Datasets: A Community Library for Natural Language Processing

Quentin Lhoest*, Albert Villanova del Moral*, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško[↑], Gunjan Chhablani[↑], Bhavitvya Malik[↑], Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf [Ω]

Hugging Face / {quentin,thomas}@huggingface.co

Abstract

The scale, variety, and quantity of publicly-available NLP datasets has grown rapidly as researchers propose new tasks, larger models, and novel benchmarks. *Datasets* is a community library for contemporary NLP designed to support this ecosystem. *Datasets* aims to standardize end-user interfaces, versioning, and documentation, while providing a lightweight front-end that behaves similarly for small datasets as for internet-scale corpora. The design of the library incorporates a distributed, community-driven approach to adding datasets and documenting usage. After a year of development, the library now includes more than 650 unique datasets, has more than 250 contributors, and has helped support a variety of novel cross-dataset research projects and shared tasks. The library is available at https://github.com/huggingface/datasets.

1 Introduction

Datasets are central to empirical NLP: curated datasets are used for evaluation and benchmarks; supervised datasets are used to train and fine-tune models; and large unsupervised datasets are necessary for pretraining and language modeling. Each dataset type differs in scale, granularity and structure, in addition to annotation methodology. Historically, new dataset paradigms have been crucial for driving the development of NLP, from the Hansard corpus for statistical machine translation (Brown et al., 1988) to the Penn Treebank for syntactic modeling (Marcus et al., 1993) to projects like OPUS and Universal Dependencies (Nivre et al., 2016; Tiedemann and Nygaard, 2004) which bring together cross-lingual data and annotations.

Contemporary NLP systems are now developed with a pipeline that utilizes many different datasets at significantly varying scale and level of annotation (Peters et al., 2018). Different datasets are used for pretraining, fine-tuning, and benchmarking. As such, there has been a large increase in the number of datasets utilized in the NLP community. These include both large text collections like C4 (Raffel et al., 2020), fine-tuning datasets like SQuAD (Rajpurkar et al., 2016), and even complex zero-shot challenge tasks. Benchmark datasets like GLUE have been central to quantifying the the advances of models such as BERT (Wang et al., 2018; Devlin et al., 2019).

The growth in datasets also brings significant challenges, including interface standardization, versioning, and documentation. A practitioner should be able to utilize N different datasets without requiring N different interfaces. In addition, N practitioners using the same dataset should know they have exactly the same version. Datasets have also grown larger, and ideally interfaces should not have to change due to this scale, whether one is using small-scale datasets like Climate Fever ($\sim$1k data points), medium-scale Yahoo Answers ($\sim$1M), or even all of PubMed ($\sim$79B). Finally, datasets are being created with a variety of different procedures, from crowd-sourcing to scraping to synthetic generation, which need to be taken into account when evaluating which is most appropriate for a given purpose and ought to be immediately apparent to prospective users (Gebru et al., 2018).

Datasets is a community library designed to address the challenges of dataset management and access, while supporting community culture and norms. The library targets the following goals:

- Ease-of-use and Standardization: All datasets can be easily downloaded with one line of

*Lead Library Maintainers, [Ω] Library Creator, [↑] Independent Research Contributor

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 175–184
November 7–11, 2021.

code. Each dataset utilizes a standard tabular format, and is versioned and cited.

- Efficiency and Scale: Datasets are computation- and memory-efficient by default and work seamlessly with tokenization and featurization. Massive datasets can even be streamed through the same interface.
- Community and Documentation: The project is community-built and has hundreds of contributors across languages. Each dataset is tagged and documented with a datasheet describing its usage, types, and construction.

Datasets is in continual development by the engineers at Hugging Face and is released under an Apache 2.0 license.[1] The library is available at https://github.com/huggingface/datasets. Full documentation is available through the project website.[2]

2 Related Work

There is a long history of projects aiming to group, categorize, version, and distribute NLP datasets which we briefly survey. Most notably, the Linguistic Data Consortium (LDC) stores, serves, and manages a variety of datasets for language and speech. In addition to hosting and distributing corpus resources, the LDC supports significant annotation efforts. Other projects have aimed to collect related annotations together. Projects like OntoNotes have collected annotations across multiple tasks for a single corpus (Pradhan and Xue, 2009) whereas the Universal Dependency treebank (Nivre et al., 2016) collects similar annotations across languages. In machine translation, projects like OPUS catalog the translation resources for many different languages. These differ from *Datasets* which collects and provides access to datasets in a content-agnostic way.

Other projects have aimed to make it easy to access core NLP datasets. The influential NLTK project (Bird, 2006) provided a data library that makes it easy to download and access core datasets. SpaCy also provides a similar loading interface (Honnibal and Montani, 2017). In recent years, concurrent with the move towards deep learning, there has been a growth in large freely available datasets often with less precise annotation standards. This has motivated cloud-based repositories of datasets. Initiatives like TensorFlow-Datasets (2021) and TorchText (2021) have collected various datasets in a common cloud format. This project began as a fork of TensorFlow-Datasets, but has diverged significantly.

Datasets differs from these projects along several axes. The project is decoupled from any modeling framework and provides a general-purpose tabular API. It focuses on NLP specifically and provides specialized types and structures for language constructs. Finally, it prioritizes community management and documentation through the dataset hub and data cards, and aims to provide access to a long-tail of datasets for many tasks and languages.

3 Library Tour and Design

We begin with a brief tour. Accessing a dataset is done simply by referring to it by a global identity.

```
dataset = load_dataset("boolq")
```

Each dataset has a features schema and metadata.

```
print(dataset.features, dataset.info)
```

Any slice of data points can be accessed directly without loading the full dataset into memory.

```
dataset["train"][start:end]
```

Processing can be applied to every data point in a batched and parallel fashion using standard libraries such as NumPy or Torch.

```
# Torch function "tokenize"
tokenized = dataset.map(tokenize,
                        num_proc=32)
```

Datasets facilitates each of these four Stages with the following technical steps.

S1. Dataset Retrieval and Building *Datasets* does not host the underlying raw datasets, but accesses hosted data from the original authors in a distributed manner.[3] Each dataset has a community contributed builder module. The builder module has the responsibility of processing the raw data, e.g. text or CSV, into a common dataset interface representation.

S2. Data Point Representation Each built dataset is represented internally as a table with typed columns. The *Dataset* type system includes a variety of common and NLP-targeted types. In addition to atomic values (int's, float's, string's,

[1] Datasets themselves may utilize different licenses which are documented in the library.

[2] https://huggingface.co/docs/datasets/

[3] For datasets with intensive preprocessing, such as Wikipedia, a preprocessed version is hosted. Datasets removed by the author are not centrally cached and become unavailable.

binary blobs) and JSON-like dicts and lists, the library also includes named categorical class labels, sequences, paired translations, and higher-dimension arrays for images, videos, or waveforms.

S3. In-Memory Access *Datasets* is built on top of Apache Arrow, a cross-language columnar data framework (Arrow, 2020). Arrow provides a local caching system allowing datasets to be backed by an on-disk cache, which is memory-mapped for fast lookup. This architecture allows for large datasets to be used on machines with relatively small device memory. Arrow also allows for copy-free hand-offs to standard machine learning tools such as NumPy, Pandas, Torch, and TensorFlow.

S4. User Processing At download, the library provides access to the typed data with minimal preprocessing. It provides functions for dataset manipulation including sorting, shuffling, splitting, and filtering. For complex manipulations, it provides a powerful *map* function that supports arbitrary Python functions for creating new in-memory tables. For large datasets, map can be run in batched, multi-process mode to apply processing in parallel. Furthermore, data processed by the same function is automatically cached between sessions.

Complete Flow Upon requesting a dataset, it is downloaded from the original host. This triggers dataset-specific builder code which converts the text into a typed tabular format matching the feature schema and caches the table. The user is given a memory-mapped typed table. To further process the data, e.g. tokenize, the user can run arbitrary vectorized code and cache the results.

4 Dataset Documentation and Search

Datasets is backed by the *Dataset Hub* [4] that helps users navigate the growing number of available resources and draws inspiration from recent work calling for better documentation of ML datasets in general (Gebru et al., 2018) and NLP datasets in particular (Bender and Friedman, 2018).

Datasets can be seen as a form of infrastructure (Hutchinson et al., 2021). NLP practitioners typically make use of them with a specific goal in mind, whether they are looking to answer a specified research question or developing a system for a particular practical application. To that end, they need to be able to not only easily identify which dataset is most appropriate for the task at hand, but also to understand how various properties of that best candidate might help with, or, conversely, run contrary to their purpose.

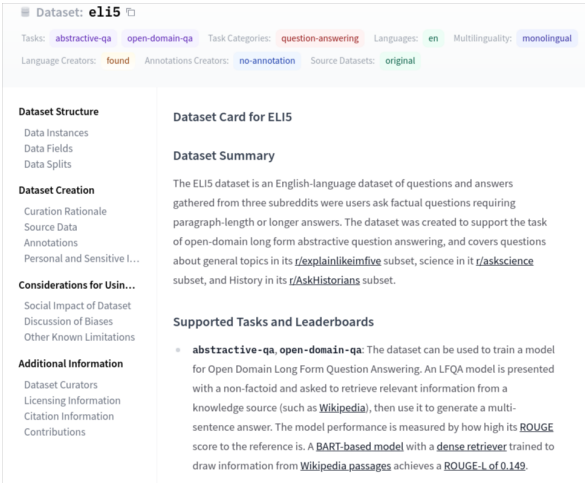

Figure 1: The data card for ELI5 (Fan et al., 2019).

The *Dataset Hub* includes all of the datasets available in the library. It links each of them together though: a set of *structured tags* holding information about their languages, tasks supported, licenses, etc.; a *data card* based on a template[5] designed to combine relevant technical considerations and broader context information (McMillan-Major et al., 2021); and a *list of models* trained on the dataset. Both the tags and data card are filled manually by the contributor who introduces the dataset to the library. Figure 1 presents an example of the dataset page on the hub.[6] Together, these pages and the search interface help users navigate the available resources.

Choosing a Dataset Given a use case, the structured tags provide a way to surface helpful datasets. For example, requesting all datasets that have the tags for Spanish language and the Question Answering task category returns 7 items at the time of writing. A user can then refine their choice by reading through the data cards, which contain sections describing the variety of language used, legal considerations including licensing and incidence of Personal Identifying Information, and paragraphs about known social biases resulting from the collection process that might lead a deployed model to cause disparate harms.

[4] https://hf.co/datasets/

[5] https://hf.co/datasets/card-guide

[6] https://hf.co/datasets/eli5

Using a Dataset The data card also contains information to help users navigate all the choices, from hardware to modeling, that go into successfully training a system. These include the number of examples in each of the dataset splits, the size on disk of the data, meaningful differences between the training, validation, and test split, and free text descriptions of the various fields that make up each example to help decide what information to use as input or output of a prediction model.

The Data Card as a Living Document A dataset's life continues beyond its initial release. As NLP practitioners interact with the dataset in various ways, they may surface annotation artifacts that affect the behavior of trained models in unexpected ways (Gururangan et al., 2018),[7] issues in the way the standard split was initially devised to test a model's ability to adapt to new settings (Krishna et al., 2021), or new understanding of the social biases exhibited therein (Hutchinson et al., 2020). The community-driven nature of *Datasets* and the versioning mechanisms provided by the GitHub backend provide an opportunity to keep the data cards up to date as information comes to light and to make gradual progress toward having as complete documentation as possible.

5 Dataset Usage and Use-Cases

Datasets is now being actively used for a variety of tasks. Figure 2 (left) shows statistics about library usage. We can see that the most commonly downloaded libraries are popular English benchmarks such as GLUE and SQuAD which are often used for teaching and examples. However there is a range of popular models for different tasks and languages.

Figure 2 (right) shows the wide coverage of the library in terms of task types, sizes, and languages, with currently 681 total datasets. During the development of the *Datasets* project, there was a public hackathon to have community members develop new Dataset builders and add them to the project. This event led 485 commits and 285 unique contributors to the library. Recent work has outlined the difficulty of finding data sources for lower-resourced languages through automatic filtering alone (Caswell et al., 2021). The breadth of languages spoken by participants in this event made it possible to more reliably bootstrap the library with datasets in a wide range of different languages. Finally while *Datasets* is designed for NLP, it is becoming used for multi-modal datasets. The library now includes types for continuous data, including multi-dimensional arrays for image and video data and an *Audio* type.

5.1 Case Studies: N-Dataset NLP

A standardized library of datasets opens up new use-cases beyond making single datasets easy to download. We highlight three use-cases in which practitioners have employed the *Datasets* library.

Case Study 1: N-task Pretraining Benchmarks Benchmarking frameworks such as NLP Decathlon and GLUE have popularized the comparison of a single NLP model across a variety of tasks (McCann et al., 2018; Wang et al., 2018). Recently benchmarking frameworks like GPT-3's test suite framework (Brown et al., 2020) have expanded this benchmarking style even further, taking on dozens of different tasks. This research has increased interest in comparison of different datasets at scale.

Datasets is designed to facilitate large-scale, N-task benchmarking beyond what might be possible for a single researcher to set up. For example, the Eleuther AI project aims to produce a massive scale open-source model. As part of this project they have released an *LM Evaluation Harness*[8] which includes nearly 100 different NLP tasks to test a large scale language model. This framework is built with the *Datasets* library as a method for retrieving and caching datasets.

Case Study 2: Reproducible Shared Tasks NLP has a tradition of shared tasks that become long-lived benchmark datasets. Tasks like CoNLL 2000 (Tjong Kim Sang and Buchholz, 2000) continue to be widely used more than 20 years after their release. *Datasets* provides a convenient, reproducible, and standardized method for hosting and maintaining shared tasks, particularly when they require multiple different datasets.

Datasets was used to support the first GEM (Generation, Evaluation, and Metrics) workshop (Gehrmann et al., 2021). This workshop ran a shared task comparing natural language generation (NLG) systems on 12 different tasks. The tasks included examples from twenty different languages and supervised datasets varying from size of 5k examples to 500k. Critically, the shared task had

[7] https://hf.co/datasets/snli#other-known-limitations

[8] https://github.com/EleutherAI/lm-evaluation-harness

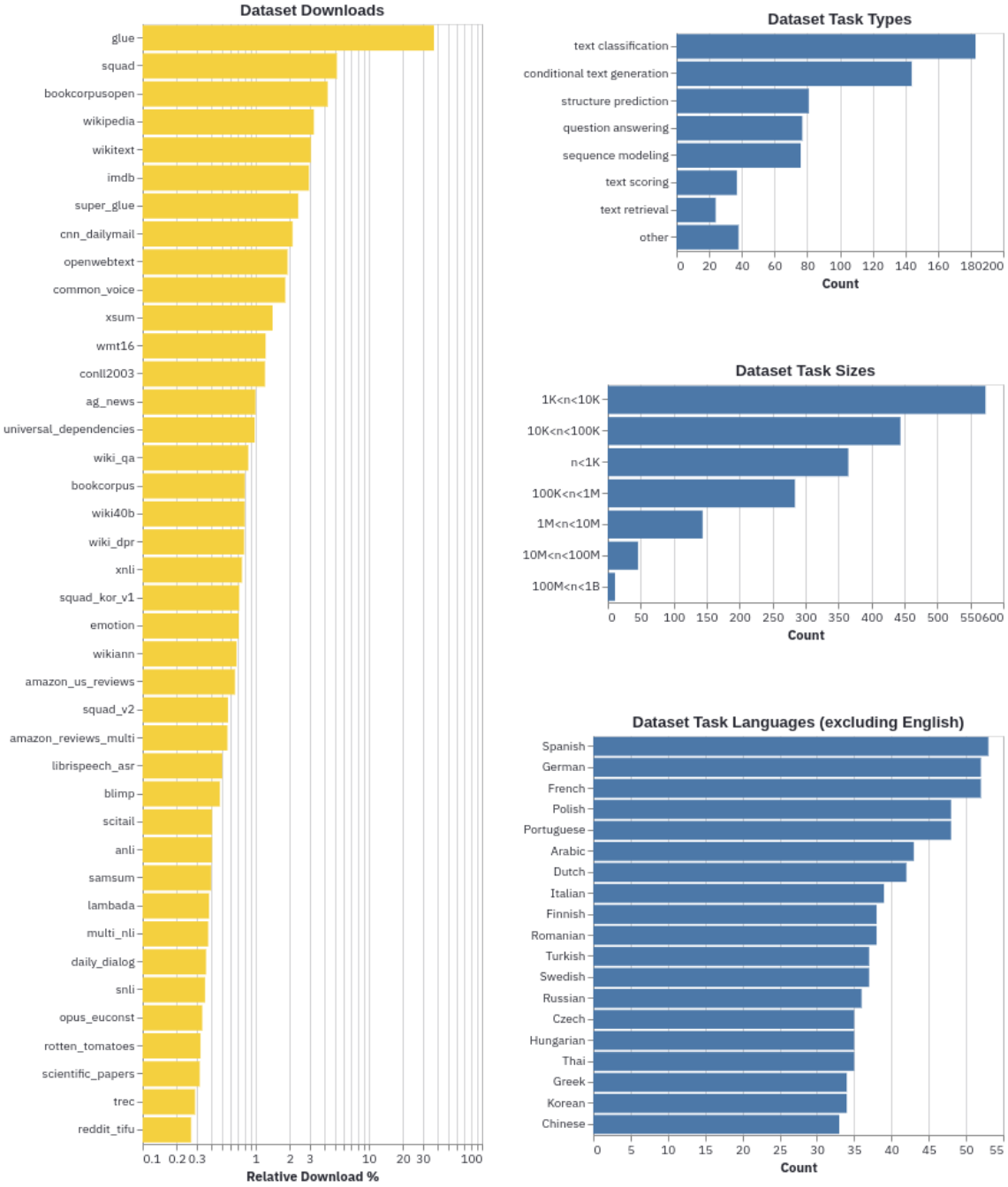

Figure 2: Summary statistics from the datasets in the library. (**Left**) The relative download numbers of the most popular datasets in the library. (**Right**) Task properties. Each dataset may have multiple sub-tasks. Task Types are the types labeled in the library. Task Sizes are the number of data points in the table. Task Languages are the languages tagged in the library (many datasets include tasks in different languages).

a large variety of different input formats including tables, articles, RDF triples, and meaning graphs. *Datasets* allows users to access all 12 datasets with a single line of code in their shared task description.

Case Study 3: Robustness Evaluation While NLP models have improved to the point that on-paper they compete with human performance, many research projects have demonstrated that these same models are fooled when given out-of-domain examples (Koehn and Knowles, 2017), simple adversarial constructions (Belinkov and Bisk, 2018), or examples that spuriously match basic patterns (Poliak et al., 2018).

Datasets can be used to support better benchmarking of these issues. The *Robustness Gym*[9] proposes a systematic way to test an NLP system across many different proposed techniques, specifically subpopulations, transformations, evaluation sets, and adversarial attacks (Goel et al., 2021). Together, these provide a robustness report that is more specific than a single evaluation measure. While developed independently, the Robustness Gym is built on *Datasets*, and "relies on a common data interface" provided by the library.

6 Additional Functionality and Uses

Streaming Some datasets are extremely large and cannot even fit on disk. *Datasets* includes a streaming mode that buffers these datasets on the fly. This mode supports the core map primitive, which works on each data batch as it is streamed. Datasets streaming helped enable recent research into distributed training of a very large open NLP model (Diskin et al., 2021).

Indexing *Datasets* includes tools for easily building and utilizing a search index over an arbitrary dataset. To construct the index the library can interface either with FAISS or ElasticSearch (Johnson et al., 2017; Elastic, 2021). This interface makes it easy to efficiently find nearest neighbors either with textual or vector queries. Indexing was used to host the open-source version of Retrieval-Augmented Generation (Lewis et al., 2020), a generation model backed by the ability to query knowledge from large-scale knowledge sources.

Metrics *Datasets* includes an interface for standardizing *metrics* which can be documented, versioned and matched with datasets. This functionality is particularly useful for benchmark datasets

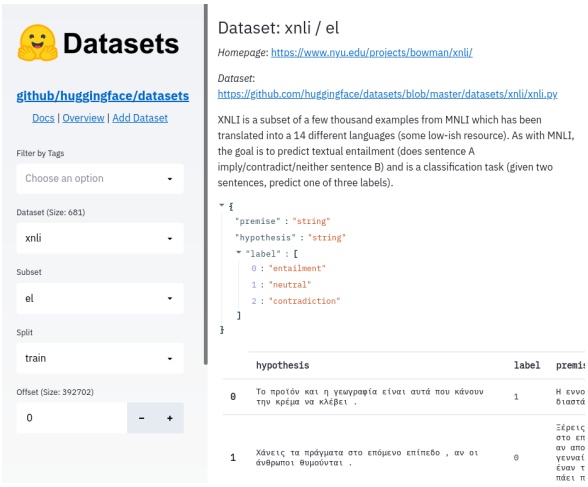

Figure 3: *Datasets* viewer is an application that shows all rows for all datasets in the library. The interface allows users to change datasets, subsets, and splits, while seeing the dataset schema and metadata.

such as GLUE that include multiple tasks each with their own metric. Some metrics like BLEU and SQuAD are included directly in the library code, whereas others are linked to external packages. The library also allows for metrics to be applied in a distributed manner over the dataset.

Data Viewer A benefit of the standardized interface of the library is that it makes it trivial to build a cross-task dataset viewer. As an example, Hugging Face hosts a generic viewer for the entirety of *datasets* (Figure 3) [10]. In this viewer, anyone on the web can open all almost 650 different datasets and view any example. Because the tables are typed, the viewer can easily show all component features, structured data, and multi-modal features.

7 Conclusion

Hugging Face *Datasets* is an open-source, community-driven library that standardizes the processing, distribution, and documentation of NLP datasets. The core library is designed to be easy to use, fast, and to use the same interface for datasets of varying size. At 650 datasets from over 250 contributors, it makes it easy to use standard datasets, has facilitated new use cases of cross-dataset NLP, and has advanced features for tasks like indexing and streaming large datasets.

[9] https://robustnessgym.com/

[10] https://huggingface.co/datasets/viewer/

Acknowledgements

While organized by Hugging Face, *Datasets* is an open-source project driven by contributors. This work was only possible thanks to Charin Polpanumas, Cahya Wirawan, Jonatas Grosman, Thomas Hudson, Zaid Alyafeai, Rahul Chauhan, Vineeth S, Sandip, Yvonnegitau, Jared T Nielsen, Michal Jamry, Bharat Raghunathan, Ceyda Cinarel, David Adelani, Misbah Khan, Steven Liu, Vasudev Gupta, Matthew Bui, Abdul Rafay Khalid, Beth Tenorio, Eduardo Gonzalez Ponferrada, Harshal Mittal, Hugo Abonizio, Moussa Kamal Eddine, Stefan Schweter, Sumanth Doddapaneni, Yavuz Kömeçoğlu, Yusuke Mori, J-chim, Ontocord, Skyprince999, Vrindaprabhu, Jonathan Bragg, Philip May, Alexander Seifert, Ivanzidov, Jake Tae, Karim Foda, Mohamed Al Salti, Nick Doiron, Vinay, Czabo, Vblagoje, Nilansh Rajput, Abdulelah S. Al Mesfer, Akshay Bhardwaj, Amit Moryossef, Basava Sai Naga Viswa Chaitanya, Darek Kłeczek, Darshan Gandhi, Gustavo Aguilar, Hassan Ismail Fawaz, Jack Morris, Jamesg, Jonathan Chang, Karthik Bhaskar, Manan Dey, Maria Grandury, Michael A. Hedderich, Mounica Maddela, Nathan Cooper, Purvi M, Richard Wang, Song Feng, Sourab Mangrulkar, Tanmoy, Vijayasaradhi, Zacharysbrown, Chameleontk, Eusip, Jeromeku, Patpizio, Tuner007, Benjamin Van Der Burgh, Bharati Patidar, George Mihaila, Olivier, Tim Isbister, Alessandro Suglia, Başak Buluz Kömeçoğlu, Boris Dayma, Dariusz Kajtoch, Frankie Robertson, Jieyu, Mihaelagaman, Nikhil Bartwal, Param Bhavsar, Paullerner, Rachelker, Ricardo Rei, Sai, Sasha Rush, Suraj Parmar, Takuro Niitsuma, Taycir Yahmed, Tuan-phong Nguyen, Vladimir Gurevich, Alex, Calpt, Idoh, Justin-yan, Katnoria, Sileod, Avinash Swaminathan, Connor Mccarthy, Jungwhan Kim, Leo Zhao, Sanjay Kamath, (bill) Yuchen Lin, 2dot71mily, 8bitmp3, Abi Komma, Adam, Adeep Hande, Aditya Siddhant, Akash Kumar Gautam, Alaa Houimel, Alex Dong, Along, Anastasia Shimorina, Andre Barbosa, Anton Lozhkov, Antonio V Mendoza, Ashmeet Lamba, Ayushi Dalmia, Batjedi, Behçet Şentürk, Bernardt Duvenhage, Binny Mathew, Birger Moëll, Blanc Ray, Bram Vanroy, Clément Rebuffel, Daniel Khashabi, David Fidalgo, David Wadden, Dhruv Kumar, Diwakar Mahajan, Elron Bandel, Emrah Budur, Fatima Haouari, Fraser Greenlee, Gergely Nemeth, Gowtham.r, Hemil Desai, Hiroki Nakayama, Ilham F Putra, Jannis Vamvas, Javier De La Rosa, Javier-jimenez99, Jeff Hale, Jeff Yang, Joel Niklaus, John Miller, John Mollas, Joshua Adelman, Juan Julián Cea Morán, Kacper Łukawski, Koichi Miyamoto, Kushal Kedia, Laxya Agarwal, Leandro Von Werra, Loïc Estève, Luca Di Liello, Malik Altakrori, Manuel, Maramhasanain, Marcin Flis, Matteo Manica, Matthew Peters, Mehrdad Farahani, Merve Noyan, Mihai Ilie, Mitchell Gordon, Niccolò Campolungo, Nihal Harish, Noa Onoszko, Nora Belrose, Or Sharir, Oyvind Tafjord, Pewolf, Pariente Manuel, Pasquale Minervini, Pedro Ortiz Suárez, Pedro Lima, Pengcheng Yin, Petros Stavropoulos, Phil Wang, Philipp Christmann, Philipp Dufter, Philippe Laban, Pierre Colombo, Rahul Danu, Rabeeh Karimi Mahabadi, Remi Calizzano, Reshinth Adithyan, Rodion Martynov, Roman Tezikov, Sam Shleifer, Savaş Yıldırım, Sergey Mkrtchyan, Shubham Jain, Shubhambindal2017, Subhendu Ranjan Mishra, Taimur Ibrahim, Tanmay Thakur, Thomas Diggelmann, Théophile Blard, Tobias Slott, Tsvetomila Mihaylova, Vaibhav Adlakha, Vegar Andreas Bergum, Victor Velev, Vlad Lialin, Wilson Lee, Yang Wang, Yasir Abdurrohman, Yenting (Adam) Lin, Yixin Nie, Yoav Artzi, Yoni Gottesman, Yongrae Jo, Yuxiang Wu, Zhong Peixiang, Zihan Wang, Aditya2211, Alejandrocros, Andy Zou, Brainshawn, Cemilcengiz, Chutaklee, Gaurav Rai, Dhruvjoshi1998, Duttahritwik, Enod, Felixgwu, Ggdupont, Jerryishere, Jeswan, Lodgi, Lorinczb, Maxbartolo, Nathan Dahlberg, Neal, Ngdodd, Kristo, Onur Güngör, Ophelielacroix, Padipadou, and Phiwi.

References

Apache Arrow. 2020. Apache Arrow, a cross-language development platform for in-memory analytics. https://arrow.apache.org/.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1988. A statistical approach to language translation. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *CoRR*, abs/2103.12028.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitry Popov, Dmitry Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Thomas Wolf, and Gennady Pekhimenko. 2021. Distributed deep learning in open collaborations.

Elastic. 2021. Elastic Search. `https://www.elastic.co/`.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondvrej Duvsek, Chris C. Emezue, Varun Gangal, Cristina Garbacea, T. Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, V. Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Rao, Vikas Raunak, Juan Diego Rodríguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, W. Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *ArXiv*, abs/2102.01672.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5491–5501. Association for Computational Linguistics.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4940–4957. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Bryan McCann, N. Keskar, Caiming Xiong, and R. Socher. 2018. The natural language decathlon: Multitask learning as question answering. *ArXiv*, abs/1806.08730.

Angelina McMillan-Major, Salomey Osey, Juan Diego Rodríguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable templates and guides for documenting datasets and models for natural language processing and generation - a case study of the huggingface and gem data and model cards.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer S. Pradhan and Nianwen Xue. 2009. OntoNotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12, Boulder, Colorado. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

TensorFlow-Datasets. 2021. TensorFlow Datasets, a collection of ready-to-use datasets. `https://www.tensorflow.org/datasets`.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

TorchText. 2021. TorchText. https://pytorch.org/text/stable/index.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.