Navigating Dataset Documentation in ML: A Large-Scale Analysis of Dataset Cards on Hugging Face

Xinyu Yang
Cornell University

Weixin Liang
Stanford University

James Zou
Stanford University

Abstract

Advances in machine learning are closely tied to the creation of datasets. While data documentation is widely recognized as essential to the reliability, reproducibility, and transparency of ML, we lack systematic empirical understanding of current dataset documentation practices. To shed light on this question, here we take Hugging Face – one of the largest platforms for sharing and collaborating on ML models and datasets – as a prominent case study. By analyzing all 7,433 dataset documentation on Hugging Face, our investigation provides an overview of the Hugging Face dataset ecosystem and insights into dataset documentation practices, yielding 5 main findings: (1) The dataset card completion rate shows marked heterogeneity correlated with dataset popularity: While 86.0% of the top 100 downloaded dataset cards fill out all sections suggested by Hugging Face community, only 7.9% of dataset cards with no downloads complete all these sections. (2) A granular examination of each section within the dataset card reveals that the practitioners seem to prioritize *Dataset Description* and *Dataset Structure* sections, accounting for 36.2% and 33.6% of the total card length, respectively, for the most downloaded datasets. In contrast, the *Considerations for Using the Data* section receives the lowest proportion of content, accounting for just 2.1% of the text. (3) By analyzing the subsections within each section and utilizing topic modeling to identify key topics, we uncover what is discussed in each section, and underscore significant themes encompassing both technical and social impacts, as well as limitations within the *Considerations for Using the Data* section. (4) Our findings also highlight the need for improved accessibility and reproducibility of datasets in the *Usage* sections. (5) In addition, our human annotation evaluation emphasizes the pivotal role of comprehensive dataset content in shaping individuals' perceptions of a dataset card's overall quality. Overall, our study offers a unique perspective on analyzing dataset documentation through large-scale data science analysis and underlines the need for more thorough dataset documentation in machine learning research.

1 Introduction

Datasets form the backbone of machine learning research [24]. The proliferation of machine learning research has spurred rapid advancements in ML dataset development, validation, and real-world deployment across academia and industry. Such growing availability of ML datasets underscores the crucial role of proper documentation in ensuring transparency, reproducibility, and data quality in research [16, 34, 22]. Documentation provides details about the dataset, including sources of data, methods used to collect it, and preprocessing or cleaning that was performed. This information holds significant value for dataset users, as it facilitates a quick understanding of the dataset's motivation

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

and its overall scope. These insights are also crucial for fostering responsible data sharing and promoting interdisciplinary collaborations [29, 17].

Despite numerous studies exploring the structure and content of dataset cards across various research domains [1, 15, 28, 3, 10], there remains a notable gap in empirical analyses of community norms and practices for dataset documentation. This knowledge gap is significant because adherence to community norms and the quality of dataset documentation directly impact the transparency, reliability, and reproducibility in the field of data-driven research. For instance, inadequate dataset descriptions, structural details, or limitations can hinder users from utilizing the dataset appropriately, potentially resulting in misuse or unintended consequences; the absence of information on data cleaning and readiness assessment practices in data documentation limits dataset reusability and productivity gains. Furthermore, without a systematic analysis of current dataset documentation practices, we risk perpetuating insufficient documentation standards, which can impede efforts to ensure fairness, accountability, and equitable use of ML research.

To address this question, we conducted a comprehensive empirical analysis of ML dataset cards hosted on Hugging Face, one of the largest platforms for sharing and collaborating on ML models and datasets, as a prominent case study. Dataset cards on the Hugging Face platform are Markdown files that serve as the README for a dataset repository. While several open-source platforms also facilitate the sharing of ML datasets, such as Kaggle, Papers with Code, and GitHub, we chose Hugging Face for two primary reasons. Firstly, it stands out as one of the most popular platforms for developers to publish, share, and reuse ML-based projects, offering a vast repository of ML datasets for study. Secondly, Hugging Face is one of the few open-source platforms that offer an official dataset card template. This feature not only enhances the accessibility and user-friendliness of the dataset card community but also makes the analysis process more efficient and informative.

By analyzing all 7,433 ML dataset documentation hosted on Hugging Face, our investigation provides an overview of the Hugging Face dataset ecosystem and insights into dataset documentation practices. Based on our research findings, we emphasize the importance of comprehensive dataset documentation and offer suggestions to practitioners on how to write documentation that promotes reproducibility, transparency, and accessibility of their datasets, which can help to improve the overall quality and usability of the dataset community. Our study aims to bridge the notable gap in the community concerning data documentation norms, taking the first step toward identifying deficiencies in current practices and offering guidelines for enhancing ML dataset documentation.

2 Overview

Finding

- **Exponential Growth of Datasets:** HuggingFace's dataset collection has grown exponentially with a weekly growth rate of 3.97% and a doubling time of 18 weeks.
- **Documentation Associated with Usage:** 95.0% of download traffic comes from the 30.9% of datasets with documentation.

Exponential Growth of Datasets Our analysis encompasses 24,065 dataset repositories on Hugging Face uploaded by 7,811 distinct user accounts as of March 16th, 2023. The number of datasets exhibits exponential growth, with a weekly growth rate of 3.97% and a doubling time of 18 weeks (**Fig. 1***a*). As a sanity check, the number of dataset repositories reached 35,973 by May 23rd, 2023, confirming the exponential trend.

Power Law in Dataset Usage Although Hugging Face has seen a significant increase in the number of dataset repositories, our analysis reveals a significant imbalance in dataset downloads, which follows a power law distribution. This means that a small proportion of the most popular datasets receive the majority of the downloads, while the vast majority of datasets receive very few downloads. In fact, our analysis shows that just the 82 datasets with the most downloads account for 80% of total downloads (**Fig. 1***b*).

Documentation Associated with Usage Despite the importance of dataset cards, only 58.2% (14,011 out of 24,065 dataset repositories contributed by 4,782 distinct user accounts) include dataset

cards as Markdown README.md files within their dataset repositories. Among these, 6,578 dataset cards are empty, resulting in only 30.9% (7,433 out of 24,065 dataset repositories contributed by 1,982 distinct user accounts) featuring non-empty dataset cards (**Fig. 1***c*). As illustrated in **Fig. 1***b*, dataset cards are prevalent among the most downloaded datasets. Notably, datasets with non-empty dataset cards make up 95.0% of total download traffic, suggesting a possible link between dataset cards and dataset popularity. For the rest of the paper, we focus our analyses on these 7,433 non-empty dataset cards. We sort these non-empty dataset cards based on the number of downloads for the corresponding datasets. So top k dataset cards (e.g. $k = 100$) refer to the dataset cards corresponding to the k most downloaded datasets.

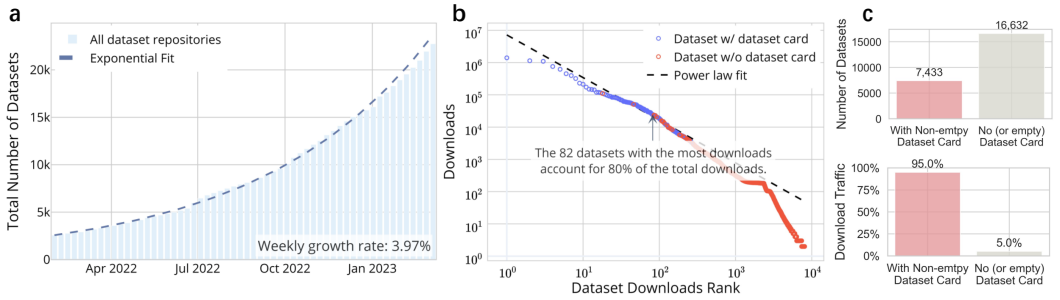

Figure 1: **Systematic Analysis of 24,065 Datasets Hosted on Hugging Face.** (**a**) *Exponential Growth of Datasets:* The Hugging Face platform has seen a remarkable surge in the number of datasets, with the count doubling approximately every 18 weeks. (**b**) *Power Law in Dataset Usage*: Dataset downloads on Hugging Face follow a power-law distribution, as indicated by the linear relationship on the log-log plot. The top 82 datasets account for 80% of the total downloads. Datasets with documentation dominate the top downloaded datasets. (**c**) *Documentation Associated with Usage:* Despite only 30.9% of dataset repositories (7,433 out of 24,065) featuring non-empty dataset cards, these datasets account for an overwhelming 95.0% of total download traffic on the platform.

3 Structure of Dataset Documentations

Finding

- **The dataset card completion rate shows marked heterogeneity correlated with dataset popularity:** While 86.0% of the top 100 downloaded datasets fill out all sections suggested by the Hugging Face community, only 7.9% of dataset cards with no downloads complete all these sections.

Community-Endorsed Dataset Card Structure Grounded in academic literature [26] and official guidelines from Hugging Face [19], the Hugging Face community provides suggestions for what to write in each section. This community-endorsed dataset card provides a standardized structure for conveying key information about datasets. It generally contains 5 sections: *Dataset Description*, *Dataset Structure*, *Dataset Creation*, *Considerations for Using the Data*, and *Additional Information* (**Table. 1**). To examine the structure of dataset cards, we used a pipeline that detects exact word matches for each section title. We then identified the section titles and checked whether they had contents. If a dataset card had all five sections completed, we considered it to be following the community-endorsed dataset card.

Adherence to Community-Endorsed Guidelines Correlates with Popularity Our evaluation revealed that popular datasets are more likely to adhere to the community-endorsed dataset card structure. **Fig. 2** illustrates significant variation in compliance with the template across datasets with different download counts. Among the 7,433 dataset cards analyzed, 86.0% of the top 100 downloaded dataset cards complete all five sections, whereas only 7.9% of

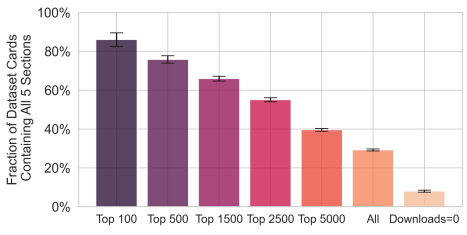

Figure 2: **Highly downloaded datasets consistently show better compliance with the community-endorsed documentation structure.**

Section Title	Subsection Title	Description
Dataset Description	Dataset Summary	A brief summary of the dataset, including its intended use, supported tasks, an overview of how and why the dataset was created, etc.
	Supported Tasks and Leaderboards	Brief description of the tag, metrics, and suggested models of the dataset.
	Languages	The languages represented in the dataset.
Dataset Structure	Data Instances	JSON-formed example and description of a typical instance in the dataset.
	Data Fields	List and describe the fields present in the dataset. Mention their data type, and whether they are used as input or output in any of the tasks the dataset currently supports.
	Data Splits	Criteria for splitting the data, descriptive statistics for the features, such as size, average length, etc.
Dataset Creation	Curation Rationale	Motivation for the creation of the dataset.
	Source Data	The source data (e.g. news text and headlines, social media posts, translated sentences, etc.), including the data collection process, and data producer.
	Annotations	Annotation process, annotation tools, annotators, etc.
	Personal and Sensitive Information	Statement of whether the dataset contains other data that might be considered sensitive (e.g., data that reveals racial or ethnic origins, financial or health data, etc.).
Considerations for Using the Data	Social Impact of Dataset	Discussion of the ways the use of the dataset will impact society.
	Discussion of Biases	Descriptions of specific biases that are likely to be reflected in the data
	Other Known Limitations	Other limitations of the dataset, like annotation artifacts.
Additional Information	Dataset Curators	The people involved in collecting the dataset and their affiliation(s)
	Licensing Information	The license and link to the license webpage if available.
	Citation Information	The BibTex-formatted reference for the dataset.
	Contributions	'Thanks to @github-username for adding this dataset.'

Table 1: **Community-Endorsed Dataset Card Structure.** This table shows the sections and their suggested subsections provided by the Hugging Face community, along with their descriptions. For more information, please refer to `https://github.com/huggingface/datasets/blob/main/templates/README_guide.md`.

dataset cards with no downloads do so. This suggests a potential correlation between adhering to community-endorsed guidelines and dataset popularity.

4 Practitioners Emphasize Description and Structure Over Social Impact and Limitations

Finding

- **Practitioners seem to prioritize on *Dataset Description* and *Dataset Structure* sections**, accounting for 36.2% and 33.6% of the total card length, respectively, on the top 100 most downloaded datasets.
- **In contrast, the *Considerations for Using the Data* section receives the lowest proportion of content, just 2.1%.** The *Considerations for Using the Data* section covers the social impact of datasets, discussions of biases, and limitations of datasets.

Social Impact, Dataset Limitations and Biases are Lacking in Most Documentations Following the community-endorsed dataset card, we conducted an analysis to determine the level of emphasis placed on each section. **Fig. 3***b* shows the word count distribution among the top 100 downloaded dataset cards, revealing their high level of comprehensiveness: 91.0% of them have a word count exceeding 200. We step further into these dataset cards to examine the emphasis placed on each section. We calculated the word count of each section and its proportion to the entire dataset card. As shown in **Fig. 3***c*, the *Dataset Description* and *Dataset Structure* sections received the most attention, accounting for 36.2% and 33.6% of the dataset card length, respectively. On the other hand, the *Considerations for Using the Data* section received a notably low proportion of only 2.1%.

Section Length Reflects Practitioner Attention The length of sections within dataset cards is reflective of practitioner attention, and it varies significantly based on the popularity of the dataset. Highly downloaded datasets tend to have more comprehensive and longer dataset cards (**Fig. 3***a*), with an emphasis on the *Dataset Description* and *Dataset Structure* sections (**Fig. 3***d*). Conversely, less popular datasets have shorter cards (**Fig. 3***a*) with a greater emphasis on the *Additional Information*

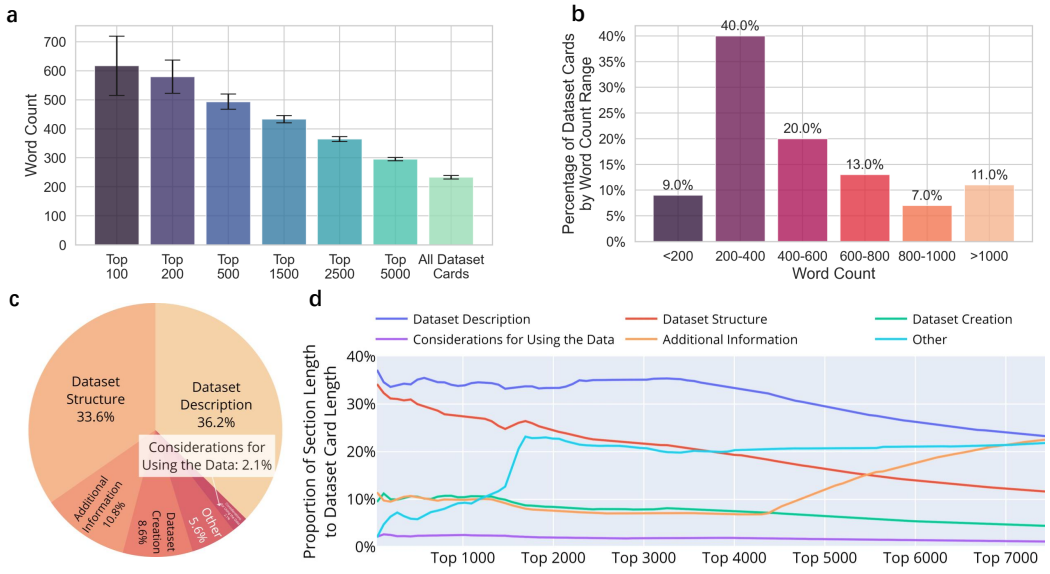

Figure 3: **Section Length Reflects Practitioner Attention.** (**a**) *Popularity Correlates with Documentation Length:* Highly downloaded dataset cards tend to be more longer. (**b**) *Distribution of Word Count Among Top 100 Downloaded Dataset Cards* (**c**) *Section Length Proportions in Top 100 Downloaded Dataset Cards:* The *Dataset Description* and *Dataset Structure* sections dominate in the top 100 downloaded dataset cards, with proportions of 36.2% and 33.6%, respectively. In contrast, the *Considerations for Using the Data* section receives the least attention, with a proportion of only 2.1%. (**d**) *Section Length Proportion Changes over Downloads:* The section length proportion changes over downloads, with *Dataset Description* and *Dataset Structure* shortening while *Additional Information* and *Other* sections lengthen. Notably, the Dataset *Creation* and *Considerations for Using the Data* sections remain consistently underemphasized across all dataset cards with different download counts.

section (**Fig. 3***d*). Despite this, sections such as *Dataset Creation* and *Considerations for Using the Data* consistently receive lower attention, regardless of download rates (**Fig. 3***d*). This suggests a need to promote more comprehensive documentation, particularly in critical sections, to enhance dataset usage and facilitate ethical considerations.

5 Practitioners Emphasize Description and Structure Over Social Impact and Limitations

Finding

- **Strong Community Adherence to Subsection Guidelines:** Practitioners in the Hugging Face community exhibit high compliance with standards, filling out 14 of the 17 recommended subsections across five main sections at a rate exceeding 50%.
- **Emergence of the *Usage* Section Beyond the Community Template:** Surprisingly, 33.2% of dataset cards includes a *Usage* section. The community template does not include such *Usage* section in its current form and should include one in the future.

Section Content Detection Pipeline To gain a deeper understanding of the topics discussed in each section, we conducted a content analysis within each section of the community-endorsed dataset card structure, which includes suggested *subsections* within the five main sections. We used exact keyword matching to identify the corresponding subsections and calculate their filled-out rates. **Fig. 4** shows that 14 out of 17 subsections have filled-out rates above 50%, indicating adherence to the community-endorsed dataset cards.

Figure 4: **Highlighting the Hugging Face Community's Compliance with Subsection Guidelines.** This figure displays subsection fill-out rates categorized by download counts in various sections. Each section contains multiple subsections, with bars indicating the fill-out rate for each. Green text highlights rates exceeding 50%, while red text indicates rates below 50%. Among the 17 subsections in the community-endorsed dataset's five sections, 14 have fill-out rates above 50%.

Limitation Section is Rare, but Long if it Exists The *Considerations for Using the Data* section (i.e., limitation section), despite being frequently overlooked and often left empty by practitioners, holds particular significance. When this section is included, it tends to adhere well to community guidelines, with subsections having a completion rate exceeding 50% and a reasonably substantial word count (98.2 words). This suggests that this section has the potential to provide valuable insights and guidance. This motivates our use of topic modeling to identify key discussion topics within this section, potentially aiding practitioners in crafting meaningful content.

a **Social Impact of Dataset**

Topic	Representative Sentences
Technical or Research Scope	• Adding a Spanish resource may help others to improve their research and educational activities. • The creation of the dataset contributes to expanding the scope of NLP research to under-explored languages across the world.
Social Scope or Background	• This dataset can be used to gain insights into the social, cultural, and political views of people in African countries. • If this matter isn't tackled with enough urgency, we might see the rise of a new dark era in Latin America politics, where many unscrupulous parties and people will manage to gain power and control the lives of many people.

b **Discussion of Biases**

Topic	Representative Sentences
Subpopulation Biases	• Gender speakers distribution is imbalanced, percentage of female speakers is mostly lower than 50% across languages. • The social biases of the time in terms of race, sex, gender, etc. might be encountered in this dataset.
Biases from Collection Procedure	• With respect to the potential risks, we note that the subjectivity of human annotation would impact on the quality of the dataset. • In terms of data collection, by using keywords and user mentions, we are introducing some bias to the data, restricting our scope to the list of keywords and users we created.

c **Other Known Limitations**

Topic	Representative Sentences
Data Quality	• The nature of the task introduce a variability in the quality of the target translations. • A number of errors, omissions and inconsistencies are expected to be found within the corpus.
Processing Limitation	• Our augmentation process can sometimes create nonexistent versions of real people. • Satellite annotation is not as accurate for pixel-level representation due to single-point annotations.

Figure 5: **Key Topics in *Considerations for Using the Data* through Topic Modeling Analysis.** This figure displays the outcomes of the topic modeling assessment on the contents of the (**a**) *Social Impact of Dataset* Subsection, (**b**) *Discussion of Biases* Subsection, and (**c**) *Other Known Limitations* Subsection. Each panel illustrates the human-assigned topic label and representative sentences for each section. Topics are generated by Latent Dirichlet Allocation (LDA).

Limitation Section Covers Diverse and Crucial Topics The *Considerations for Using the Data* section (i.e., limitation section) encompasses diverse and crucial topics. The Hugging Face community emphasizes three major themes within this section: *Social Impact of Dataset*, *Discussion of Biases*, and *Other Known Limitations*.

The *Social Impact of Dataset* aspect explores not only societal implications but also the potential benefits to technology and research communities. In this section, practitioners discuss issues like how the dataset can expand the scope of NLP research [2], and increase access to natural language technology across diverse regions and cultures [35]. Additionally, the subsection covers sensitive topics related to politics, ethics, and culture within the social scope.

Discussion of Biases delves into subpopulation bias and data collection biases, highlighting the importance of addressing bias-related issues. Previous research have identified numerous technical and social biases such as subgroup bias [7], data collection bias [36], and label bias [23]. Our topic modeling results reveal that two primary biases are discussed by practitioners in this subsection. The first is subpopulation bias, which includes biases related to gender, age, or race. For instance, an audio dataset [27] notes that female speakers are underrepresented, comprising less than 50% of the dataset. The second major bias arises from the data collection process, specifically the annotation process, which is often a significant bottleneck and source of errors.

Lastly, *Other Known Limitations* focuses on technical limitations, particularly data quality and processing limitations. This comprehensive coverage underscores the multifaceted nature of considerations related to dataset usage. Data quality is often a focus in other disciplines, such as the social sciences and biomedicine, and there are many insights to draw upon [30, 13, 12]. Meanwhile, processing limitations encompass a broader range of issues beyond biases from the collection procedure, such as inaccuracies or the absence of some data points.

Emergence of the *Usage* Section Beyond the Community Template While Hugging Face's community-endorsed dataset card structure comprises five main sections, there are instances where practitioners encounter valuable information that doesn't neatly fit into these sections. These additional sections, referred to as *Other* sections, can contain important content. Notably, among these *Other* sections, discussions related to *Usage* emerge as a frequent (nearly one-third of the time, 33.2%) and significant theme. These *Usage* sections offer a diverse range of information, including details on downloading, version specifications, and general guidelines to maximize the dataset's utility. This highlights the importance of considering content that falls outside the predefined template and suggests a potential area for improvement in dataset card templates.

Quantifying the Impact of *Usage* Section on Dataset Downloads To assess the influence of a *Usage* section in dataset documentation, we conducted a counterfactual analysis experiment (**Appendix. D**). We trained a BERT [11] model using dataset card content and download counts, which were normalized to fall within the range of [0, 1] for meaningful comparisons. When a dataset card that initially included a *Usage* section had this section removed, there was a substantial decrease of 1.85% in downloads, with statistical significance. This result underscores the significant impact of the *Usage* section in bolstering dataset accessibility and popularity.

6 Analyzing Human Perceived Dataset Documentation Quality

Finding

- **Our human annotation evaluation emphasizes the pivotal role of *comprehensive dataset content* in shaping individuals' perceptions of a dataset card's overall quality.**

Human Annotations for Comprehensive Evaluation of Dataset Card Quality We utilized human annotations to evaluate the quality of dataset cards, considering seven distinct aspects, drawing from prior research in dataset documentation literature [1, 15, 28, 3, 10]: (1) Structural Organization, (2) Content Comprehensiveness, (3) Dataset Description, (4) Dataset Structure, (5) Dataset Preprocessing, (6) Usage Guidance, and (7) Additional Information. Structural Organization and Content Comprehensiveness constitute the overall presentation of the dataset card. Dataset Description, Dataset Structure, and Additional Information can be found in sections of community-endorsed

dataset cards, while Data Preprocessing and Usage Guidance are prominent aspects highlighted in the literature. To conduct this assessment, we randomly selected 150 dataset cards and enlisted five human annotators. These annotators were tasked with evaluating each card across these seven aspects and providing an overall quality score within a 5-point range. The overall quality score reflects the subjective judgment of the annotators, considering the seven aspects as well as their overall impression. This evaluation approach aims to provide a comprehensive assessment of dataset card quality, reflecting the importance of these aspects in effective dataset documentation.

Human Perception of Documentation Quality Strongly Aligns with Quantitative Analysis Human annotation evaluation of dataset cards shows varying scores across different aspects. While Dataset Description (2.92/5), Structural Organization (2.82/5), Data Structure (2.7/5), and Content Comprehensiveness (2.48/5) received relatively higher scores, areas like Data Preprocessing (1.21/5) and Usage Guidance (1.14/5) scored lower. This aligns with the quantitative analysis that indicates a greater emphasis on the *Dataset Description* and *Dataset Structure* sections. Notably, even the highest-scoring aspect, Dataset Description, falls below 60% of the highest possible score, indicating room for improvement in dataset documentation.

Content Comprehensiveness has the strongest positive correlation with the overall quality of a dataset card (Coefficient: 0.3935, p-value: 3.67E-07), emphasizing the pivotal role of comprehensive dataset content in shaping individuals' perceptions of a dataset card's overall quality. Additionally, aspects like Dataset Description (Coefficient: 0.2137, p-value: 3.04E-07), Structural Organization (Coefficient: 0.1111, p-value: 2.17E-03), Data Structure (Coefficient: 0.0880, p-value: 6.49E-03), and Data Preprocessing (Coefficient: 0.0855, p-value: 2.27E-03) also significantly contribute to people's evaluations of dataset documentation quality. Moreover, the length of a dataset card is positively related to Content Comprehensiveness (p-value: 1.89E-011), reinforcing the importance of detailed documentation in enhancing dataset quality and usability.

7 Related Works

Dataset documentation has long been discussed, but a systematic analysis of the current dataset documentation practices is lacking in the literature. A long-standing problem in the literature is that there is no industry standard being formed about data documentation. Therefore, most existing work in the literature has been in exploring, conceptualizing and proposing different dataset documentation frameworks. Data-focused tools such as datasheets for datasets and data nutrition labels have been proposed to promote communication between dataset creators and users, and address the lack of industry-wide standards for documenting AI datasets [5, 6, 31, 15, 18, 9]. These tools provide detailed information on the composition, collection process, recommended uses, and other contextual factors of datasets, promoting greater transparency, accountability, and reproducibility of AI results while mitigating unwanted biases in AI datasets. Additionally, they enable dataset creators to be more intentional throughout the dataset creation process. Consequently, datasheets and other forms of data documentation are now commonly included with datasets, helping researchers and practitioners to select the most appropriate dataset for their particular needs. Additionally, some documentation tools gradually have a focus on the data lifecycle, which includes aspects such as assembly, collection, and annotation [21]. Such documentation tools have since expanded to cover the entire AI development lifecycle in a more comprehensive manner, with an emphasis on an iterative design process that ensures accessibility for users with diverse backgrounds and goals when interacting with dataset cards. Researchers are also advocating topics on how to create high-quality and responsible documentation to be incorporated into the AI curriculum [33, 14, 32, 4, 25], which could improve the scope and usage of dataset documentation by education. Despite the proliferation of dataset documentation tools and the growing emphasis on them, the current landscape of dataset documentation remains largely unexplored. Specifically, a systematic analysis of the current dataset documentation practices is lacking in the literature. To address this gap, we present a comprehensive analysis of ML dataset documentation on Hugging Face to provide insights into current dataset documentation practices.

8 Discussion

In this paper, we present a comprehensive large-scale analysis of 7,433 ML dataset documentation on Hugging Face. The analysis offers insights into the current state of adoption of dataset cards by the

community, evaluates the effectiveness of current documentation efforts, and provides guidelines for writing effective dataset cards. Overall, our main findings cover 5 aspects:

- *Varied Adherence to Community-Endorsed Dataset Card:* We observe that high-downloaded dataset cards tend to adhere more closely to the community-endorsed dataset card structure.
- *Varied Emphasis on Sections:* Our analysis of individual sections within dataset cards reveals that practitioners place varying levels of emphasis on different sections. For instance, among the top 100 downloaded dataset cards, *Dataset Description* and *Dataset Structure* sections receive the most attention. In contrast, the *Considerations for Using the Data* section garners notably lower engagement across all downloads, with only approximately 2% of dataset cards containing this section. This discrepancy can be attributed to the section's content, which involves detailing limitations, biases, and the societal impact of datasets – a more complex and nuanced endeavor. An internal user study conducted by Hugging Face [20] also identified the *Limitation* section within this category as the most challenging to compose.
- *Topics Discussed in Each Section:* Our examination of subsections within each section of dataset cards reveals a high completion rate for those suggested by the Hugging Face community. This highlights the effectiveness of the community-endorsed dataset card structure. We pay particular attention to the *Considerations for Using the Data* section in our study, employing topic modeling to identify key themes, including technical and social aspects of dataset limitations and impact.
- *Importance of Including Usage Sections:* We observe that many dataset card creators go beyond the recommended structure by incorporating *Usage* sections, which provide instructions on effectively using the dataset. Our Empirical experiment showcases the potential positive impact of these *Usage* sections in promoting datasets, underscoring their significance.
- *Human Evaluation of Dataset Card Quality:* Our human evaluation of dataset card quality aligns well with our quantitative analysis. It underscores the pivotal role of Content Comprehensiveness in shaping people's assessments of dataset card quality. This finding offers clear guidance to practitioners, emphasizing the importance of creating comprehensive dataset cards. Moreover, we establish a quantitative relationship between Content Comprehensiveness and the word length of dataset cards, providing a measurable method for evaluation.

Limitations and Future Works Our analysis of ML dataset documentation relies on the distinctive community-curated resource, Hugging Face, which may introduce biases and limitations due to the platform's structure and coverage. For example, Hugging Face's NLP-oriented concentration could introduce biases into the dataset categories. Additionally, our analysis of completeness and informativeness is based on word count and topic modeling, which may not fully capture the nuances of the documentation. Furthermore, measuring dataset popularity based on downloads alone may not fully reflect the dataset's impact. Future research could consider additional factors, such as the creation time of the dataset and research area of the dataset (**Appendix. E**). Lastly, our human evaluation serves as a preliminary evaluation. Future analyses could involve a more diverse group of annotators with varying backgrounds and perspectives.

Research Significance To summarize, our study uncovers the current community norms and practices in ML dataset documentation, and demonstrates the importance of comprehensive dataset documentation in promoting transparency, accessibility, and reproducibility in the ML community. We hope to offer a foundation step in the large-scale empirical analysis of dataset documentation practices and contribute to the responsible and ethical use of ML datasets while highlighting the importance of ongoing efforts to improve dataset documentation practices.

Acknowledgments and Disclosure of Funding

We thank Y. Yin for his helpful comments and discussions. J.Z. is supported by the National Science Foundation (CCF 1763191 and CAREER 1942926), the US National Institutes of Health (P30AG059307 and U01MH098953) and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative.

References

[1] S. Afzal, R. C, M. Kesarwani, S. Mehta, and H. Patel. Data readiness report, 2020.

[2] R.-A. Armstrong, J. Hewitt, and C. Manning. Jampatoisnli: A jamaican patois natural language inference dataset. *arXiv preprint arXiv:2212.03419*, 2022.

[3] N. Barman, Y. Reznik, and M. Martini. Datasheet for subjective and objective quality assessment datasets, 2023.

[4] J. Bates, D. Cameron, A. Checco, P. Clough, F. Hopfgartner, S. Mazumdar, L. Sbaffi, P. Stordy, and A. de la Vega de León. Integrating fate/critical data studies into data science curricula: where are we going and how do we get there? In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 425–435, 2020.

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

[6] E. M. Bender, B. Friedman, and A. McMillan-Major. A guide for writing data statements for natural language processing, 2021.

[7] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[8] I. Chalkidis, T. Passini, S. Zhang, L. Tomada, S. F. Schwemer, and A. Søgaard. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, 2022.

[9] K. S. Chmielinski, S. Newman, M. Taylor, J. Joseph, K. Thomas, J. Yurkofsky, and Y. C. Qiu. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954*, 2022.

[10] M. R. Costa-jussà, R. Creus, O. Domingo, A. Domínguez, M. Escobar, C. López, M. Garcia, and M. Geleta. Mt-adapted datasheets for datasets: Template and repository, 2020.

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

[12] W. Fan and F. Geerts. Foundations of data quality management. *Synthesis Lectures on Data Management*, 4(5):1–217, 2012.

[13] V. Fedorov. Optimal experimental design. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):581–589, 2010.

[14] C. Fiesler, N. Garrett, and N. Beard. What do we teach when we teach tech ethics? a syllabi analysis. In *Proceedings of the 51st ACM technical symposium on computer science education*, pages 289–295, 2020.

[15] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, nov 2021.

[16] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, M. A. Q. C. M. S. B. of Directors Shraddha Thakkar 35 Kusko Rebecca 36 Sansone Susanna-Assunta 37 Tong Weida 35 Wolfinger Russ D. 38 Mason Christopher E. 39 Jones Wendell 40 Dopazo Joaquin 41 Furlanello Cesare 42, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, et al. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16, 2020.

[17] L. Hemphill, A. Pienta, S. Lafia, D. Akmon, and D. A. Bleckley. How do properties of data, their curation, and their funding relate to reuse? *Journal of the Association for Information Science and Technology*, 73(10):1432–1444, 2022.

[18] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

[19] HuggingFace. Huggingface dataset card guidebook, 2021. Accessed: 2023-05-23.

[20] HuggingFace. Model card user studies. `https://huggingface.co/docs/hub/model-cards-user-studies`, 2022.

[21] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575, 2021.

[22] M. Hutson. Artificial intelligence faces reproducibility crisis, 2018.

[23] H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[24] B. Koch, E. Denton, A. Hanna, and J. G. Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[25] P. M. Leidig and L. Cassel. Acm taskforce efforts on computing competencies for undergraduate data science curricula. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*, pages 519–520, 2020.

[26] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[27] E. O. Nsoesie and S. Galea. Towards better Data Science to address racial bias and health equity. *PNAS Nexus*, 1(3), 07 2022. pgac120.

[28] O. Papakyriakopoulos, A. S. G. Choi, W. Thong, D. Zhao, J. Andrews, R. Bourke, A. Xiang, and A. Koenecke. Augmented datasheets for speech datasets and ethical decision-making. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2023.

[29] I. V. Pasquetto, B. M. Randles, and C. L. Borgman. On the reuse of scientific data. 2017.

[30] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, nov 2021.

[31] M. Pushkarna, A. Zaldivar, and O. Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

[32] R. Reich, M. Sahami, J. M. Weinstein, and H. Cohen. Teaching computer ethics: A deeply multidisciplinary approach. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 296–302, 2020.

[33] H. Shen, W. H. Deng, A. Chattopadhyay, Z. S. Wu, X. Wang, and H. Zhu. Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 850–861, 2021.

[34] V. Stodden, J. Seiler, and Z. Ma. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11):2584–2589, 2018.

[35] A. M. Tache, M. Gaman, and R. T. Ionescu. Clustering word embeddings with self-organizing maps. application on laroseda – a large romanian sentiment data set. *ArXiv*, 2021.

[36] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.

A Code and Data Availability

We have assembled a collection of dataset cards as a community resource, which includes extracted metadata such as the number of downloads and textual analyses. This resource along with our analysis code can be accessed at `https://anonymous.4open.science/r/HuggingFace-Dataset-Card-Analysis`.

The repository comprises two main components: the $Data$ folder and the $Scripts$ folder. The $Data$ folder contains data on 7,433 dataset cards that have been analyzed, along with metadata for each dataset and dataset card. Details about this metadata can be found in **Fig. S1**. The $Scripts$ folder contains the code used to conduct the analysis, which includes instructions for accessing the data through the Hugging Face API, an overview of the dataset community on Hugging Face, and an analysis of the dataset cards.

a

	dataset_name	author	dataset_creation_time	downloads	has_card	has_nonempty_card	task	domain
0	super_glue	huggingface	Tue Jan 25 16:34:18 2022 +0100	1403269.0	True	True	text-classification,token-classification,quest...	nlp
1	glue	huggingface	Tue Jan 25 16:34:03 2022 +0100	1140355.0	True	True	text-classification	nlp
...	...	...	...	...	...	...	...	...
24063	fffhyyhh666/Mouth-64	fffhyyhh666	Mon Aug 29 14:52:42 2022 +0000	0.0	False	False	None	None
24064	IronDice/esdeath	IronDice	Fri Mar 10 21:46:51 2023 +0000	0.0	False	False	None	None

24065 rows × 8 columns

b

	dataset_name	author	dataset_creation_time	downloads	task	domain	dataset_card	total_word_cnt	follow_template
0	super_glue	huggingface	Tue Jan 25 16:34:18 2022 +0100	1403269.0	text-classification,token-classification,quest...	nlp	---\nannotations_creators:\n- expert-generated...	517.0	1.0
1	glue	huggingface	Tue Jan 25 16:34:03 2022 +0100	1140355.0	text-classification	nlp	---\nannotations_creators:\n- other\nlanguage_...	1388.0	1.0
...	...	...	...	...	...	...	...	...	...
7431	irds/mmarco_v2_vi_train	irds	Thu Jan 5 03:29:58 2023 +0000	0.0	text-retrieval	None	---\npretty_name: "mmarco/v2/vi/train"\nview...	74.0	0.0
7432	autoevaluate/autoeval-staging-eval-project-976...	autoevaluate	Thu Jul 21 15:35:27 2022 +0000	0.0	None	None	---\ntype: predictions\ntags:\n- autotrain\n- ...	48.0	0.0

7433 rows × 9 columns

c

	dataset description					dataset structure		
	has_section	section_length_proportion	subsection_title	section_content	word_cnt	not_empty	has_section	section_leng
super_glue	1	0.268182	Dataset Summary	Dataset Description\nHomepage: https://github....	118	1	1	
glue	1	0.712919	Supported Tasks and Leaderboards;Languages;Dat...	Dataset Description\nHomepage: https://nyu-mll...	894	1	1	
...	...	...	...	...	...	...	...	...
irds/mmarco_v2_vi_train	0	0.000000	None	None	0	0	0	

7433 rows × 36 columns

Figure S1: **Metadata Provided by the Repository for the Datasets and Dataset Cards.** (**a**) *Metadata for the Datasets:* The $dataset\_info.parquet$ in the $Data$ folder stores the metadata we extracted of the 24,065 datasets as of Mar 16th, 2023. The metadata include the creation time, author, downloads, whether the dataset has a (non-empty) dataset card, the task category, and the task domain of the dataset. (**b**) *Metadata for the Datasets Cards:* The $datasetcard\_info.parquet$ in the $Data$ folder stores the information we extracted of the 7,433 dataset cards. The information include the dataset name, author, creation time, number of downloads, task category, task domain, content ot the dataset card, total word count, and whether the dataset card follows the template. (**c**) *Information about the Sections of the Dataset Cards:* The $datasetcard\_sections\_info.parquet$ in the $Data$ folder stores the information of the sections of the dataset cards. The sections include Dataset Description, Dataset Structure, Dataset Creation, Considerations for Using the Data, Additional Information. For each section, we provide whether a dataset card has this section (and whether it's empty), the subsections of the section, section length proportion of the section, the content of the section, and the word count of the section.

B Illustrations for Dataset Cards Suggested by Hugging Face Community

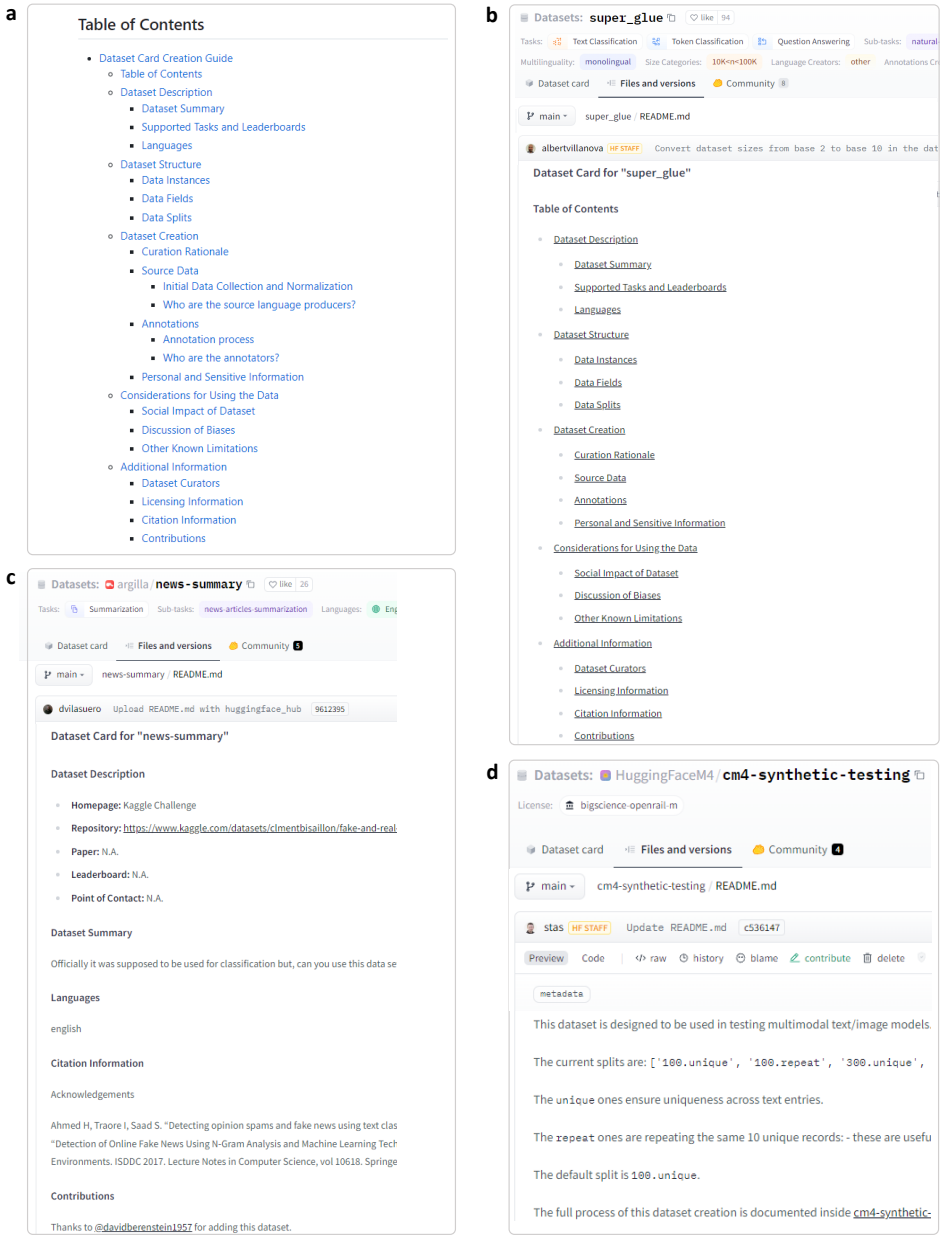

Figure S2: **Illustration of Adherence to Community-Endorsed Dataset Card.** (**a**) *Community-Endorsed Dataset Card Struture:* Hugging Face community provides a suggested dataset card structure, which contains five main sections: *Dataset Description*, *Dataset Structure*, *Dataset Creation*, *Considerations for Using the Data*, and *Additional Information*. (**b**) *Example of a Dataset Card Conforming to the Community Guidelines:* A dataset card is considered to conform to the community guidelines when it includes the five main sections outlined in the community guidelines, with the corresponding content provided for each section. (**c**) *Example of Dataset Cards Not Following Community Guidelines (1):* A dataset card is considered non-conforming if it omits any of the five main sections provided in the suggested dataset card structure. (**d**) *Example of Dataset Cards Not Following Community Guidelines (2):* This dataset card contains only a few words and does not follow the structure at all.

C Method

C.1 Accessing and Parsing Dataset Cards

In this work, we analyze datasets hosted on Hugging Face, a popular platform that provides a wealth of tools and resources for AI developers. One of its key features is the Hugging Face Hub API, which grants access to a large library of pre-trained models and datasets for various tasks. With this API, we obtained all 24,065 datasets hosted on the Hub as of March 16th, 2023.

Dataset cards are Markdown files that serve as the README for a dataset repository. They provide information about the dataset and are displayed on the dataset's homepage. We downloaded all dataset repositories hosted on Hugging Face and extracted its README file to get the dataset cards. For further analysis of the documentation content, we utilized the Python package mistune (`https://mistune.readthedocs.io/en/latest/`) to parse the README file and extract the intended content. The structure of dataset cards typically consists of five sections: *Dataset Description*, *Dataset Structure*, *Dataset Creation*, *Additional Information*, and *Considerations for Using the Data*, as recommended by Hugging Face community. Examples of dataset cards, as shown in **Fig. S2**, illustrate the essential components and information provided by dataset cards. We identified and extracted different types of sections through parsing and word matching of the section heading.

C.2 Human-Annotated Dataset Card Evaluation Methodology and Criteria

We conducted an evaluation on a sample of 150 dataset cards from a total of 7,433. The assessment involved five human annotators and focused on seven key aspects of the dataset cards:

- **Structural Organization:** How well is the documentation structured with headings, sections, or subsections?
- **Content Comprehensiveness:** How comprehensive is the information provided in the documentation?
- **Dataset Description:** How effectively does the documentation describe the dataset?
- **Dataset Structure:** How well does the documentation explain the underlying data structure of the dataset?
- **Dataset Preprocessing:** How well does the documentation describe any preprocessing steps applied to the data?
- **Usage Guidance:** How well does the documentation offer guidance on using the dataset?
- **Additional Information:** How well does the documentation provide extra details such as citations and references?

Each aspect received a score on a scale from 0 to 5, with the following score metrics:

Score	Description
5	Exceptionally comprehensive and effective
4	Very good and thorough
3	Moderately satisfactory
2	Insufficient
1	Poor and inadequate
0	Absent

Table S1: **Metrics of the Scores**

D Additional Analysis of *Usage* Section

Among 7,433 dataset cards, there are 567 dataset cards uploaded by 52 distinct practitioners that contain a *Usage* section, instructing how to use the dataset through text and codes. A specific example of *Usage* section is from ai4bharat/naamapadam, which has 469 downloads and has a *Usage* section to instruct how to use the dataset (**Fig. S3**).

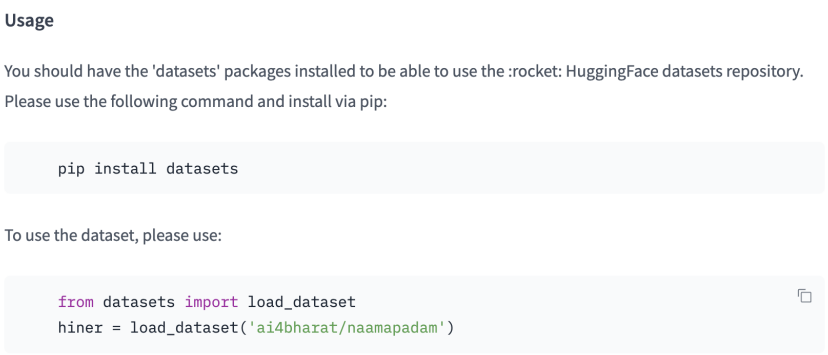

Figure S3: **Example of a *Usage* Section**

Intuitively, a *Usage* section could give users quick instructions on how to use the dataset, which could make the dataset more accessible, transparent, and reproducible. To verify this intuition, we conduct an experiment to quantify how the *Usage* section will affect the dataset's popularity.

In our experiment, we trained a BERT [11] Model using the content of dataset cards and their corresponding download counts. To ensure comparability, the download counts were normalized to a range of [0,1] and stratified monthly based on the dataset's creation time. This ranking system assigned a rank of 1 to the dataset with the highest downloads within a given month, and a rank of 0 to the dataset with the lowest downloads.

Using the dataset card content, the trained BERT Model predicted the download counts. Subsequently, we conducted a test using 567 dataset cards that included a *Usage* section. For this test, we deliberately removed the *Usage* section from the dataset cards and employed the BERT Model to predict the download counts for these modified cards. The resulting predictions are summarized in the table below:

	Predicted Score of Downloads
Dataset Card with Usage Section	0.3917
Remove the Usage Section	0.3732
Reduction upon Removal	**-0.0185**

Table S2: **Impact of Usage Section on Predicted Score of Downloads**

The average predicted score of downloads after removing the *Usage* section is 0.0185 lower compared to the original dataset card. This indicates a decrease in the number of downloads, highlighting the negative impact of not including a *Usage* section.

In future research, it would be valuable to further investigate the effect of adding a *Usage* section to the dataset cards that do not have one originally. A randomized controlled trial (RCT) experiment could be conducted to assess whether the inclusion of a *Usage* section leads to an increase in downloads.

E Optional Metrics for Datasets

In our analysis, we employ downloads as a metric to gauge the popularity of the dataset. Numerous factors can influence the download count, including the dataset's publication date and its associated research field. Moreover, aside from dataset downloads, we can incorporate other indicators of dataset popularity, such as the count of models utilizing the datasets and the corresponding download counts.

To address the concerns of factors that might affect downloads, we expanded our dataset analysis by extracting more metadata from the Hugging Face dataset information. We collected data such as the models utilizing the corresponding dataset, the total number of downloads for these models, and the dataset's task domain. The primary dataset tasks recognized by Hugging Face encompass Multimodal, Computer Vision, Natural Language Processing, Audio, Tabular and Reinforcement Learning. Among the total of 7,433 dataset cards, 1,988 are categorized as NLP dataset cards, 198 are related to computer vision, and 102 pertain to multimodal datasets. We proceeded with additional analysis by employing the following metrics:

1. We integrated dataset downloads with the downloads of models employing the dataset, which can be termed as *"secondary usage of the dataset"*.
2. Task domains were specified.
3. A time range (measured in months) was selected, encompassing dataset cards created within the designated time frame and domain.
4. Selected dataset cards were ranked within each domain for each time range and then normalized to a range of $[0, 1]$.

By adopting this approach, we account for the dataset's publication time, task domain, secondary dataset usage, as well as the number of downloads. We conducted a word count analysis using this new metric and attained results consistent with our prior analysis that datasets with higher rankings tend to have longer dataset cards, as shown in **Fig. S4**.

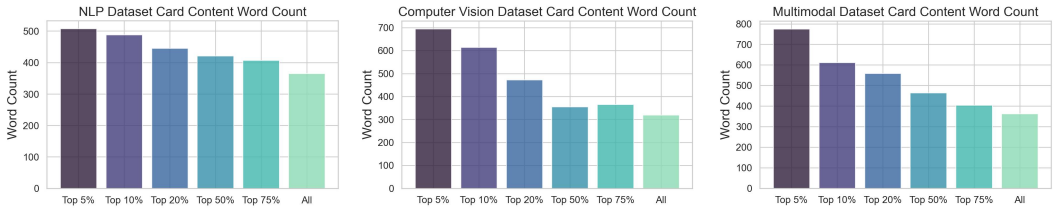

Figure S4: **Length Correlates with Dataset Quality.** In the updated metrics, there's a notable trend where higher-ranked dataset cards tend to be longer. This suggests that these dataset cards encompass more comprehensive and detailed information.

The finding enables us to contemplate an alternative metric option, factoring in publication time, research area, and secondary dataset usage. However, the results remain aligned with our previous analysis, which solely considered download counts, highlighting the reasonableness of using download counts as metrics.

F Additional Analysis of Each Section in the Dataset Card

Section. 5 offers a concise summary of each section, complemented by topic modeling results for the most engaging section, *Considerations for Using the Data*. In addition, **Table. 1** provides a clear presentation of the community-endorsed dataset card, including suggested sections, subsections, and their corresponding descriptions. The completion rates of subsections within each section are depicted in **Fig. 4**, which suggests a general adherence to the community-endorsed dataset card. In the subsequent paragraph, a comprehensive analysis of each section is provided, offering further insight into the content covered.

Dataset Description The *Dataset Description* section contains the fundamental information about a dataset, and is comprised of three subsections: *Dataset Summary*, *Supported Tasks and Leaderboards*, and *Languages*. As depicted in **Fig. 4**, *Dataset Summary* is the most frequently filled-out subsection in the *Dataset Description* section, with a filled-out rate of 94.5% and 80.0% in the top 100 downloaded dataset cards and all 7,433 dataset cards, respectively. This underscores the importance of providing a brief summary of the dataset, which can enhance its accessibility to users and, in turn, promote its use. On the other hand, the finer-grained subsections of *Dataset Description*, such as *Supported Tasks and Leaderboards* and *Languages*, have a relatively low filled-out rate. This may be due to the fact that people tend to provide only a brief mention of this information in the *Dataset Summary* section, instead of elaborating on it in a separate section. However, separating this information into distinct subsections can help to emphasize its importance. Given that tasks and languages are essential features of a dataset, it could be better for developers to follow the guidelines and write the information in the corresponding sections.

Dataset Structure Overall, dataset cards conform well to the official guidelines in the *Dataset Structure* section, particularly in the case of the top 100 downloaded dataset cards. Specifically, 95.3% of the top 100 downloaded dataset cards contain *Data Instances* in the *Dataset Structure* section, 98.8% of them contain *Data Fields*, and 97.7% of them contain *Data Splits*. The *Dataset Structure* section offers detailed information about the dataset's composition, with *Data Instances* providing examples and descriptions of typical instances in the dataset, *Data Fields* describing the fields present in the dataset, and *Data Splits* providing information about the criteria for splitting the data, as well as the size and name of each split. The high filled-out rate of these subsections highlights their importance and serves as an example for practitioners to follow when providing information about the *Dataset Structure*.

Dataset Creation *Dataset Creation* encompasses both technical and ethical considerations. Technical aspects, such as *Source Data*, which provides information about the initial data collection and normalization, and the source language producers, have the highest filled-out rate, at 70.8% and 70.6% for all datasets and the top 100 downloaded datasets, respectively. The *Annotations* subsection, which includes information about the annotation process and annotators, receives moderate attention, with a filled-out rate of 59.5% and 52.8% for all dataset cards and the top 100 downloaded dataset cards, respectively. Subjective issues, such as *Curation Rationale*, which outlines the motivation and reasons behind dataset curation, are included in 55.8% of dataset cards within the *Dataset Creation* section. Notably, the *Personal and Sensitive Information* subsection has a low filled-out rate, with only 35.3% of dataset cards discussing it in the *Dataset Creation* section. This is understandable, as limited datasets contain sensitive data that reveals information such as racial or ethnic origins, religious beliefs, political opinions, and so on. Nevertheless, this subsection is indispensable, as it helps ensure that the dataset is being handled ethically and in compliance with relevant regulations and laws. By providing information about any personal or sensitive data in the dataset, researchers and data scientists can take appropriate measures to protect the privacy and security of individuals represented in the data.

Considerations for Using the Data **Section. 4** highlights that *Considerations for Using the Data* is the section of a dataset card that receives the lowest attention. However, despite this, three prominent topics discussed in this section have been identified by the community: *Social Impact of Dataset*, *Discussion of Biases*, and *Other Known Limitations*. These topics are prevalent among both the entire set of 7,433 dataset cards and the top 100 downloaded dataset cards, all have a filled-out rate larger than 50%. Specifically, 80.0% of the top 100 downloaded dataset cards that include *Considerations for Using the Data* discuss the *Social Impact of Dataset*, describing the potential ways that the dataset

may impact society. For example, the datasets for evaluating the fairness of pre-trained legal language models and techniques [8] states the following sentence in its *Social Impact of Dataset* section: "This work can help practitioners to build assisting technology for legal professionals with respect to the legal framework (jurisdiction) they operate." Additionally, 73.3% of the top 100 downloaded dataset cards discuss the biases of the dataset, such as biases of the data distribution or data collection process. (e.g. "This dataset is imbalanced"; "Since the data is from human annotators, there are likely to be biases.") The *Other Known Limitations* subsection outlines other limitations of the dataset, such as annotation artifacts, and is present in 57.2% of the *Considerations for Using the Data* sections. This subsection is important because it helps potential users understand the potential limitations and drawbacks of the dataset, which can inform their decision-making process when selecting a dataset for their research.

Overall, the high filled-out rate of the subsections of *Considerations for Using the Data* underscores the importance of considering the potential biases and limitations of a dataset, as well as its potential impact on society, when selecting and using a dataset for research purposes, and suggests researchers and data scientists are increasingly put more emphasis on the ethical and technical implications of their work.

Additional Information The *Additional Information* section of the dataset card includes details about the dataset curators, licensing information, citation information, and contributions. Our analysis shows a high rate of completion for citation information and contributions among the top 100 downloaded dataset cards that include this section. Of the top 100 downloaded dataset cards that contain *Additional Information*, 95.6% include the *Contributions* section, which typically acknowledges contributors with a statement like "Thanks to @github-username for adding this dataset", as suggested by the community-endorsed dataset card. Additionally, 94.5% of these dataset cards include citation information in BibTex format.

These findings emphasize the importance that researchers place on community sharing and recognition of contributions. Such emphasis can promote a healthy community ecosystem for sharing and discussing ideas and therefore prompt the development of the research field.

Other The *Other* section in a dataset card includes topics that are not covered by the five sections of the community-endorsed dataset card. Our analysis identifies two prominent topics that people discuss in this section. The first is *About*, which is similar to the *Dataset Description* section and accounts for 16.6% of *Other* sections. The second is *Usage*, which has a 33.2% filled-out rate of all discussions in the *Other* section. Indeed, the *Usage* section in a dataset card is important because it could provide users with information on how to use the dataset, including instructions on how to download and access the data, as well as how to preprocess or transform the data for various use cases. A clear and detailed *Usage* section can help users avoid common pitfalls or errors, saving time and effort for researchers and developers who are using the dataset for their projects. This, in turn, increases the reproducibility, transparency, and usage of the dataset. We suggest that dataset creators include a comprehensive *Usage* section in their dataset card to facilitate the use and reproducibility of the dataset. Furthermore, we recommend that the community incorporates this key information into their suggested dataset card to better serve the needs of the community.