Appendix

Contents

A 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.

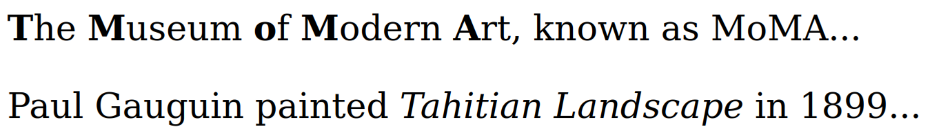
B Details on tools used to obtain crowdsourced dataset

B.1 Pseudocode to recreate the text structure from the HTML code

The HTML code of a web page provides information about the structure of the text. The final structure of a web page is, however, the one produced by the rendering engine of the web browser and any CSS instructions. The latter two elements, which can vary enormously from one situation to another, always use the tag types for their rendering rules (Figure 9. Therefore, we have used a fairly simple heuristic on tag types to reconstruct the structure of the text extracted from an HTML code. To reconstruct the text, the HTML DOM, which can be represented as a tree (Figure 10), is traversed with an depth-first search algorithm. The text is initially empty and each time a new node with textual content is reached its content is concatenated according to the rules presented in the Algorithm 1. Block-type tags are for us: <*address*>, <*article*>, <*aside*>, <*blockquote*>, <*body*>, <*br*>, <*button*>, <*canvas*>, <*caption*>, <*col*>, <*colgroup*>, <*dd*>, <*div*>, <*dl*>, <*dt*>, <*embed*>, <*fieldset*>, <*figcaption*>, <*figure*>, <*footer*>, <*form*>, <*h1*>, <*h2*>, <*h3*>, <*h4*>, <*h5*>, <*h6*>, <*header*>, <*hgroup*>, <*hr*>, <*li*>, <*map*>, <*noscript*>, <*object*>, <*ol*>, <*output*>, <*p*>, <*pre*>, <*progress*>, <*section*>, <*table*>, <*tbody*>, <*textarea*>, <*tfoot*>, <*th*>, <*thead*>, <*tr*>, <*ul*>, and <*video*>. Inline-type tags are for us: <*address*>, <*cite*>, <*details*>, <*datalist*>, <*iframe*>, <*img*>, <*input*>, <*label*>, <*legend*>, <*optgroup*>, <*q*>, <*select*>, <*summary*>, <*tbody*>, <*td*>, and <*time*>.

```
<html>
<body>
<div><p><b>T</b>he <b>M</b>useum <b>o</b>f <b>M</b>odern <b>A</b>rt,known as MoMA...</p><p>Paul Gauguin
painted <cite>Tahitian Landscape</cite> in 1899...</p></div>
</body>
</html>
```

(a) HTML code



The **M**useum **o**f **M**odern **A**rt, known as MoMA...

Paul Gauguin painted *Tahitian Landscape* in 1899...

(b) Web browser rendering

Figure 9: Example showing how a single line of HTML code is rendered by a browser's renderer. In this example, we can see that the tags <*p*> delimit different blocks which are therefore spaced by line breaks while other tags, such as <*cite*>, are rendered on the same line of text that precedes and follows them.

```
<div>
    <h1>Heading</h1>
    <p>
        p-inner
    </p>
    p-trailing
</div>
```

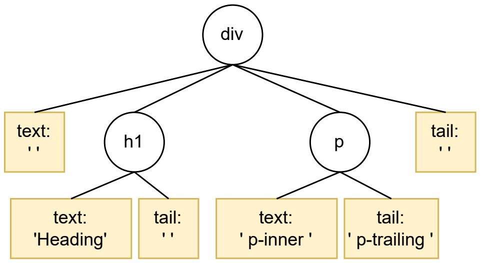


Figure 10: Simplified version of HTML DOM model on an example. Left: snippet of HTML code. Right: corresponding DOM. The yellow squares represent nodes with textual content.

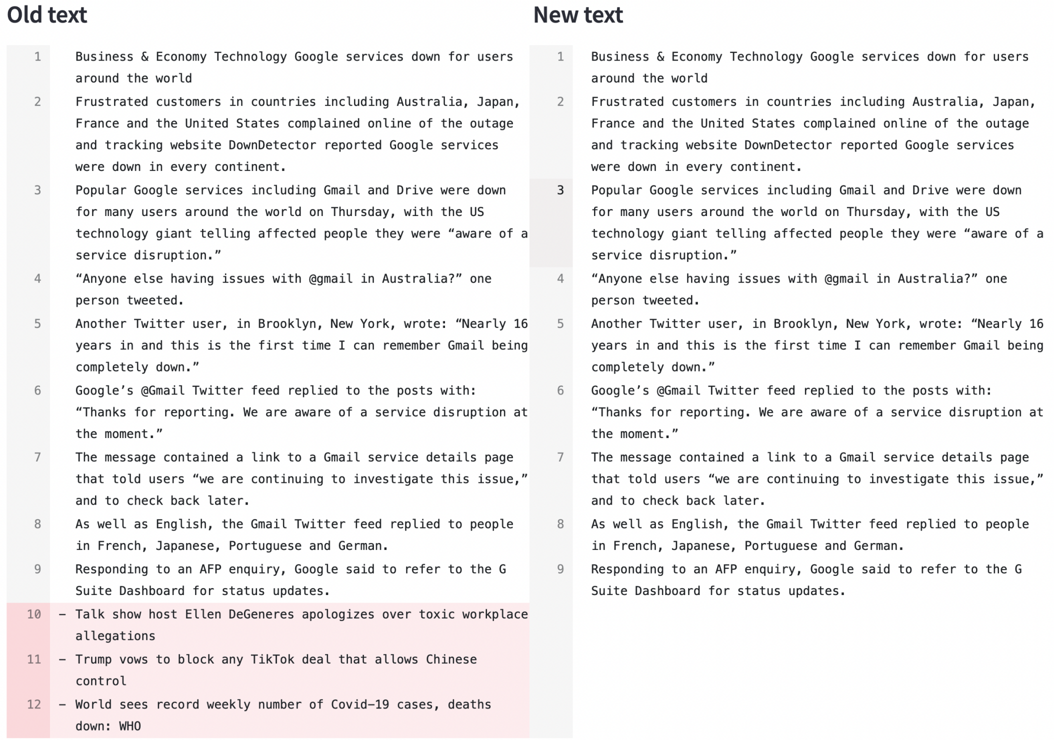
B.2 Visualization tool use cases

The visualisation tool was for us an iterative tool that we used to define new cleaning and filtering methods by visualising their effect on a subset of documents. This visualisation allowed us understand the impact of functions on the dataset at every stage of the processing pipeline (Figure 11 for advertisement detection for example), prompted us to adapt pipelines as well as introduce new functions for specific cases.

As a typical usage of the visualisation tool as a development tool, for documents coming from pseudo-crawls, we wanted to create a method to remove the parts of the documents that looked like a template, based on the principle that these templates would be identifiable by the fact that they would be repeated lines between documents. With the first version of the pipeline we could see from the estimates of the size of the final dataset (Figure 12) that a lot of content was removed. Looking at the examples (Figure 12c), we could confirm that a large part of the article text was removed. The

Algorithm 1 Pseudo-code to concatenate texts retrieved from the HTML DOM

```
$text \leftarrow$ empty string
for $new\_text$ in list of texts retrieved by the DFS traversal do
    if $new\_text$ is attached to a block-type tag then
        # Block elements are separated from the rest by a line break
        if $text$ ends with a breaking line then
            $text \leftarrow text + new\_text$
        else if $text$ ends with a space then
            $text \leftarrow text$ without end space
            $text \leftarrow text$ + breaking line + $new\_text$
        else
            $text \leftarrow text$ + breaking line + $new\_text$
        end if
    else if $new\_text$ is attached to a inline-type tag then
        #  Inline elements are separated from the rest by a line break or a space
        if $text$ ends with a space or a breaking line then
            $text \leftarrow text + new\_text$
        else
            $text \leftarrow text$ + space + $new\_text$
        end if
    else
        $text \leftarrow text + new\_text$
    end if
end for
```



Old text

```
Business & Economy Technology Google services down for users around the world
Frustrated customers in countries including Australia, Japan, France and the United States complained online of the outage and tracking website DownDetector reported Google services were down in every continent.
Popular Google services including Gmail and Drive were down for many users around the world on Thursday, with the US technology giant telling affected people they were “aware of a service disruption.”
“Anyone else having issues with @gmail in Australia?” one person tweeted.
Another Twitter user, in Brooklyn, New York, wrote: “Nearly 16 years in and this is the first time I can remember Gmail being completely down.”
Google’s @Gmail Twitter feed replied to the posts with: “Thanks for reporting. We are aware of a service disruption at the moment.”
The message contained a link to a Gmail service details page that told users “we are continuing to investigate this issue,” and to check back later.
As well as English, the Gmail Twitter feed replied to people in French, Japanese, Portuguese and German.
Responding to an AFP enquiry, Google said to refer to the G Suite Dashboard for status updates.
- Talk show host Ellen DeGeneres apologizes over toxic workplace allegations
- Trump vows to block any TikTok deal that allows Chinese control
- World sees record weekly number of Covid-19 cases, deaths down: WHO
```

New text

```
Business & Economy Technology Google services down for users around the world
Frustrated customers in countries including Australia, Japan, France and the United States complained online of the outage and tracking website DownDetector reported Google services were down in every continent.
Popular Google services including Gmail and Drive were down for many users around the world on Thursday, with the US technology giant telling affected people they were “aware of a service disruption.”
“Anyone else having issues with @gmail in Australia?” one person tweeted.
Another Twitter user, in Brooklyn, New York, wrote: “Nearly 16 years in and this is the first time I can remember Gmail being completely down.”
Google’s @Gmail Twitter feed replied to the posts with: “Thanks for reporting. We are aware of a service disruption at the moment.”
The message contained a link to a Gmail service details page that told users “we are continuing to investigate this issue,” and to check back later.
As well as English, the Gmail Twitter feed replied to people in French, Japanese, Portuguese and German.
Responding to an AFP enquiry, Google said to refer to the G Suite Dashboard for status updates.
```

Figure 11: Example of showing sample changes throughout each step of the processing pipeline. In the following example, users can notice that advertisement text were removed from the main article.

cause of this behaviour was due to the fact that the same article was appearing at several different URLs as the website hierarchy had changed between the different common crawl dumps. For the final pipeline, we therefore added a custom deduplication of the urls as a first operation to target this change of addresses. With the final pipeline developed, less content was removed. By manually inspecting the examples, we could observe that the content removed from the documents was indeed the one initially targeted.

B.3 Exhaustive list of functions used in (Crowd)Sourced dataset

We provide an exhaustive list of functions used in each of the processing pipeline for the crowdsourced dataset[14]:

replace_newline_with_space Takes in a batch of texts and for each text replaces the newline character "

n" with a single space.

remove_lines_with_code Takes in a batch of texts and removes lines with the following substrings: *"{", "}", "[if", "<script",*

remove_html_spans Takes in a batch of texts and removes lines with the following substrings: *"<span", "</span>", "<div", "<a", "</div>", "</a>", "br>",*

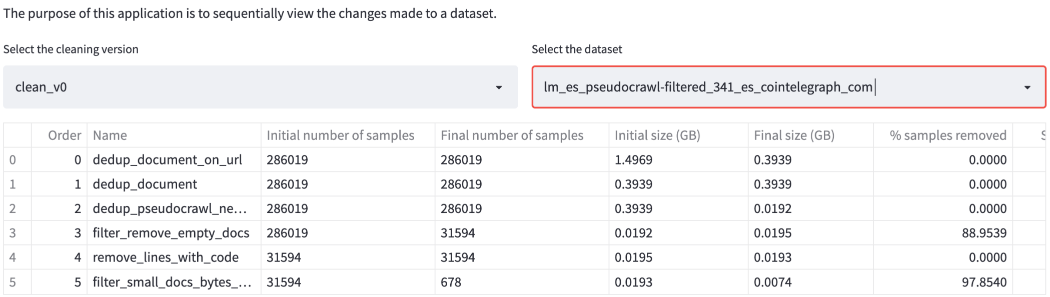
remove_html_spans_sanad Takes in a batch of texts and removes lines with the following substrings: *"<img", "]]>", "<![CDATA", "//DW", "var ", "xtImg", "To view this video please enable JavaScript",*

remove_wiki_mojibake Takes in a batch of texts and removes lines with the following substrings: *"À À"*

strip_substrings_en_wiktionary Takes in a batch of texts and removes the following substrings:

- *This entry needs pronunciation information*
- *Please try to find a suitable image on Wikimedia Commons or upload one there yourself!This entry need pronunciation information*
- *You may continue to edit this entry while the discussion proceeds, but please mention significant edits at the RFD discussion and ensure that the intention of votes already cast is not left unclear*
- *This entry is part of the phrasebook project, which presents criteria for inclusion based on utility, simplicity and commonality*
- *If you are a native speaker with a microphone, please record some and upload them*
- *If you are familiar with the IPA then please add some!*
- *Feel free to edit this entry as normal, but do not remove rfv until the request has been resolved*
- *This entry needs quotations to illustrate usage*
- *If you are familiar with the IPA then please add some!This entry needs audio files*
- *Please see that page for discussion and justifications*
- *If you are familiar with the IPA or enPR then please add some!A user has added this entry to requests for verification(+) If it cannot be verified that this term meets our attestation criteria, it will be deleted*
- *This entry needs a photograph or drawing for illustration*
- *A user has added this entry to requests for deletion(+)*
- *Do not remove the rfd until the debate has finished*
- *This entry needs audio files*
- *If you come across any interesting, durably archived quotes then please add them!This entry is part of the phrasebook project, which presents criteria for inclusion based on utility, simplicity and commonality*

[14] Code is available at https://github.com/bigscience-workshop/data-preparation/blob/main/preprocessing/training/clean.py



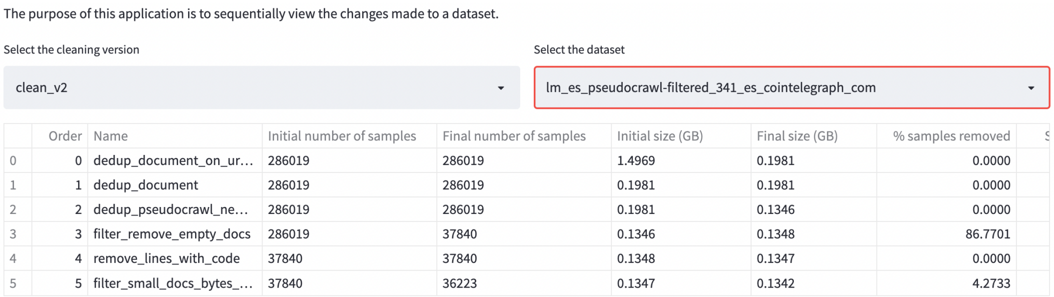
The purpose of this application is to sequentially view the changes made to a dataset.

Select the cleaning version: clean_v0

Select the dataset: lm_es_pseudocrawl-filtered_341_es_cointelegraph_com

	Order	Name	Initial number of samples	Final number of samples	Initial size (GB)	Final size (GB)	% samples removed
0	0	dedup_document_on_url	286019	286019	1.4969	0.3939	0.0000
1	1	dedup_document	286019	286019	0.3939	0.3939	0.0000
2	2	dedup_pseudocrawl_ne…	286019	286019	0.3939	0.0192	0.0000
3	3	filter_remove_empty_docs	286019	31594	0.0192	0.0195	88.9539
4	4	remove_lines_with_code	31594	31594	0.0195	0.0193	0.0000
5	5	filter_small_docs_bytes_…	31594	678	0.0193	0.0074	97.8540

(a) Pipeline v0



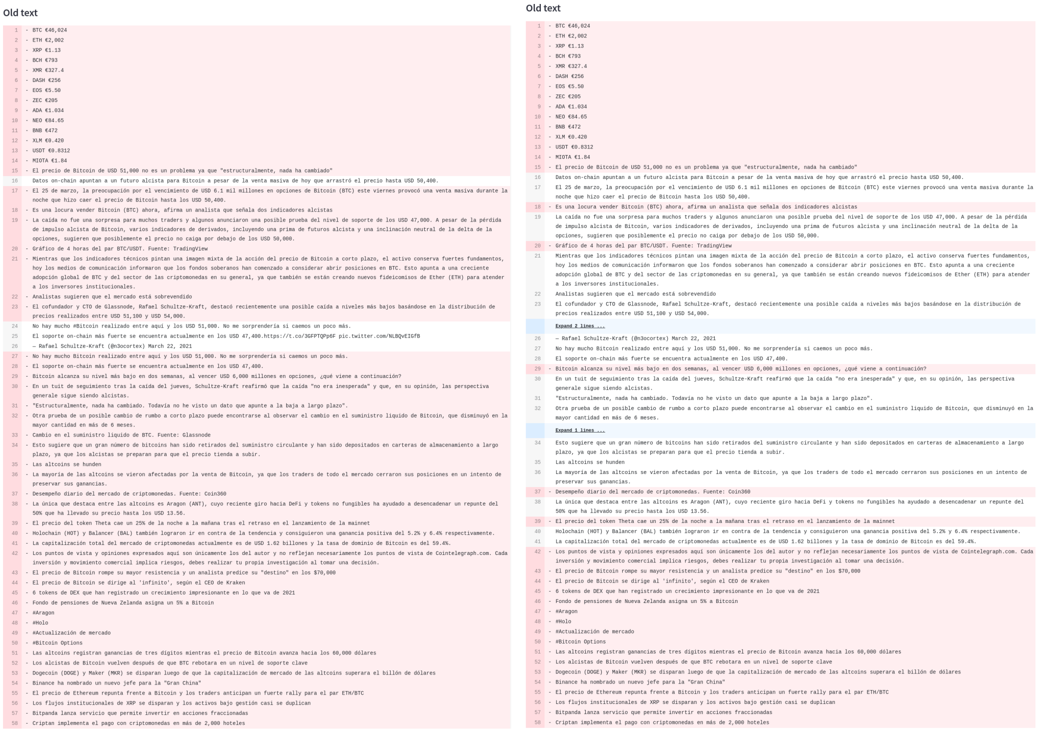
The purpose of this application is to sequentially view the changes made to a dataset.

Select the cleaning version: clean_v2

Select the dataset: lm_es_pseudocrawl-filtered_341_es_cointelegraph_com

	Order	Name	Initial number of samples	Final number of samples	Initial size (GB)	Final size (GB)	% samples removed
0	0	dedup_document_on_ur…	286019	286019	1.4969	0.1981	0.0000
1	1	dedup_document	286019	286019	0.1981	0.1981	0.0000
2	2	dedup_pseudocrawl_ne…	286019	286019	0.1981	0.1346	0.0000
3	3	filter_remove_empty_docs	286019	37840	0.1346	0.1348	86.7701
4	4	remove_lines_with_code	37840	37840	0.1348	0.1347	0.0000
5	5	filter_small_docs_bytes_…	37840	36223	0.1347	0.1342	4.2733

(b) Pipeline v2



Old text

```
- BTC €46,024
- ETH €2,002
- XRP €1.13
- BCH €793
- XMR €327.4
- DASH €256
- EOS €5.50
- ZEC €205
- ADA €1.034
- NEO €84.65
- BNB €472
- XLM €0.420
- USDT €0.8312
- MIOTA €1.84
- El precio de Bitcoin de USD 51,000 no es un problema ya que "estructuralmente, nada ha cambiado"
  Datos on-chain apuntan a un futuro alcista para Bitcoin a pesar de la venta masiva de hoy que arrastró el precio hasta USD 50,400.
- El 25 de marzo, la preocupación por el vencimiento de USD 6.1 mil millones en opciones de Bitcoin (BTC) este viernes provocó una venta masiva durante la
  noche que hizo caer el precio de Bitcoin hasta los USD 50,400.
- Es una locura vender Bitcoin (BTC) ahora, afirma un analista que señala dos indicadores alcistas
- La caída no fue una sorpresa para muchos traders y algunos anunciaron una posible prueba del nivel de soporte de los USD 47,000. A pesar de la pérdida
  de impulso alcista de Bitcoin, varios indicadores de derivados, incluyendo una prima de futuros alcista y una inclinación neutral de la delta de la
  opciones, sugieren que posiblemente el precio no caiga por debajo de los USD 50,000.
- Gráfico de 4 horas del par BTC/USDT. Fuente: TradingView
- Mientras que los indicadores técnicos pintan una imagen mixta de la acción del precio de Bitcoin a corto plazo, el activo conserva fuertes fundamentos,
  hoy los medios de comunicación informaron que los fondos soberanos han comenzado a considerar abrir posiciones en BTC. Esto apunta a una creciente
  adopción global de BTC y del sector de las criptomonedas en su general, ya que también se están creando nuevos fideicomisos de Ether (ETH) para atender
  a los inversores institucionales.
- Analistas sugieren que el mercado está sobrevendido
- El cofundador y CTO de Glassnode, Rafael Schultze-Kraft, destacó recientemente una posible caída a niveles más bajos basándose en la distribución de
  precios realizados entre USD 51,100 y USD 54,000.
  No hay mucho #Bitcoin realizado entre aquí y los USD 51,000. No me sorprendería si caemos un poco más.
  El soporte on-chain más fuerte se encuentra actualmente en los USD 47,400.https://t.co/3GFPTQPp6F pic.twitter.com/NLBQvEIGfB
  — Rafael Schultze-Kraft (@n3ocortex) March 22, 2021
- No hay mucho Bitcoin realizado entre aquí y los USD 51,000. No me sorprendería si caemos un poco más.
- El soporte on-chain más fuerte se encuentra actualmente en los USD 47,400.
- Bitcoin alcanza su nivel más bajo en dos semanas, al vencer USD 6,000 millones en opciones, ¿qué viene a continuación?
- En un tuit de seguimiento tras la caída del jueves, Schultze-Kraft reafirmó que la caída "no era inesperada" y que, en su opinión, las perspectiva
  generale sigue siendo alcistas.
- "Estructuralmente, nada ha cambiado. Todavía no he visto un dato que apunte a la baja a largo plazo".
- Otra prueba de un posible cambio de rumbo a corto plazo puede encontrarse al observar el cambio en el suministro liquido de Bitcoin, que disminuyó en la
  mayor cantidad en más de 6 meses.
- Cambio en el suministro liquido de BTC. Fuente: Glassnode
- Esto sugiere que un gran número de bitcoins han sido retirados del suministro circulante y han sido depositados en carteras de almacenamiento a largo
  plazo, ya que los alcistas se preparan para que el precio tienda a subir.
- Las altcoins se hunden
- La mayoría de las altcoins se vieron afectadas por la venta de Bitcoin, ya que los traders de todo el mercado cerraron sus posiciones en un intento de
  preservar sus ganancias.
- Desempeño diario del mercado de criptomonedas. Fuente: Coin360
- La única que destaca entre las altcoins es Aragon (ANT), cuyo reciente giro hacia DeFi y tokens no fungibles ha ayudado a desencadenar un repunte del
  50% que ha llevado su precio hasta los USD 13.56.
- El precio del token Theta cae un 25% de la noche a la mañana tras el retraso en el lanzamiento de la mainnet
- Holochain (HOT) y Balancer (BAL) también lograron ir en contra de la tendencia y consiguieron una ganancia positiva del 5.2% y 6.4% respectivamente.
- La capitalización total del mercado de criptomonedas actualmente es de USD 1.62 billones y la tasa de dominio de Bitcoin es del 59.4%.
- Los puntos de vista y opiniones expresados aquí son únicamente los del autor y no reflejan necesariamente los puntos de vista de Cointelegraph.com. Cada
  inversión y movimiento comercial implica riesgos, debes realizar tu propia investigación al tomar una decisión.
- El precio de Bitcoin rompe su mayor resistencia y un analista predice su "destino" en los $70,000
- El precio de Bitcoin se dirige al 'infinito', según el CEO de Kraken
- 6 tokens de DEX que han registrado un crecimiento impresionante en lo que va de 2021
- Fondo de pensiones de Nueva Zelanda asigna un 5% a Bitcoin
- #Aragon
- #Holo
- #Actualización de mercado
- #Bitcoin Options
- Las altcoins registran ganancias de tres dígitos mientras el precio de Bitcoin avanza hacia los 60,000 dólares
- Los alcistas de Bitcoin vuelven después de que BTC rebotara en un nivel de soporte clave
- Dogecoin (DOGE) y Maker (MKR) se disparan luego de que la capitalización de mercado de las altcoins superara el billón de dólares
- Binance ha nombrado un nuevo jefe para la "Gran China"
- El precio de Ethereum repunta frente a Bitcoin y los traders anticipan un fuerte rally para el par ETH/BTC
- Los flujos institucionales de XRP se disparan y los activos bajo gestión casi se duplican
- Bitpanda lanza servicio que permite invertir en acciones fraccionadas
- Criptan implementa el pago con criptomonedas en más de 2,000 hoteles
```

Old text

```
- BTC €46,024
- ETH €2,002
- XRP €1.13
- BCH €793
- XMR €327.4
- DASH €256
- EOS €5.50
- ZEC €205
- ADA €1.034
- NEO €84.65
- BNB €472
- XLM €0.420
- USDT €0.8312
- MIOTA €1.84
- El precio de Bitcoin de USD 51,000 no es un problema ya que "estructuralmente, nada ha cambiado"
  Datos on-chain apuntan a un futuro alcista para Bitcoin a pesar de la venta masiva de hoy que arrastró el precio hasta USD 50,400.
  El 25 de marzo, la preocupación por el vencimiento de USD 6.1 mil millones en opciones de Bitcoin (BTC) este viernes provocó una venta masiva durante la
  noche que hizo caer el precio de Bitcoin hasta los USD 50,400.
- Es una locura vender Bitcoin (BTC) ahora, afirma un analista que señala dos indicadores alcistas
  La caída no fue una sorpresa para muchos traders y algunos anunciaron una posible prueba del nivel de soporte de los USD 47,000. A pesar de la pérdida
  de impulso alcista de Bitcoin, varios indicadores de derivados, incluyendo una prima de futuros alcista y una inclinación neutral de la delta de la
  opciones, sugieren que posiblemente el precio no caiga por debajo de los USD 50,000.
- Gráfico de 4 horas del par BTC/USDT. Fuente: TradingView
  Mientras que los indicadores técnicos pintan una imagen mixta de la acción del precio de Bitcoin a corto plazo, el activo conserva fuertes fundamentos,
  hoy los medios de comunicación informaron que los fondos soberanos han comenzado a considerar abrir posiciones en BTC. Esto apunta a una creciente
  adopción global de BTC y del sector de las criptomonedas en su general, ya que también se están creando nuevos fideicomisos de Ether (ETH) para atender
  a los inversores institucionales.
  Analistas sugieren que el mercado está sobrevendido
  El cofundador y CTO de Glassnode, Rafael Schultze-Kraft, destacó recientemente una posible caída a niveles más bajos basándose en la distribución de
  precios realizados entre USD 51,100 y USD 54,000.
  Expand 2 lines ...
  — Rafael Schultze-Kraft (@n3ocortex) March 22, 2021
  No hay mucho Bitcoin realizado entre aquí y los USD 51,000. No me sorprendería si caemos un poco más.
  El soporte on-chain más fuerte se encuentra actualmente en los USD 47,400.
- Bitcoin alcanza su nivel más bajo en dos semanas, al vencer USD 6,000 millones en opciones, ¿qué viene a continuación?
  En un tuit de seguimiento tras la caída del jueves, Schultze-Kraft reafirmó que la caída "no era inesperada" y que, en su opinión, las perspectiva
  generale sigue siendo alcistas.
  "Estructuralmente, nada ha cambiado. Todavía no he visto un dato que apunte a la baja a largo plazo".
  Otra prueba de un posible cambio de rumbo a corto plazo puede encontrarse al observar el cambio en el suministro liquido de Bitcoin, que disminuyó en la
  mayor cantidad en más de 6 meses.
  Expand 1 lines ...
  Esto sugiere que un gran número de bitcoins han sido retirados del suministro circulante y han sido depositados en carteras de almacenamiento a largo
  plazo, ya que los alcistas se preparan para que el precio tienda a subir.
  Las altcoins se hunden
  La mayoría de las altcoins se vieron afectadas por la venta de Bitcoin, ya que los traders de todo el mercado cerraron sus posiciones en un intento de
  preservar sus ganancias.
- Desempeño diario del mercado de criptomonedas. Fuente: Coin360
  La única que destaca entre las altcoins es Aragon (ANT), cuyo reciente giro hacia DeFi y tokens no fungibles ha ayudado a desencadenar un repunte del
  50% que ha llevado su precio hasta los USD 13.56.
- El precio del token Theta cae un 25% de la noche a la mañana tras el retraso en el lanzamiento de la mainnet
  Holochain (HOT) y Balancer (BAL) también lograron ir en contra de la tendencia y consiguieron una ganancia positiva del 5.2% y 6.4% respectivamente.
  La capitalización total del mercado de criptomonedas actualmente es de USD 1.62 billones y la tasa de dominio de Bitcoin es del 59.4%.
- Los puntos de vista y opiniones expresados aquí son únicamente los del autor y no reflejan necesariamente los puntos de vista de Cointelegraph.com. Cada
  inversión y movimiento comercial implica riesgos, debes realizar tu propia investigación al tomar una decisión.
- El precio de Bitcoin rompe su mayor resistencia y un analista predice su "destino" en los $70,000
- El precio de Bitcoin se dirige al 'infinito', según el CEO de Kraken
- 6 tokens de DEX que han registrado un crecimiento impresionante en lo que va de 2021
- Fondo de pensiones de Nueva Zelanda asigna un 5% a Bitcoin
- #Aragon
- #Holo
- #Actualización de mercado
- #Bitcoin Options
- Las altcoins registran ganancias de tres dígitos mientras el precio de Bitcoin avanza hacia los 60,000 dólares
- Los alcistas de Bitcoin vuelven después de que BTC rebotara en un nivel de soporte clave
- Dogecoin (DOGE) y Maker (MKR) se disparan luego de que la capitalización de mercado de las altcoins superara el billón de dólares
- Binance ha nombrado un nuevo jefe para la "Gran China"
- El precio de Ethereum repunta frente a Bitcoin y los traders anticipan un fuerte rally para el par ETH/BTC
- Los flujos institucionales de XRP se disparan y los activos bajo gestión casi se duplican
- Bitpanda lanza servicio que permite invertir en acciones fraccionadas
- Criptan implementa el pago con criptomonedas en más de 2,000 hoteles
```

(c) Sample example difference between pipeline versions

Figure 12: High level statistics between two seperate pipelines and a sample example of the difference between two pipelines. First iteration (Figure 12a) generated a 7Mb dataset. After some careful tweaking, and some observed samples, we proposed a new pipeline in order to preserve previously wrongly removed data (Figure 12b) which generated a 134Mb dataset (x18). A example sample is available in Figure 12c

- *(For audio required quickly, visit WT:APR)*

remove_references_{lang} Removes lines that do not contain a minimum ratio of stopwords, as defined for each language[15]. Note, currently does not support languages with different segmentation (e.g. Chinese). Designed for academic datasets.

split_sentences_{lang} Builds a sentence splitter depending on the language passed: For Arabic, Catalan, Basque, Indonesian, and Chinese (both simplified and traditional), we use the Stanza tokenizer (Qi et al., 2020). For English, French, Portuguese, and Spanish, we use the NLTK tokenizer (Bird et al., 2009). For Bengalic, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and Telugu, we use the Indic NLP library tokenizer (Kunchukuttan, 2020). For Vietnamese, we use the Underthesea tokenizer [16].

filter_remove_empty_docs Removes documents that have a length of 0 when whitespace is removed.

filter_wiki_user_titles Removes documents where the Wikimedia metadata title starts with *"user"*,

filter_wiki_non_text_type Removes documents where the Wikimedia metadata type is not "text"

filter_small_docs Discards documents with less than 15 words. Tokenization is done via whitespace tokenizer.

filter_small_docs_bytes_{i} Discards documents with less than either 300 or 1024 bytes of text

dedup_template_soft Removes lines that are a minimum of 15 characters long and occur 10 or more times.

dedup_pseudocrawl_newspapers Removes lines that occur 2 or more times.

dedup_document Removes duplicate documents ignoring whitespaces and punctuation so only keeping characters and keeps one occurrence.

dedup_document_on_url Removes duplicate documents based on matched url while ignoring query parameters and keeps one occurrence.

dedup_document_on_url_lm_es_pseudocrawl-filtered_341_es_cointelegraph_com Removes duplicate documents based on the normalized urls (e.g., $URL and $URL/amp are treated as the same) without the query parameters and keeps one occurrence.

dedup_document_on_url_lm_en_pseudocrawl_filtered_619_www_qut_edu_au Removes duplicate documents based on the url without query parameters except for the "id" and "new-id" query parameters. The "new-id" query parameter is changed into a simple "id" parameter.

concatenate_lm_fr_ester Concatenate the text sorted by the id number in the metadata.

C Exhaustive list of human curated filters used on OSCAR

Before performing the filtering step, we did a cleaning step to modify the documents by standardizing whitespace and removing links, non-printable characters, and long words beyond a character threshold. These steps were designed to remove "non natural" language parts of the document (i.e. texts that are machine generated or not language, such as URLs).

Many of these filters require to split a document into words. For Chinese, we used the SentencePiece unigram tokenizer. For Vietnamese, since a word can be composed of two or three sub-words separated by spaces, we augmented the list of space separated tokens by the list of two and three consecutive space separated tokens.

Filter on number of words We discarded documents with too few words, as they often contain incorrect sentences, or contain no context for a model to learn correctly.

Filter on character repetition ratio To remove documents containing many repetitions, for a given n (determined in practice according to the language by native speakers), we counted the occurrence of each *character* n-gram present in the document. We defined the character repetition ratio as the ratio

[15] https://github.com/bigscience-workshop/catalogue_data/blob/master/clean_helpers/stopwords.py

[16] https://github.com/undertheseanlp/underthesea

of the sum of the k largest occurrences by the sum of all occurrences, and we discarded documents with a too high ratio.
If $k = 1$, short sentences are much more likely to have a high character repetition ratio, since the most frequent n-gram represents a larger proportion of the sentence. If k is the number of occurrences greater than or equal to 2, very long documents, but not necessarily including repetitions, tend to have a high character repetition ratio, since these texts inherently have a wide diversity of n-grams. We found that $k = \lfloor\sqrt{N}\rfloor$, with N the number of different n-grams found in the document, counterbalances well this effect in practice.

Example: Take the sentence "`ok_ok_good_ok`" and $n = 3$. Character n-grams, with their frequencies, are given in the following table.

`ok_`	`_ok`	`k_o`	`k_g`	`_go`	`goo`	`ood`	`od_`	`d_o`
2	2	1	1	1	1	1	1	1

Since we have 9 different character n-grams, $N = 9$ and $k = \lfloor\sqrt{N}\rfloor = 3$.

The sum of the k largest occurrences is $2 + 2 + 1 = 5$ and the sum of all occurrences is 11. Thus, the character repetition ratio for this sentence is $\frac{5}{11}$.

Filter on word repetition ratio As a complement to the previous filter, we remove documents that have commonly repeated similar long sentences. More specifically, we create a filter for the repetitions by looking this time at the occurrences of the *word* n-grams, for a chosen n parameter. We define the word repetition ratio as the ratio of the sum of the occurrences greater than or equal to 2 to the sum of all occurrences, and we discard documents with too high of a ratio. Contrary to the filter on the character repetition ratios, we did not find a bias of this method giving systematically higher or lower scores to longer or short documents. This filter is more robust in finding documents with long exact duplicated sentences in them, while the previous one is used to find short to medium sized repetitions.

Filter on special character ratio We established a list of special characters, including emojis, and simply discard documents with a special character ratio above a certain threshold.

Filter on closed class word ratio We found that having a low closed class word ratio in a document was one of the best indicators of a non-human generated content. We built lists of closed class words for each language by taking pre-existing lists, for example from Universal Dependencies[17], which were then reviewed by native speakers. We discard documents with a too low closed class word ratio.

Filter on flagged word ratio To limit the over-representation of pornographic documents, which are in practice much more likely to have shocking and sexist content, and to contain only buzzwords for SEO, we built lists of flagged words for each language by gathering existing lists, and filtering them by native speakers with precise instructions. We are then able to compute the flagged word ratio of a document and discard it if it is too high. About 1% of the documents for each language are removed by this filter.

Instructions for building the lists of flagged words: Keep only the words associated with porn and systematically used in a sexual context. Remove words that can be used in medical, scientific, colloquial (without referring systematically to porn), or everyday contexts. Remove all insults. Remove all words referring to race or sexual orientation.

Filter on language identification prediction score We used fastText (Joulin et al., 2017) to perform language identification and getting confidence scores for each document. If a score is below a specific threshold, we discard the document. We chose to eliminate few documents with this filter, because the language identification does not perform as well on low-resource languages.

Filter on perplexity score Following Wenzek et al. (2020), we trained SentencePiece unigram tokenizers (Kudo, 2018) followed by KenLM 5-gram models after tokenization (Heafield, 2011) on Wikipedia article openings for every language that was extracted from OSCAR. As in De la Rosa

[17] https://universaldependencies.org/

et al. (2022), we discarded documents to move the perplexity distribution towards the median, to avoid too high perplexity scores (deemed as not useful for the model), but subsampling was done by perplexity thresholding, not by reshaping the distribution as in De la Rosa et al. (2022). This thresholding was done lightly, by having native speakers manually establish the cutoff values per language[18], so as not to be too biased by the Wikipedia content and keep the dataset diverse.

D PII filtering initiative

Even if not eventually used in our final pipeline, we have released `muliwai`[19] a library for text pre-processing, augmentation, anonymization, and synthesis. It relies on transformer models and back-translation to perform NER and associated augmentation and anonymization over 100+ languages (i.e., we rely on XLMRoberta Fan et al. (2021) and M2M100 Conneau et al. (2020)). We either use a specific model for the chosen language or a model with cross-lingual capabilities. `Muliwai` tags using the aforementioned transformer then translate the sentence to a target language (e.g., English) and test to see if the translation preserves the NER tagging and discounts or increases the weight of a NER decision accordingly. It then performs NER in the target language and back translates to the source language. Finally it matches the translated sentence to the original sentence to determine which text spans in the source language sentence should be NER tagged based on the target language NER. We also use spacy and regex as added signals for NER tags.

We also include in the library specific regexes for detecting age, email, date, time, personal addresses, phone numbers and government-issued identifiers (such as license plates). Some regex matches use also the surrounding text context to improve precision.

However, the scale of the data, the fact that the impact on the resulting text could not be fully assessed in terms of language modeling and the time constraint due to compute allocation, meant this approach could not be operationalized on ROOTS. Instead we fell back to a simpler approach, see Section 3.3.

E Data Sources

Dataset	Language	Source
AraBench	ar	Sajjad et al. (2020)
1.5 billion words Arabic Corpus	ar	El-Khair (2016)
BanglaLM	bn	Kowsher et al. (2021)
bangla sentiment classification datasets	bn	Rahman et al. (2018)
Question answering in Bengali	bn	Mayeesha et al. (2020)
Binhvq News Corpus	vi	Bình (2021)
Books by Book Dash	en, fr, xh, zu	https://bookdash.org/books/
Bloom Library	ak, bm, fon, ki, lg, ln, nso, rw, st, sw, tn, ts, xh, zu	bloomlibrary.org
BRAD 2.0	ar	Elnagar et al. (2018)
brWaC Corpus	pt	
EusCrawl	eu	Artetxe et al. (2022)
Catalan General Crawling	ca	Armengol-Estapé et al. (2021)
Catalan Government Crawling	ca	Armengol-Estapé et al. (2021)

[18] Native speakers used an ad-hoc visualization tool built for the occasion: https://huggingface.co/spaces/huggingface/text-data-filtering

[19] Pronounced *"mu-lee-why"*, Hawaiian for river. https://github.com/ontocord/muliwai/tree/main

Dataset	Language	Source
Catalan Textual Corpus	ca	Armengol-Estapé et al. (2021)
Data on COVID-19 News Coverage in Vietnam	vi	Vuong et al. (2021)
DuReader	zhs	He et al. (2018)
Enriched CONLLU Ancora for ML training	ca	Rodriguez-Penagos and Armentano-Oller (2021a)
ESTER	fr	Galliano et al. (2006)
Github	code	Github on BigQuery
Habibi	ar	El-Haj (2020)
HAL	fr	hal.archives-ouvertes.fr/
IIT Bombay English-Hindi Parallel Corpus	hi	Kunchukuttan et al. (2018)
IndicNLP Corpus	bn, gu, hi, kn, ml, mr, or, pa, ta, te	Kunchukuttan et al. (2020)
Indo4B BPPT	id	Budiono et al. (2009)
Indo4B OPUS JW300	id	Agić and Vulić (2019)
Indo4B Kompas	id	Sakti et al. (2008)
Indo4B Parallel Corpus	id	Pisceldo et al. (2009)
Indo4B TALPCo	id	Nomoto et al. (2018)
Indo4B Tempo	id	Sakti et al. (2008)
Indonesian Frog Storytelling corpus	id	Moeljadi (2012)
Indonesian News Articles Published at 2017	id	Ashari (2018)
Indonesian News Corpus	id	Rahutomo and Miqdad Muadz Muzad (2018)
IndoNLI	id	Mahendra et al. (2021)
Indosum	id	Kurniawan and Louvan (2018)
KALIMAT	ar	El-Haj and Koulali (2013)
KSUCCA	ar	Alrabiah et al. (2013)
LABR	ar	Aly and Atiya (2013)
Language modeling data for Swahili	sw	Shikali and Refuoe (2019)
Leipzig Corpora Collection	ur	Goldhahn et al. (2012)
Mann Ki Baat	bn, gu, hi, ml, mr, or, ta, te, ur	Siripragada et al. (2020)
Masakhaner	ig, lg, rw, sw, wo, yo	Adelani et al. (2021)
MultiUN v2	en, ar, es, fr, zhs	Chen and Eisele (2012)
Odiencorp	en, or	Parida et al. (2020)
Opensubtitles2016	ca, en, ar, es, eu, fr, id, bn, hi, ml, ta, te, ur, pt, vi, zhs	Lison and Tiedemann (2016)
OpenITI proc	ar	Belinkov et al. (2019)
OPUS-100	ca, ar, eu, id, as, bn, gu, hi, ig, kn, ml, mr, or, pa, rw, ta, te, ur, pt, vi, xh, yo, zu	Zhang et al. (2020)
ParlamentParla	ca	Külebi (2021)
PIB	bn, gu, hi, ml, mr, or, pa, ta, te, ur	Siripragada et al. (2020)

Dataset	Language	Source
Project Gutenberg	en, es, fr, pt, zhs	gutenberg.org
QED (formely AMARA Corpus)	ar, en, es, fr, hi, pt, zhs, zht	Abdelali et al. (2014)
Recibrew	id	Wibowo (2020)
The Royal Society Corpus	en	Kermes et al. (2016)
S2ORC	en	Lo et al. (2020)
Samanantar	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	Ramesh et al. (2021)
SANAD	ar	Einea et al. (2019)
SciELO	en, es, pt	scielo.org
Stack Exchange	code	Gao et al. (2020)
Swahili News Classification Dataset	sw	David (2020)
Tashkeela	ar	Zerrouki and Balla (2017)
TeCLa	ca	Carrino et al. (2021)
WIT^3	ca, ar, en, es, eu, fr, id, as, bn, gu, hi, kn, ml, mr, pa, sw, ta, te, ur, pt, vi, zhs	Cettolo et al. (2012)
The Pile: EuroParl	en, es, fr, pt	Gao et al. (2020)
The Pile: USPTO	en	Gao et al. (2020)
UIT-VSMEC	vi	Ho et al. (2020)
United Nations Parallel Corpus	ar, en, es, fr, zhs	Ziemski et al. (2016)
Unsupervised Cross-lingual Representation Learning at Scale Common Crawl Corpus	ne	Conneau et al. (2020)
Urdu Monolingual Corpus	ur	Jawaid et al. (2014)
VietAI SAT	vi	Ngo and Trinh (2021)
Vietnamese Poetry Corpus	vi	Nguyen et al. (2021)
UIT-VSFC	vi	Nguyen et al. (2018)
VilaQuAD	ca	Rodriguez-Penagos and Armentano-Oller (2021b)
VinBigdata-VSLP ASR Challenge 2020	vi	institute.vinbigdata.org/events/vinbigdata-chia-se-100-gio-du-lieu-tieng-noi-cho-cong-dong/
VinBigdata-VSLP Monolingual Corpus 2020	vi	institute.vinbigdata.org/events/vinbigdata-chia-se-100-gio-du-lieu-tieng-noi-cho-cong-dong/
VinBigdata-VSLP Bilingual Corpus 2020	vi	institute.vinbigdata.org/events/vinbigdata-chia-se-100-gio-du-lieu-tieng-noi-cho-cong-dong/
ViquiQuAD	ca	Rodriguez-Penagos and Armentano-Oller (2021c)
VNTQcorpus(big)	vi	http://viet.jnlp.org/download-du-lieu-tu-vung-corpus
Wikibooks	ca, ar, en, es, eu, fr, id, bn, hi, ml, mr, pa, ta, te, ur, pt, vi, zhs	wikibooks.org
Wikimedia	ca, id, hi, pt	wikimedia.org

Dataset	**Language**	**Source**
Wikinews	ar, ca, en, es, fr, ta, pt, zhs	wikinews.org
Wikipedia	ak, ar, as, bm, bn, ca, en, es, eu, fr, id, ig, gu, hi, ki, kn, lg, ln, ml, mr, nso, ny, or, pa, pt, rn, rw, sn, st, sw, ta, te, tn, ts, tum, tw, ur, vi, wo, yo, zhs, zht, zu	wikipedia.org
Wikiquote	ar, ca, en, es, eu, fr, id, gu, hi, kn, ml, mr, ta, te, ur, pt, vi, zhs	wikiquote.org
Wikisource	ar, ca, es, eu, fr, id, as, bn, gu, hi, kn, ml, mr, or, pa, ta, te, pt, vi	wikisource.org
Wikiversity	ar, en, es, fr, hi, pt, zhs	wikiversity.org
Wikivoyage	en, es, fr, bn, hi, pt, vi, zhs	wikivoyage.org
Wiktionary	ar, ca, en, es, eu, fr, id, as, bn, gu, hi, kn, ml, mr, or, pa, ta, te, ur, pt, vi	wiktionary.org
WuDaoCorpora	zhs	Yuan et al. (2021)
XQUAD-ca	ca	Armengol-Estapé et al. (2021)

Table 2: List of datasets used in crowdsourced dataset.

Language	ISO-639-3	catalog-ref	Genus	Family	Macroarea	Size in Bytes
Akan	aka	ak	Kwa	Niger-Congo	Africa	70,1554
Arabic	arb	ar	Semitic	Afro-Asiatic	Eurasia	74,854,900,600
Assamese	asm	as	Indic	Indo-European	Eurasia	291,522,098
Bambara	bam	bm	Western Mande	Mande	Africa	391,747
Basque	eus	eu	Basque	Basque	Eurasia	2,360,470,848
Bengali	ben	bn	Indic	Indo-European	Eurasia	18,606,823,104
Catalan	cat	ca	Romance	Indo-European	Eurasia	17,792,493,289
Chi Chewa	nya	ny	Bantoid	Niger-Congo	Africa	1,187,405
Chi Shona	sna	sn	Bantoid	Niger-Congo	Africa	6,638,639
Chi Tumbuka	tum	tum	Bantoid	Niger-Congo	Africa	170,360
English	eng	en	Germanic	Indo-European	Eurasia	484,953,009,124
Fon	fon	fon	Kwa	Niger-Congo	Africa	2,478,546
French	fra	fr	Romance	Indo-European	Eurasia	208,242,620,434
Gujarati	guj	gu	Indic	Indo-European	Eurasia	1,199,986,460
Hindi	hin	hi	Indic	Indo-European	Eurasia	24,622,119,985
Igbo	ibo	ig	Igboid	Niger-Congo	Africa	14078,521
Indonesian	ind	id	Malayo-Sumbawan	Austronesian	Papunesia	19,972,325,222
Isi Zulu	zul	zu	Bantoid	Niger-Congo	Africa	8,511,561
Kannada	kan	kn	Southern Dravidian	Dravidian	Eurasia	2,098,453,560
Kikuyu	kik	ki	Bantoid	Niger-Congo	Africa	359,615
Kinyarwanda	kin	rw	Bantoid	Niger-Congo	Africa	40,428,299
Kirundi	run	rn	Bantoid	Niger-Congo	Africa	3,272,550
Lingala	lin	ln	Bantoid	Niger-Congo	Africa	1,650,804
Luganda	lug	lg	Bantoid	Niger-Congo	Africa	4,568,367
Malayalam	mal	ml	Southern Dravidian	Dravidian	Eurasia	3,662,571,498
Marathi	mar	mr	Indic	Indo-European	Eurasia	1,775,483,122
Nepali	nep	ne	Indic	Indo-European	Eurasia	2,551,307,393
Northern Sotho	nso	nso	Bantoid	Niger-Congo	Africa	1,764,506
Odia	ori	or	Indic	Indo-European	Eurasia	1,157,100,133
Portuguese	por	pt	Romance	Indo-European	Eurasia	79,277,543,375
Punjabi	pan	pa	Indic	Indo-European	Eurasia	1,572,109,752
Sesotho	sot	st	Bantoid	Niger-Congo	Africa	751,034
Setswana	tsn	tn	Bantoid	Niger-Congo	Africa	1,502,200
Simplified Chinese	—	zhs	Chinese	Sino-Tibetan	Eurasia	261,019,433,892
Spanish	spa	es	Romance	Indo-European	Eurasia	175,098,365,045
Swahili	swh	sw	Bantoid	Niger-Congo	Africa	236,482,543
Tamil	tam	ta	Southern Dravidian	Dravidian	Eurasia	7,989,206,220
Telugu	tel	te	South-Central Dravidian	Dravidian	Eurasia	2993407,159
Traditional Chinese	—	zht	Chinese	Sino-Tibetan	Eurasia	762,489,150
Twi	twi	tw	Kwa	Niger-Congo	Africa	1,265,041
Urdu	urd	ur	Indic	Indo-European	Eurasia	2,781,329,959
Vietnamese	vie	vi	Viet-Muong	Austro-Asiatic	Eurasia	43,709,279,959
Wolof	wol	wo	Wolof	Niger-Congo	Africa	3,606,973
Xhosa	xho	xh	Bantoid	Niger-Congo	Africa	14,304,074
Xitsonga	tso	ts	Bantoid	Niger-Congo	Africa	707,634
Yoruba	yor	yo	Defoid	Niger-Congo	Africa	89,695,835
Programming Languages	—	—	—	—		174,700,245,772

Table 3: Linguistic makeup of the corpus.

Language	**Number of Pseudocrawled Domains**	**Size in Bytes**
Spanish	108	29,440,210,712
English	22	4,537,031,408
Swahili	5	109,110,002
Indonesian	4	770,023,233
Basque	4	281,610,312
French	3	1,416,682,404
Hindi	2	1,536,649,276
Simplified Chinese	2	173,884,238
Yoruba	2	6,198,347
Igbo	2	2,650,116
Arabic	1	694,455,304
Portuguese	1	30,615,557
Kinyarwanda	1	9,301,301

Table 4: Pseudocrawled data per language sorted by number of domains crawled

F Author contributions

Author contributions in alphabetical order.

Aaron Gokaslan set a `pre-commit` (for code formatting) in a repository and helped with the writing of the Related Work section of the paper.

Aitor Soroa integrated one dataset into crowdsourced data.

Albert Villanova del Moral led the gathering of identified sources, implemented loading scripts in the *datasets* library in a single unified interface, and integrated the most datasets into crowdsourced data.

Angelina McMillan-Major gathered the lists of closed class words for many languages used for the filtering of OSCAR.

Anna Rogers contributed to the writing of the paper.

Chenghao Mou was the main contributor for OSCAR deduplication.

Christopher Akiki advised on analysis aspects of the project, integrated over a hundred datasets into crowdsourced data, wrote dataset loading scripts, participated in cleaning and filtering efforts, helped with visualization, and contributed to the writing of the paper.

Daniel van Strien integrated one dataset into crowdsourced data.

David Ifeoluwa Adelani participated in the PII filtering initiative (see Appendix D).

Eduardo González Ponferrada trained SentencePiece and KenLM models used for the filtering of OSCAR.

Francesco De Toni led one crowdsourcing hackathon, analyzed the distribution of the sources in the catalogue, participated in the PII filtering initiative (see Appendix D), and contributed to the writing of the paper.

Giada Pistilli helped write the Ethical Considerations section of the paper.

Gérard M Dupont contributed to data tooling and sourcing and advised on analysis aspects of the project.

Hieu Tran helped set the filtering parameters for Vietnamese and contributed to the list of Vietnamese closed-class words.

Hugo Laurençon developed the filtering library used for the cleaning of OSCAR and the visualization tool to help choose the filtering parameters, and ran OSCAR filtering jobs. He was involved in the cleaning of some crowdsourced and pseudo-crawled datasets and the deduplication of OSCAR. He also contributed to the writing of the paper.

Huu Nguyen contributed to the data tooling and lead the PII filtering initiative (see Appendix D).

Ian Yu participated in the PII filtering initiative (see Appendix D) and helped to choose the filtering parameters for Chinese.

Itziar Gonzalez-Dios, as a Basque native speaker, helped choose the filtering parameters for this language.

Javier De la Rosa contributed with perplexity sampling efforts for OSCAR (not used in final pipeline).

Jenny Chim participated in the PII filtering initiative (see Appendix D) and helped to choose the filtering parameters and closed-class words for Chinese.

Jian Zhu integrated two datasets into crowdsourced data.

Jörg Frohberg integrated multiple datasets into crowdsourced data and reached out to license holders.

Khalid Almubarak integrated some datasets into crowdsourced data.

Kyle Lo integrated one dataset into crowdsourced data.

Leandro von Werra participated in cleaning and filtering efforts, built the code dataset, contributed to its analysis, and participated in the sourcing effort for target datasets.

Leon Weber integrated one dataset into crowdsourced data.

Long Phan participated in the PII filtering initiative (see Appendix D).

Loubna Ben allal contributed the analysis of the code dataset.

Lucile Saulnier co-led pseudo-crawled data acquisition, contributed to filtering, cleaning, and deduplication for the crowdsourced datasets, built visualization tools to inspect the results of preprocessing, scaled the PII filtering process, performed the document size analysis, and participated in paper writing.

Manan Dey, as a Bengali native speaker, helped to choose the filtering parameters for this language.

Manuel Romero Muñoz contributed to KenLM models and to corpus visualization with those models, and participated in the PII filtering initiative (see Appendix D).

Maraim Masoud contributed to the sourcing of some Arabic datasets, the list of Arabic closed-class words, and the writing of the paper.

Margaret Mitchell co-led the final regex-based PII efforts.

Mario Šaško integrated multiple datasets into crowdsourced data.

Olivier Nguyen helped build the first blocks of the OSCAR filtering pipeline.

Paulo Villegas participated in the PII filtering initiative (see Appendix D) and the perplexity sampling efforts for OSCAR.

Pedro Ortiz Suarez contributed to crowdsourced data and provided high-level metrics on OSCAR.

Pierre Colombo participated in the PII filtering initiative (see Appendix D).

Quentin Lhoest integrated multiple datasets into crowdsourced data.

Sasha Luccioni co-led the final regex-based PII efforts. participated in filtering and cleaning efforts, and contributed to the writing of the paper.

Sebastian Nagel helped to implement the pseudo-crawled data acquisition step.

Shamik Bose participated in the PII filtering initiative (see Appendix D) and contributed the list of Bengali closed-class words.

Shayne Longpre contributed to the writing of the paper.

Somaieh Nikpoor co-led the development of ethical/legal charter and contributed to the section on ethical considerations.

Stella Biderman contributed to the writing of the paper.

Suhas Pai participated in the PII filtering initiative (see Appendix D).

Suzana Ilić coordinated the organization of BigScience working groups.

Teven Le Scao led final quality control checks, contributed to filtering, cleaning, and deduplication for all components of the corpus, contributed to the OSCAR filtering visualization tool, contributed and repaired several datasets for crowdsourced data, performed the tokenizer-based analysis, and participated in the writing of the paper.

Thomas Wang co-led pseudo-crawled data acquisition, built the distributed cleaning pipelines for pseudo-crawled and crowdsourced datasets, handled job monitoring for crowdsourced dataset filtering, and participated in paper writing.

Tristan Thrush participated in the final regex-based PII efforts.

Violette Lepercq was the primary project manager for the final dataset cleaning efforts and helped to reach out to the native speakers to tune the filtering parameters.

Vu Minh Chien participated in the PII filtering initiative (see Appendix D).

Yacine Jernite defined the primary goals of the project, advised on data collection and filtering efforts, contributed sourcing tools and early cleaning scripts, and contributed to the writing of the paper.

Zaid Alyafeai helped find a list of Arabic datasets that should be integrated into crowdsourced data.