvan, F. Rahman, and C. Vania (2021). Indonli: A natural language inference dataset for indonesian. *arXiv preprint arXiv:2110.14566*.

Manber, U. and G. Myers (1993). Suffix arrays: A new method for on-line string searches. *SIAM Journal on Computing 22*(5), 935–948.

Manku, G. S., A. Jain, and A. Das Sarma (2007). Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, New York, NY, USA, pp. 141–150. Association for Computing Machinery.

Mayeesha, T. T., A. M. Sarwar, and R. M. Rahman (2020, November). Deep learning based question answering system in Bengali.

McMillan-Major, A., Z. Alyafeai, S. Biderman, K. Chen, F. De Toni, G. Dupont, H. Elsahar, C. Emezue, A. F. Aji, S. Ilić, N. Khamis, C. Leong, M. Masoud, A. Soroa, P. O. Suarez, Z. Talat, D. van Strien, and Y. Jernite (2022). Documenting geographically and contextually diverse data sources: The bigscience catalogue of language data and resources.

Moeljadi, D. (2012). Usage of indonesian possessive verbal predicates: A statistical analysis based on questionnaire and storytelling surveys. In *5th Conference on Austronesian and Papuan Languages and Linguistics (APLL5)*, SOAS, University of London.

Mohr, G., J. Kunze, and M. Stack (2008). The warc file format 1.0 (iso 28500).

Nekoto, W., V. Marivate, T. Matsila, T. E. Fasubaa, T. Fagbohungbe, S. O. Akinola, S. H. Muhammad, S. K. Kabenamualu, S. Osei, F. Sackey, R. A. Niyongabo, R. Macharm, P. Ogayo, O. Ahia, M. M. Berhe, M. Adeyemi, M. Mokgesi-Selinga, L. Okegbemi, L. Martinus, K. Tajudeen, K. Degila, K. Ogueji, K. Siminyu, J. Kreutzer, J. Webster, J. T. Ali, J. Z. Abbott, I. Orife, I. Ezeani, I. A. Dangana, H. Kamper, H. Elsahar, G. Duru, G. Kioko, E. Murhabazi, E. V. Biljon, D. Whitenack, C. Onyefuluchi, C. C. Emezue, B. F. P. Dossou, B. Sibanda, B. I. Bassey, A. Olabiyi, A. Ramkilowan, A. Öktem, A. Akinfaderin, and A. Bashir (2020). Participatory research for low-resourced machine translation: A case study in african languages. In T. Cohn, Y. He, and Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Volume EMNLP 2020 of *Findings of ACL*, pp. 2144–2160. Association for Computational Linguistics.

Ngo, C. and T. H. Trinh (2021). Styled augmented translation (sat). *https://github.com/vietai/SAT*.

Nguyen, K. V., V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L.-T. Nguyen (2018). Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 19–24.

Nguyen, T., H. Pham, M. Truong, H. Duc, and P. Tan (2021). Vietnamese poem generator. https://github.com/fsoft-ailab/Poem-Generator.

Nomoto, H., K. Okano, D. Moeljadi, and H. Sawada (2018). Tufs asian language parallel corpus (talpco). In *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*, pp. 436–439. Association for Natural Language Processing.

Ortiz Suárez, P. J., L. Romary, and B. Sagot (2020, July). A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 1703–1714. Association for Computational Linguistics.

Ortiz Suárez, P. J., B. Sagot, and L. Romary (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Mannheim, pp. 9 – 16. Leibniz-Institut für Deutsche Sprache.

Parida, S., O. Bojar, and S. R. Dash (2020). Odiencorp: Odia–english and odia-only corpus for machine translation. In *Smart Intelligent Computing and Applications*, pp. 495–504. Springer.

Pisceldo, F., R. Manurung, and M. Adriani (2009). Probabilistic part-of-speech tagging for bahasa indonesia. In *Third International Workshop on Malay and Indonesian Language Engineering (MALINDO)*, Suntec, Singapore.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners.

Rae, J. W., S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, H. F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. M. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. A. Hechtman, L. Weidinger, I. Gabriel, W. S. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving (2021). Scaling language models: Methods, analysis & insights from training gopher. *CoRR abs/2112.11446*.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research 21*, 1–67.

Rahman, M., E. Kumar Dey, et al. (2018). Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data 3*(2), 15.

Rahutomo, F. and A. Miqdad Muadz Muzad (2018). Indonesian news corpus.

Rajkomar, A., M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine 169*, 866–872.

Ramesh, G., S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, S. Nagaraj, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, and M. S. Khapra (2021). Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Rodriguez-Penagos, C. G. and C. Armentano-Oller (2021a, June). Enriched conllu ancora for ml training.

Rodriguez-Penagos, C. G. and C. Armentano-Oller (2021b, February). VilaQuAD: an extractive QA dataset from Catalan newswire.

Rodriguez-Penagos, C. G. and C. Armentano-Oller (2021c, February). ViquiQuAD: an extractive QA dataset from Catalan Wikipedia.

Rogers, A. (2021, August). Changing the world by changing the data. 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)*, Online, pp. 2182–2194. Association for Computational Linguistics.

Sajjad, H., A. Abdelali, N. Durrani, and F. Dalvi (2020, December). AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), pp. 5094–5107. International Committee on Computational Linguistics.

Sakti, S., E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura (2008). Development of Indonesian large vocabulary continuous speech recognition system within a-STAR project. In *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*.

Scheuerman, M. K., A. Hanna, and E. Denton (2021). Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction 5*, 1–37.

Shikali, S. C. and M. Refuoe (2019, November). Language modeling data for swahili.

Siripragada, S., J. Philip, V. P. Namboodiri, and C. V. Jawahar (2020, May). A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 3743–3751. European Language Resources Association.

Sloane, M., E. Moss, O. Awomolo, and L. Forlano (2020). Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423*.

Smith, S., M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zhang, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, and B. Catanzaro (2022). Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv*.

Talat, Z., A. Névéol, S. Biderman, M. Clinciu, M. Dey, S. Longpre, S. Luccioni, M. Masoud, M. Mitchell, D. Radev, S. Sharma, A. Subramonian, J. Tae, S. Tan, D. Tunuguntla, and O. van der Wal (2022). You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Challenges & Perspectives in Creating Large Language Models*.

Vuong, Q.-H., V.-P. La, T.-H. T. Nguyen, M.-H. Nguyen, T.-T. Le, and M.-T. Ho (2021). An ai-enabled approach in analyzing media data: An example from data on covid-19 news coverage in vietnam. *Data 6*(7).

Wenzek, G., M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and É. Grave (2020). Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 4003–4012.

Wibowo, H. A. (2020). Recibrew. https://github.com/haryoa/ingredbrew.

Winner, L. (2017). Do artifacts have politics? In *Computer Ethics*, pp. 177–192. Routledge.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR abs/2010.11934*.

Yuan, S., H. Zhao, Z. Du, M. Ding, X. Liu, Y. Cen, X. Zou, Z. Yang, and J. Tang (2021). Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open 2*, 65–68.

Zerrouki, T. and A. Balla (2017). Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. *Data in brief 11*, 147.

Zhang, B., P. Williams, I. Titov, and R. Sennrich (2020). Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Zhang, C., D. Ippolito, K. Lee, M. Jagielski, F. Tramèr, and N. Carlini (2021). Counterfactual memorization in neural language models. *arXiv preprint arXiv:2112.12938*.

Zhang, S., S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Ziemski, M., M. Junczys-Dowmunt, and B. Pouliquen (2016, May). The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, pp. 3530–3534. European Language Resources Association (ELRA).