Raison d'être of the benchmark dataset:
A Survey of Current Practices of Benchmark Dataset Sharing Platforms

Jaihyun Park [*]
School of Information Sciences
University of Illinois Urbana-Champaign
jaihyun2@illinois.edu

Sullam Jeoung [*]
School of Information Sciences
University of Illinois Urbana-Champaign
sjeoung2@illinois.edu

Abstract

This paper critically examines the current practices of benchmark dataset sharing in NLP and suggests a better way to inform reusers of the benchmark dataset. As the dataset sharing platform plays a key role not only in distributing the dataset but also in informing the potential reusers about the dataset, we believe data sharing platforms should provide a comprehensive context of the datasets. We survey four benchmark dataset sharing platforms: HuggingFace, PaperswithCode, Tensorflow, and Pytorch to diagnose the current practices of how the dataset is shared - *which metadata is shared and omitted*. To be specific, drawing on the concept of *data curation* which considers the future reuse when the data is made public, we advance the direction that benchmark dataset sharing platforms should take into consideration. We identify that four benchmark platforms have different practices of using metadata and there is a lack of consensus on what social impact metadata is. We believe the problem of missing a discussion around social impact in the dataset sharing platforms has to do with the failed agreement on who should be in charge. We propose that the benchmark dataset should develop social impact metadata and data curator should take a role in managing the social impact metadata.

1 Introduction

Benchmark datasets play a crucial role in developing the model. Publicly available benchmark datasets serve as a baseline proxy to measure the model's performance and an evaluation as the machine learning (ML) and natural language processing (NLP) scholarship competes for the higher ground. Recent works have started to question the validity of such benchmark datasets regarding their generalizability (Bowman and Dahl, 2021; Paullada et al., 2021), documentation practices (Bender and Friedman, 2018), and social impact (Hovy and Spruit, 2016; Sap et al., 2021), amongst others. Paullada et al. (2021) focus on the way how benchmark datasets are collected and used and advocate cautious understanding of data in order to address ethical issues of using such datasets. Bowman and Dahl (2021) suggest the criteria benchmarks should qualify, namely the robustness, statistical power, and considerations of social impact. However, despite the fact that the documentation of benchmark datasets and the role of the dataset sharing platform are pivotal not only in informing the users about the benchmark dataset but also soliciting a safe use, it has been relatively understudied. We believe that critically examining the current practices of dataset sharing platforms - which metadata is documented and omitted - and suggesting desiderata for data sharing platforms can serve as a practical guide for users and researchers in encouraging a safe environment.

Our findings show that current practices of dataset sharing platforms are highly centered on *reusable* purpose, which focuses on the convenience of the users in making use of the dataset. For example, it provides detailed explanations of how to load the dataset into actual development, how the test and train split are made. It was hard, on the other hand, to find the documentation of the limitations of the dataset (e.g. which societal impacts it may bring); even if there were, the concepts and definitions were often elusive. We introduce the concept of *social impact metadata* which is the documenting practice done in Library and Information Science in order to advocate mitigating possible social harms.

We propose desiderata for documenting benchmark datasets. Beyond descriptive and administrative metadata, the documents of the metadata should also include the social impact metadata. To make it possible, we highly encourage developing

[*] Both authors contributed equally to this research.

Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP, pages 1 - 10
May 26, 2022 ©2022 Association for Computational Linguistics

the social impact metadata (e.g. demographic statistics of the data) and also emphasize a role of the data curator who is responsible for documenting in terms of the data sharing platforms.

2 Definitions

In order to narrow down the conceptual difference that may conflict between the ML (and NLP) community and Library and Information Science community, we introduce the definition of the key terms that will be used throughout the paper.

Data documentation Data documentation (sometimes called a "codebook") is helpful in understanding and interpreting the dataset (Vardigan et al., 2008). A document can be defined as 'anything in which knowledge is recorded' and documentation is 'any process which serves to make a document available to the user after knowledge.' (Woledge, 1983). With this sense, Data Documentation Initiatives (DDI) defines data documentation as 'document and manage data across the entire data life cycle, from conceptualization to data publication, analysis and beyond' (DDI, 2020). ML community defines the data documentation as 'annotating various demographic characteristics for disaggregated testing, gathering representative data, and providing documentation pertaining to the data gathering and annotation process' (Jo and Gebru, 2020).

Benchmark dataset Benchmark dataset refers to the typical set of datasets that are commonly used for evaluating the model's baseline performance on specific tasks (Bowman and Dahl, 2021). Some of the widely used benchmark datasets in NLP are GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) for natural language understanding, SQuAD (Rajpurkar et al., 2016) for question and answering, and Seteroset (Nadeem et al., 2021), CrowS (Nangia et al., 2020) for checking biased natural language models, amongst others.

Data sharing platform We consider a data-sharing platform, that provides access to the datasets or the metadata of the datasets. Normally the benchmark datasets are shared through the third party provider rather than the data creators themselves. Generally, the data sharing platforms offer the users direct access to the datasets by their pre-defined methods that are compatible with their libraries used for developing NLP models (e.g. HuggingFace, PyTorch, Tensorflow). Apart from the concept of data curator, who collects, selects, and participates in the data creation process, contributors are the ones that upload and document the dataset to the data-sharing platform. This could be voluntary individuals (e.g. HuggingFace), or in part supported by automated algorithms (e.g. PaperswithCode).

Descriptive metadata Descriptive metadata is considered to contain information that can help users to find, identify, select, and obtain the resource. Title, creator, keywords, subject, type of resource, and other attributes that describe what the resource is about are considered as descriptive metadata (Liu and Qin, 2014; Pomerantz, 2015). In practice, descriptive metadata in the archives can be used to catalog entities, events, time, and space to answer the queries that the users want to find (Dobreski et al., 2020). Descriptive metadata in the benchmark dataset can be derived from variables inside of the dataset. The domain (e.g., social media, news media), scope (e.g., the topic covered by the dataset) can be additional descriptive metadata of the benchmark dataset.

Administrative metadata Administrative metadata is required to house information about managing and administering collections. Administrative metadata includes information about rights, versions, and preservation (NISO, 2004). For the benchmark dataset, the version can be appropriate administrative metadata.

3 Social Impact of Benchmark Dataset Sharing Platforms

Once the benchmark dataset is made available for others, *data friction* comes into play. *Data friction* explains a point of resistance where data can be garbled, misinterpreted, or lost (Edwards et al., 2011). As Edwards et al. (2011) argue, researchers' main interest is in *using* data, not in *describing* the dataset for the benefit of invisible, unknown future users. The problem of benchmark dataset sharing arise because text-as-data and computation is no longer exclusive field of NLP and ML (Monroe et al., 2008). The benchmark dataset can be easily used by researchers outside of NLP and ML community. For instance, now in the name of digital humanities, researchers in humanities also use computational approaches and technologies with historical text data (Connolly, 2020; Soni et al., 2021; Smith et al., 2014).

Likewise, if the benchmark dataset is shared through the sharing platforms, ML and NLP researchers will make derivative models based on the benchmark dataset and this will lead the researchers outside of ML and NLP to indirectly impacted by benchmark dataset without knowing the social impact of the dataset. Even though humanists and social scientists may not fine-tune the parameters of the model itself, their research will be impacted by how the benchmark dataset is designed and constructed. The use of pre-trained model from NLP community by researchers outside of NLP and ML can be found in the case of *politeness detection* model. The original idea of developing a NLP model for detecting politeness from language is from Danescu-Niculescu-Mizil et al. (2013). As the automatic scoring of the politeness in the language has benefits regardless of the field, Hoffman et al. (2017) attempted to reproduce and validate Danescu-Niculescu-Mizil et al. (2013). In doing so, Hoffman et al. (2017) applied the same model to Wikipedia, which is the identical domain and found unexpected results that led them to question the quality of the dataset. Their conclusion called for an investigation on research which reused the dataset that Danescu-Niculescu-Mizil et al. (2013) developed. If the quality of dataset is spurious, then it is hard to say the following research building on the questionable dataset can avoid critics. Nonetheless, the politeness corpus of Danescu-Niculescu-Mizil et al. (2013) is now incorporated into the R package (Yeomans et al., 2018), allowing researchers from outside of ML community can easily load the package and analyze the data. There are some papers already utilized politeness corpus from outside of NLP and ML for social science research purpose (Sun et al., 2021; G Moore et al., 2020). At this point, we do not know how to measure the social impact of politeness detection dataset and the derivative package will bring.

For ML and NLP researchers, identifying who is responsible for assessing the social impact and alarming the benchmark dataset reusers is now more than important. We can find another example of social impact of malfunctioned benchmark dataset in the recent development of a chatbot called ‘Lee ruda’. ’Lee ruda’ showed how artificial intelligence systems can jeopardize sexual minorities by exposing them to toxic communication space (McCurry, 2021). If the data documentation process is not shared in the benchmark dataset sharing platforms and discussion around social impact of the dataset is not mature enough to alert reusers, the social impact of benchmark dataset can be catastrophic. In this vein, Hovy and Spruit (2016) also emphasizes how naive use of the datasets may cause problems on the society by directly deploying the trained model into the society.

4 Current Practices of Benchmark Dataset Sharing Platforms

We investigated the platforms that practitioners and researchers largely accessed for datasets. This resulted in four main platforms: HuggingFace[1], PaperswithCode [2], Tensorflow [3], and PyTorch [4]. We focused on whether it provides users easy access to datasets along with its metadata. As for PaperswithCode, it did not provide direct access to datasets however, it offered detailed information of data such as the papers that used (cited) the datasets. We excluded the platforms that were managed by the users themselves, such as Github, as it was mostly uploaded by the data creators themselves, rather than other contributors that curated the dataset for ease of use.

HuggingFace HuggingFace provides an infrastructure so ML researchers can easily leverage models and datasets. The idea of HuggingFace is similar to Github, where the codes and data are shared. In HuggingFace, it is the language model trained by different groups of researchers in NLP that is shared. Of many language models available in HuggingFace, what made HuggingFace famous is Transformers, which enabled loading thousands of deep learning frameworks (PyTorch, Tensorflow, JAX) as well as language models (e.g., BERT, RoBERTa, GPT) with a single line of code.

[1] https://huggingface.co/

[2] https://paperswithcode.com/

[3] https://www.tensorflow.org/datasets

[4] https://pytorch.org/text/stable/datasets.html

The Dataset card for social impact, biases, and unknown limitations is developed to reflect the growing concern around the social impact of the ML benchmark dataset (McMillan-Major et al., 2021).

PaperswithCode PaperswithCode organizes the research works from the ML community by providing three access points: tasks, datasets, and methods. PaperswithCode do not house the datasets but rather provide a reference point where you can find the research worked on the specific benchmark dataset. The Dataset section was organized with brief information about the dataset, relevant papers which reused the benchmark dataset, on which tasks the benchmark dataset was used, and where researchers can find the benchmark dataset. For instance, PaperswithCode introduces GLUE dataset with additional information that it can be found from Hugging Face and Tensorflow.

Tensorflow Tensorflow is an open-source library that helps the users to develop and train ML models developed by the Google Brain Team. It serves as the core platform and library for machine learning by allowing the users to customize their own models. In addition to the model library, Tensorflow also provides datasets as a collection of ready-to-use libraries. Ranging from audio, graphs, image, and texts, it offers widely used datasets including benchmarks (e.g. GLUE, SQuAD). The merit of Tensorflow datasets lies in their easy-to-use nature, as users can simply load and make use of the datasets by importing the library, except for a few exceptions that require a manual download.

PyTorch Analogous to the Tensorflow library, PyTorch is an open-source tensor library for deep learning using GPUs and CPUs, primarily developed by Facebook's AI research lab. In addition to the modules for operation, it also provides datasets and tools that make data loading easy, mainly for usability purposes. The dataset it provides is the most widely used benchmark, such as WikiText-2, CoNLL2000Chunking, for a variety of tasks including language modeling, sequence tagging, and text classification amongst others.

5 Results

As one would expect, the essential role of these platforms was focused on helping the users easily fetch the dataset and use it without putting in an extra endeavor. For example, the datasets were well-curated into train and test set splits, so that users can readily reproduce, and custom it to their own task. However, when it came to sharing auxiliary information (*metadata*) regarding the dataset, such as its limitations, and societal impacts, a large portion of the platforms lacked providing detailed information. In this section, we introduce the metadata types used in benchmark dataset sharing platforms and summarized in Table 1 of Appendix A.

5.1 Confusing concepts in terminology and metadata

HuggingFace placed 'Personal and Sensitive Information' into the big category called Dataset Creation. However, given that dataset creation includes information about source and annotation, Dataset Creation is the section for descriptive metadata. Following HuggingFace's rule of categorization, it is hard to identify whether 'Personal and Sensitive Information' is descriptive metadata that can be recorded directly from the dataset. Furthermore, the terminology that HuggingFace is using can be misleading. HuggingFace uses 'Curators' to show *"people involved in collecting the dataset and their affiliation(s)"* [5]. Using the term 'curator' to indicate people who created (collected) the dataset can be confusing. In Library and Information Science (especially the documentation field), the museum curators are people establishing collecting policies to guide the future acquisition of objects (Roberts and Light, 1980). With this sense, HuggingFace is equivalent to a museum where the virtual place houses multiple objects (datasets) and curators are people who put the dataset in the benchmark dataset sharing platform. We believe the confusing concept of curator stems from the fact that the dataset is also collected from various sources. However, a curator is the person who works at the museum or library to facilitate access or circulation of the object, not the writer or creator of the book

[5] https://github.com/huggingface/datasets/blob/master/templates/README_guide.md

or object.

5.2 Lack of documenting the limitations of the benchmark datasets

Among the platforms we investigated, only a few platforms provided the information of limitations of the benchmark datasets. HuggingFace data cards have a section that links the contributors, those who upload the dataset to the platform to write down what the data curation rationale is and how the annotations were made. However, when it comes to the limitations of the benchmark datasets, the detailed explanations about what the limitations are unclear. It is hard to find coherent concept of what limitation should be addressed. For example, one of the datasets mentioned the contextual limitation - monolingual dataset as the limitation *"(the) issue is the focus on English language and lack of multilingual hate speech."* (*hatexplain*[6]) - while the other noted the technical issue, the data size, as its limitation *"The dataset is relatively small and should be used combined with larger datasets."* (*ethos*[7]).

Similar to the Hugging Face, PaperswithCode showed the related papers, however, as the related papers were based on the citation information - whether the dataset was cited in the paper or not - it did not explicitly distinguish the papers that mentioned the limitations of the datasets. Even though one particular paper cited the dataset, it does not necessarily mean that the paper used the dataset for improving their own models. It could have been the paper discussing the caveats of using the dataset. However, this demarcation was not clear to help reusers notice whether there is a potential harm of leveraging this dataset.

As Tensorflow and PyTorch were focused more on its technical use of the datasets, only the information pertaining to how to practically use the datasets was documented. For example, the test and train splits, and the functions that were used to load the data. This different metadata recording practice in benchmark dataset sharing platforms shows that there is no consensus on what metadata to use to inform the reusers of the benchmark dataset.

5.3 Discussions of social impacts

We denote two prominent points when investigating the platforms overall regarding the discussions of social impacts. First, the platforms that documented the social impacts of the benchmark datasets barely existed. Even if there were sections for limitations, it was not clear whether the section is for discussing the social impact of dataset or technical aspect of dataset. Second, the definition of what social impacts it is referring to was obscure if there were any sections allocated to document it. For Tensorflow and PyTorch, as the main focus of these platforms are on redistribution, and enhancing the reusability of the users, the documentation did not include any discussions of the social impacts of the datasets. PaperswithCode has its unique feature, 'leaderboard' that demonstrates the state-of-the-art models that were tested on the given datasets. It allows the users to easily check the model performance based on this leaderboard. This practice, however, is far from discussing the social impacts the datasets and it does not provide the audience with potential caveats that may arise when using the dataset.

HuggingFace, on the other hand, provided the data cards which is the format the contributors need to fill when sharing the dataset, and there is a section that deals with the possible social impact of the dataset. According to the HuggingFace data card guidelines, the range of what social impact is broad ranging from positive impact to potential risks it may have to the society. One of the dataset explanations mentioned the positive social impact it can bring: *"The dataset could prove beneficial to develop models which are more explainable and less biased."* (*hatexplain*[8]), while the others focused on the functional effectiveness: *"This dataset is part of an effort to encourage text classification research in languages other than English."* (*amazon reviews*[9]), and few on the negative impacts: *"..it necessarily requires confronting online content that may be offensive or disturbing but argue that deliberate avoidance does not eliminate such problems"* (*social bias frames*[10]). This lack

[6] https://huggingface.co/datasets/hatexplain

[7] https://huggingface.co/datasets/ethos#other-known-limitations

[8] https://huggingface.co/datasets/hatexplain#social-impact-of-dataset

[9] https://huggingface.co/datasets/amazon_reviews_multi#social-impact-of-dataset

[10] https://huggingface.co/datasets/social_bias_frames#social-impact-of-dataset

of consensus on what limitation is with respect to benchmark dataset and social impact the benchmark dataset can bring can lead to haphazard organization of benchmark dataset and in turn lead to the failed control of managing implicit bias slipping into derivative NLP models and the findings of the scholars who simply utilize the NLP models.

6 Desiderata of Data Sharing Platforms

The caveat of using benchmark dataset and *vis-à-vis* social impact is gaining attention from the ML community (Hovy and Spruit, 2016; Sap et al., 2021). However, we believe the benchmark dataset sharing platform is not currently up-to-date because the current discussion around the benchmark dataset is missing. We acknowledge that the endeavor of dataset creators is crucial in developing a safe benchmark ecosystem, however, in this work we typically focus on the data sharing platforms. From our analysis, there are many loopholes to fill. PaperswithCode, Tensorflow, and PyTorch emphasized descriptive and administrative metadata while neglecting the importance of the social impact that the benchmark dataset can bring. We want to reiterate that even though data documentation recorded the entire process of dataset creation perfectly, *data friction* (Edwards et al., 2011) could happen when it was made available for others for reuse purposes. Therefore, dataset sharing platforms should take initiative to inform the social impact of the benchmark datasets by critically assessing the datasets. From a data curation perspective, it is unclear who is responsible for organizing the information of social impact, biases, and other limitations.

6.1 Beyond descriptive and administrative metadata

Metadata for administrative purposes which does not describe the dataset itself but may be of use to clarify rights and version were well-developed in four platforms. Although administrative metadata that each platform used was varied, we were able to identify that platforms tried to record licenses (HuggingFace, PaperswithCode) and versions (Tensorflow). However, metadata for social impact were absent in PaperswithCode, Tensorflow, and PyTorch. This may indicate that the discussion around the social impact of reusing the benchmark datasets stays in scholarly communication. Practitioners (both in the ML community and outside of the community) deserve the right to know the potential social impact that the benchmark dataset they are using can bring.

6.2 Data curator for social impact metadata

The next will be answering who is responsible for providing social impact metadata. We propose that the data curator specialists working for the benchmark dataset sharing platform should take a role to announce and organize the social impact of the dataset. As we discussed in the 5. Results section, the role of a curator is to manage the dataset and critically assess the social impact of reuse. It is a lack of understanding the importance of metadata and the role of the curator that made the climate of putting less emphasis on sharing social impact information. HuggingFace placed 'personal and sensitive information' into descriptive metadata section (Dataset Creation), confusing who is responsible for filling out the field of 'personal and sensitive information'. We believe sensitive information is an aspect of the dataset after critically reviewing it. Additional information can either be detected during the collection or after it is completed collecting process. However, it is more likely that sensitive information can slip into the dataset without dataset creators' notice. This makes the nature of 'personal and sensitive information' fall under metadata that needs to be addressed afterward, which is far from descriptive metadata. For instance, if the dataset was collected from social media, data curators should critically assess the dataset to identify if it contains personally identifiable information and complete the metadata section for it.

6.3 Developing social impact metadata

The benchmark dataset may have an impact on society with *exclusion* and *overgeneralization* (Hovy and Spruit, 2016). Hovy and Spruit (2016) explain that the exclusion of certain demographics in the dataset may exacerbate as the models overfit these factors. For example, models that are overfitted to *standard white English* may have the propensity to fail when applied to the products by marginalizing other demographics and their use of language can be overgeneralized. Concretely, below we list some of the possible social impact metadata that needs to be included: *which* metadata should be included, and *why* it should be considered important in terms of social impact.

Demographic statistics is about *the population from whom the data comes*. As the data for NLP deals with language, it carries contextual information beyond its face value. For example, text data retrieved from news wire may represent a typically white, educated, middle-upper class man (Garimella et al., 2019) while text data retrieved from certain social media platforms may convey the language spoken by the platform users. Likewise, the data itself may represent certain socio-demographic groups for the language models to be trained on. Thus, it is important to document the demographic statistics of the dataset. Resonating our recommendation, there is a scholarship claiming the importance of ensuring demographic variation in order to mitigate potential bias upon deployment (Hovy and Prabhumoye, 2021; Rogers et al., 2021; Ardehaly and Culotta, 2014).

Annotators demographics is about *the population who added values (labels) on top of the collected raw data*. Recording metadata about annotators demographics is related to *selection bias* and demographics of annotators accord with *label bias* (Hovy and Prabhumoye, 2021). As annotators (e.g. crowdsource workers) contribute to form the labels, their social norms can be systematically encoded in the dataset, inducing a label bias. Sap et al. (2021) demonstrates how the annotations are highly dependent on the annotator's demographics. To be specific, the task of annotating whether the text is a type of hate speech or not is hinged much on the annotators' ethnic group. It is important to document the annotators' demographics, not only because it informs the users about the representation of annotators but also it also steers future data creators to take into consideration when crowdsourcing annotators.

Besides these items, we also note the initiatives of NAACL (*the discussion of the broader impacts*[11]) and GDPR (*privacy issues of collected data*[12]) are also highly recommendable for starting a discussion on making a consensus about what social impact metadata the benchmark dataset sharing platforms should reflect.

7 Conclusion

We believe the documentation of the benchmark dataset plays an important role as it introduces the pitfalls as well as the usage of the dataset. To this end, we examine current practices of widely accessed benchmark dataset sharing platforms - *what is documented and what is omitted* -. Our findings suggest the need for documenting the social impact of the benchmark dataset as well as assigning the data curators for data sharing platforms to be in-charge of documenting relevant metadata.

References

Ehsan Mohammady Ardehaly and Aron Culotta. 2014. Using county demographics to infer attributes of twitter users. In *Proceedings of the joint workshop on social dynamics and personal attributes in social media*, pages 7–16.

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.

Samuel Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.

Randy Connolly. 2020. Why computing belongs within the social sciences. *Communications of the ACM*, 63(8):54–59.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259.

DDI. 2020. Ddi lifecycle 3.3.

Brian Dobreski, Jaihyun Park, Alicia Leathers, and Jian Qin. 2020. Remodeling archival metadata descriptions for linked archives. In *International Conference on Dublin Core and Metadata Applications*, pages 1–11.

Paul N Edwards, Matthew S Mayernik, Archer L Batcheller, Geoffrey C Bowker, and Christine L Borgman. 2011. Science friction: Data, metadata, and collaboration. *Social studies of science*, 41(5):667–690.

[11] https://2021.naacl.org/ethics/faq/

[12] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

Sarah G Moore, Gopal Das, and Anirban Mukhopadhyay. 2020. Emotional echo chambers: Observed emoji clarify individuals' emotions and responses to social media posts. *ACR North American Advances*.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Association for Computational Linguistics*.

Erin R Hoffman, David W McDonald, and Mark Zachry. 2017. Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. *Proceedings of the ACM on Human-computer Interaction*, 1(CSCW):1–14.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316.

Xiaozhong Liu and Jian Qin. 2014. An interactive metadata model for structural, descriptive, and referential representation of scholarly output. *Journal of the Association for Information Science and Technology*, 65(5):964–983.

Justin McCurry. 2021. South korean ai chatbot pulled from facebook after hate speech towards minorities.

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, 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. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

NISO. 2004. Understanding metadata. *National Information Standards Organization*, 20.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Jeffrey Pomerantz. 2015. *Metadata*. MIT Press.

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.

D Andrew Roberts and Richard B Light. 1980. Progress in documentation: museum documentation. *Journal of documentation*.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?'a checklist for responsible data use in nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

David A Smith, Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. Detecting and modeling local text reuse. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 183–192. IEEE.

Sandeep Soni, Lauren F Klein, and Jacob Eisenstein. 2021. Abolitionist networks: Modeling lan guage change in nineteenthcentury activist newspapers. *Journal of Cultural Analytics*, 1(1):43.

Shujing Sun, Yang Gao, and Huaxia Rui. 2021. Chronic complainers or increased awareness? the dynamics of social media customer service. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 6525.

Mary Vardigan, Pascal Heus, and Wendy Thomas. 2008. Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, 3(1).

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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.

Geoffrey Woledge. 1983. Historical studies in documentation:'bibliography'and 'documentation': words and ideas. *Journal of Documentation*.

Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness package: Detecting politeness in natural language. *R Journal*, 10(2).

A Appendix

Platforms	Metadata type	Items
HuggingFace	Descriptive metadata	Dataset Creation: Curation Rationale, Source Data, Annoatations
	Social impact metadata	Dataset Creation: Personal and Sensitive Information Considerations for Using the data: Social Impact of dataset, Discussions of Biases, Other known Limitations
	Administrative metadata	Additional Information Dataset: Curators, Licensing Information, Citation Information, Contributions
PaperswithCode	Descriptive metadata	Description
	Social impact metadata	
	Administrative metadata	Homepage (Link to paper), Usage (Number of papers using this dataset by year), Benchmark Leader Board (Task, Dataset variant, Best Model, Paper, Code), License, List of papers
Tensorflow	Descriptive metadata	Description, Download size, Dataset size, Auto-cached, Splits, Supervised keys, Figure
	Social impact metadata	
	Administrative metadata	Homepage (Link to paper), Source code (Example code for deployment), Versions, Examples, Citation
PyTorch	Descriptive metadata	Number of lines per split, Number of classes, Parameters
	Social impact metadata	
	Administrative metadata	Code (example code for deployment)

Table 1: Metadata of benchmark dataset sharing platforms